MULTIPLE IINEAR REGRESSION ON MARKETTING DATA IN DATARIUM USING R

Srimanta Singha

2023-07-04

DESCRIPTION OF DATASET:

“Marketing” data set is a data frame containing the impact of three advertising medias (YouTube, facebook and newspaper) on sales. Data are the advertising budget in thousands of dollars along with the sales (in thousands of units). The advertising experiment has been repeated 200 times. This is a simulated data.

Dataset Loading:

```
library(datarium)
data=marketing # This is our dataset
head(data) # showing first six rows of dataset
```

```
##   youtube facebook newspaper sales
## 1  276.12    45.36     83.04 26.52
## 2   53.40    47.16     54.12 12.48
## 3   20.64    55.08     83.16 11.16
## 4  181.80    49.56     70.20 22.20
## 5  216.96    12.96     70.08 15.48
## 6   10.44    58.68     90.00  8.64
```

```
str(data) # structure of dataset
```

```
## 'data.frame':    200 obs. of  4 variables:
##  $ youtube  : num  276.1 53.4 20.6 181.8 217 ...
##  $ facebook : num  45.4 47.2 55.1 49.6 13 ...
##  $ newspaper: num  83 54.1 83.2 70.2 70.1 ...
##  $ sales    : num  26.5 12.5 11.2 22.2 15.5 ...
```

This shows that our dataset contains three variables which are all continuos variable.

AIM:

The aim of this project is to find out the features which influences the sales and build a multiple linear regression(if possible) model to predict the amount of sales in future.

Here we want to predict the sales based on the advertising budget invested in given three platform(YouTube, facebook and newspaper) so we will consider the variable “sales” as response variable and others variables as predictor variable.

Features selection:

Since we have three features in our dataset but it is possible that all of them may not influences the sales so ,in this part, we will extract the features which are the most important to increase the sales.

For this we will create the cor-relation matrix to find out the features which are the most linearly dependent variable to the response variable and we will take those features in our model later.

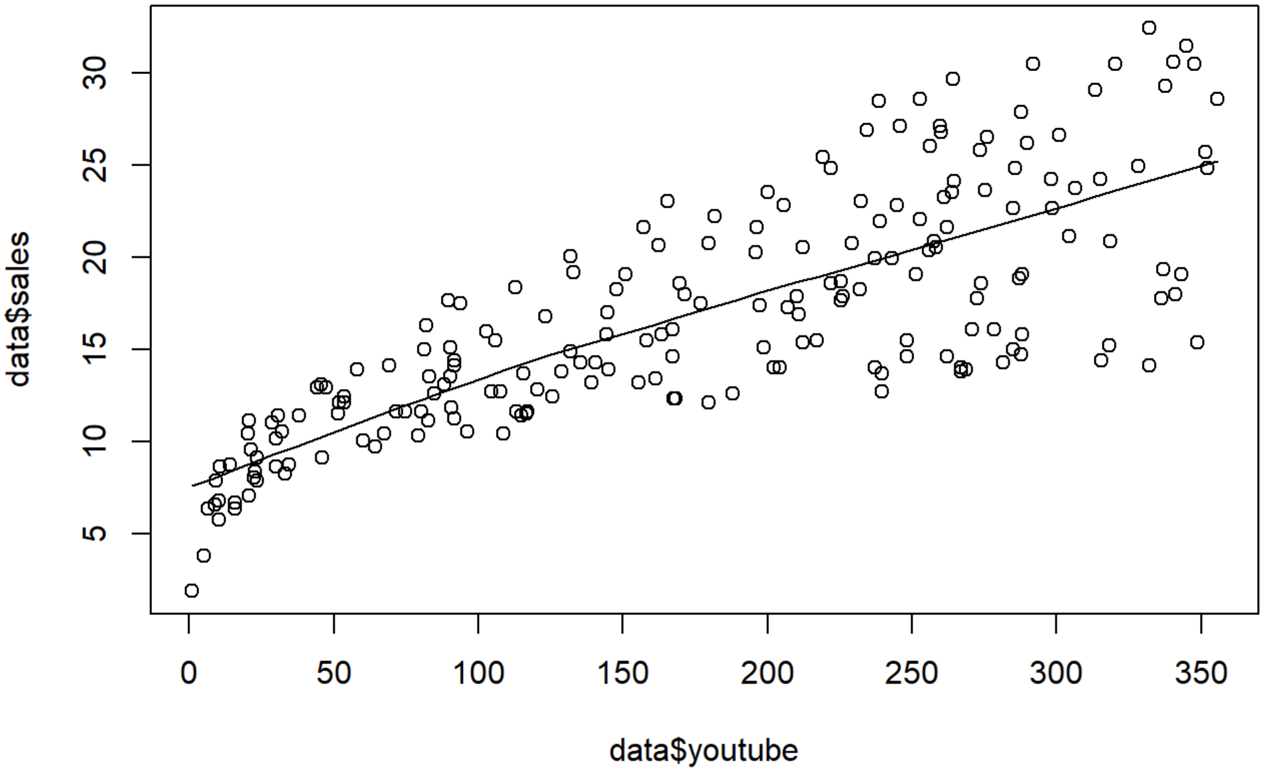
```
cor(data)
```

```
##              youtube   facebook  newspaper     sales
## youtube   1.00000000 0.05480866 0.05664787 0.7822244
## facebook  0.05480866 1.00000000 0.35410375 0.5762226
## newspaper 0.05664787 0.35410375 1.00000000 0.2282990
## sales     0.78222442 0.57622257 0.22829903 1.0000000
```

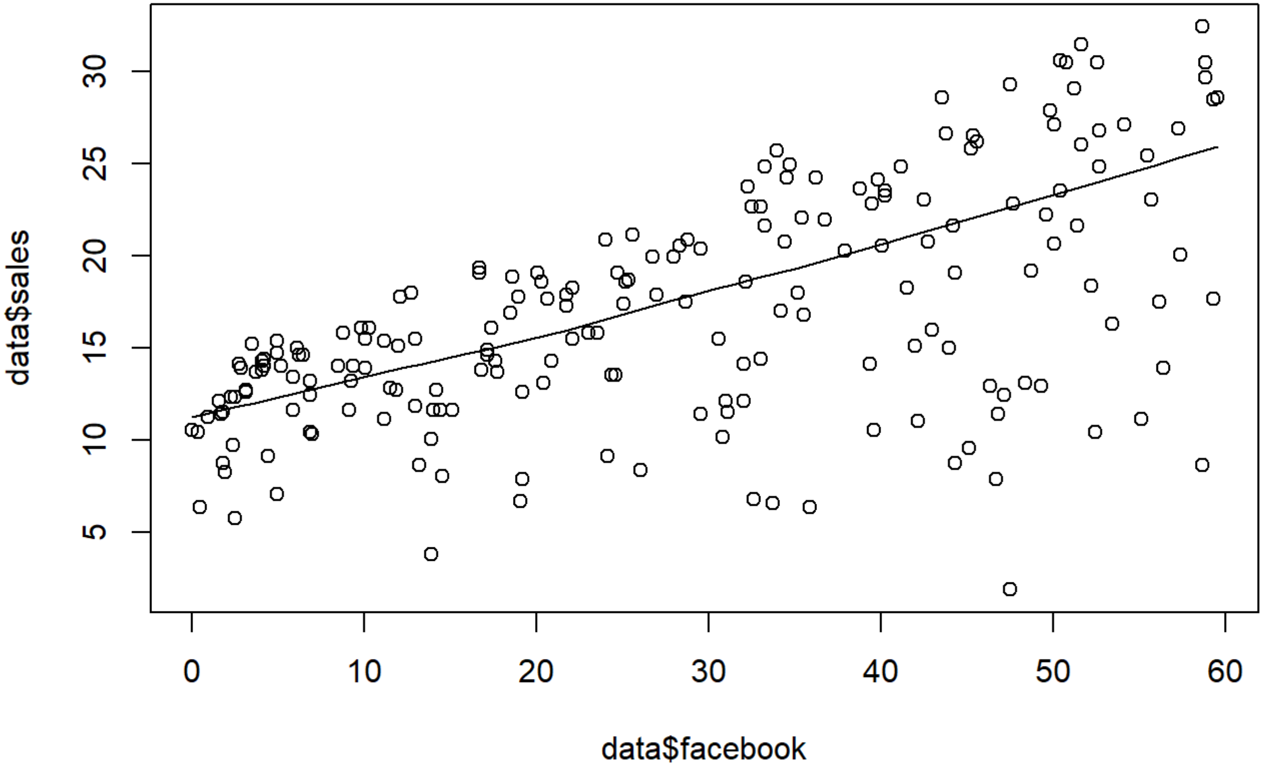
The cor-relation matrix shows that the variables facebook and youtube are highly correlated to the sales but newspaper is not too much correlated to the sales compared to others. Again the correlation between youtube and facebook is very less so they are likely to be linearly independent.

we can also see the following plots to clarify the linearity.

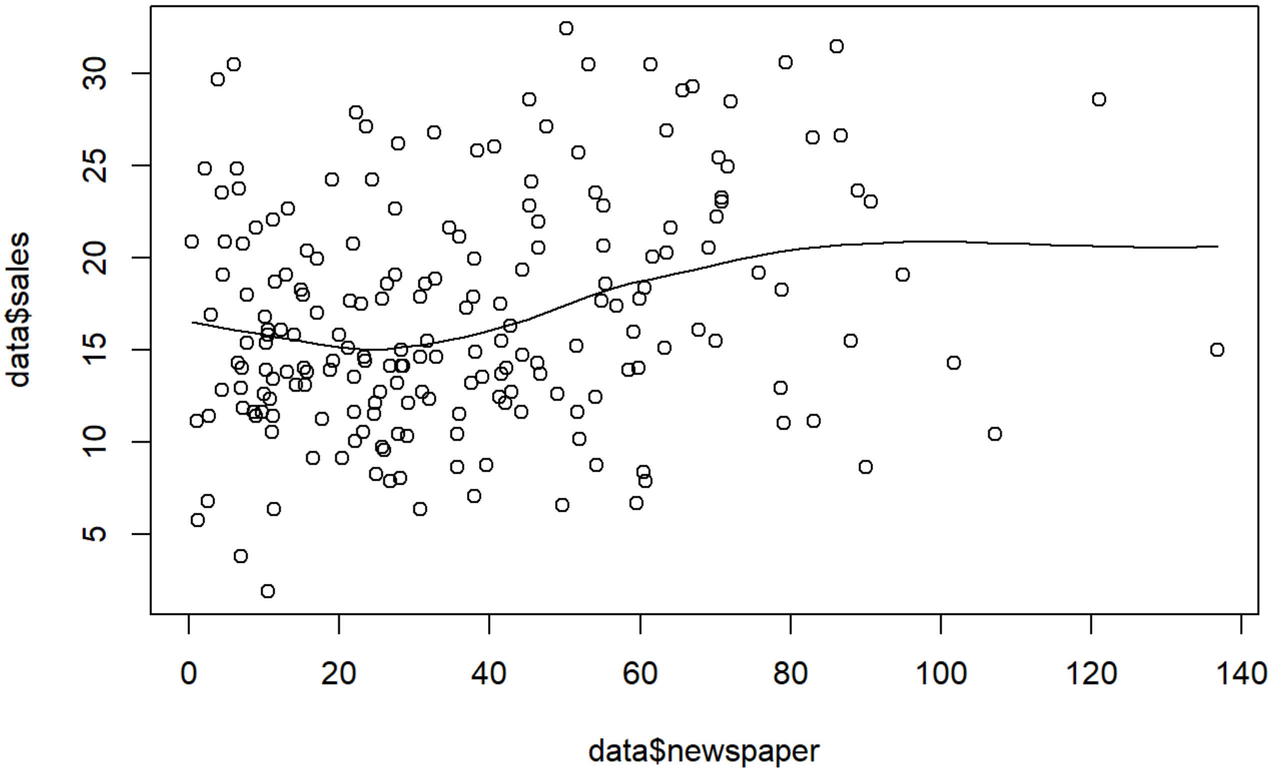
```
scatter.smooth(data$youtube,data$sales) # Plotting youtube vs sales
```



```
scatter.smooth(data$facebook,data$sales) #plotting facebook vs sales
```



```
scatter.smooth(data$newspaper,data$sales) # plotting newspaper vs sales
```



Multiple Linear Regression Model Building:

As we discussed above we will build the multiple linear regression model with sales as response variable and youtube and facebook are as predictors.

MODEL:

```
model=lm(sales~youtube+facebook,data=data)
summary(model)
```

```
## 
## Call:
## lm(formula = sales ~ youtube + facebook, data = data)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -10.5572  -1.0502   0.2906   1.4049   3.3994 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  3.50532    0.35339   9.919   <2e-16 ***
## youtube      0.04575    0.00139  32.909   <2e-16 ***
## facebook     0.18799    0.00804  23.382   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.018 on 197 degrees of freedom
## Multiple R-squared:  0.8972, Adjusted R-squared:  0.8962 
## F-statistic: 859.6 on 2 and 197 DF,  p-value: < 2.2e-16
```

Clearly we see that all the predictors are significant in our model and p-value: < 2.2e-16 indicates that our model is acceptable. The Adjusted R-squared: 0.8962 indicates that 89.62% of the variation in the response is explained by the predictors in our model. # ASSUMPTIONS CHECKING:

We will check the following assumptions:

1. Linearity relation between response and regressors/predictors.

2.Homogeneity of variance(Homoscedasticity) in residuals.

3.Normality of Residuals

4.independence of error terms(No auto correlation)

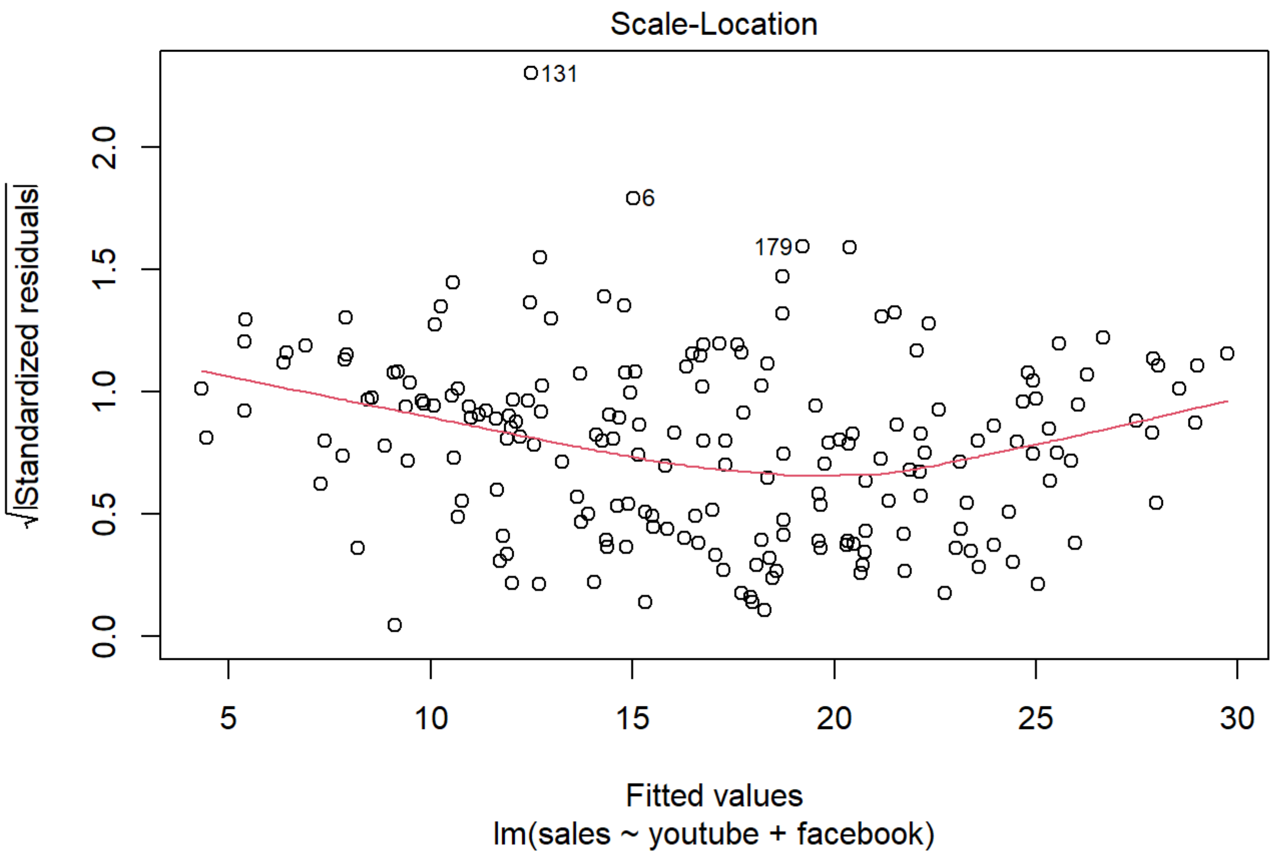
5.No multicolinearity.

Linearity:

We already check that there is relationship between sales and each predictors variable in our model and hence linearity assumption satisfied.

Homoscedasticity:

```
plot(model,3)
```



This plot shows that there is no constant variance in residuals that is, no homoscedasticity as the fitted line(red colored) in the above plot is not horizontal and to resolve this issue we need some transformation. Since the fitted red colored line looks like parabolic so we may use the square root or cube root or logarithmic transformation of response variable.

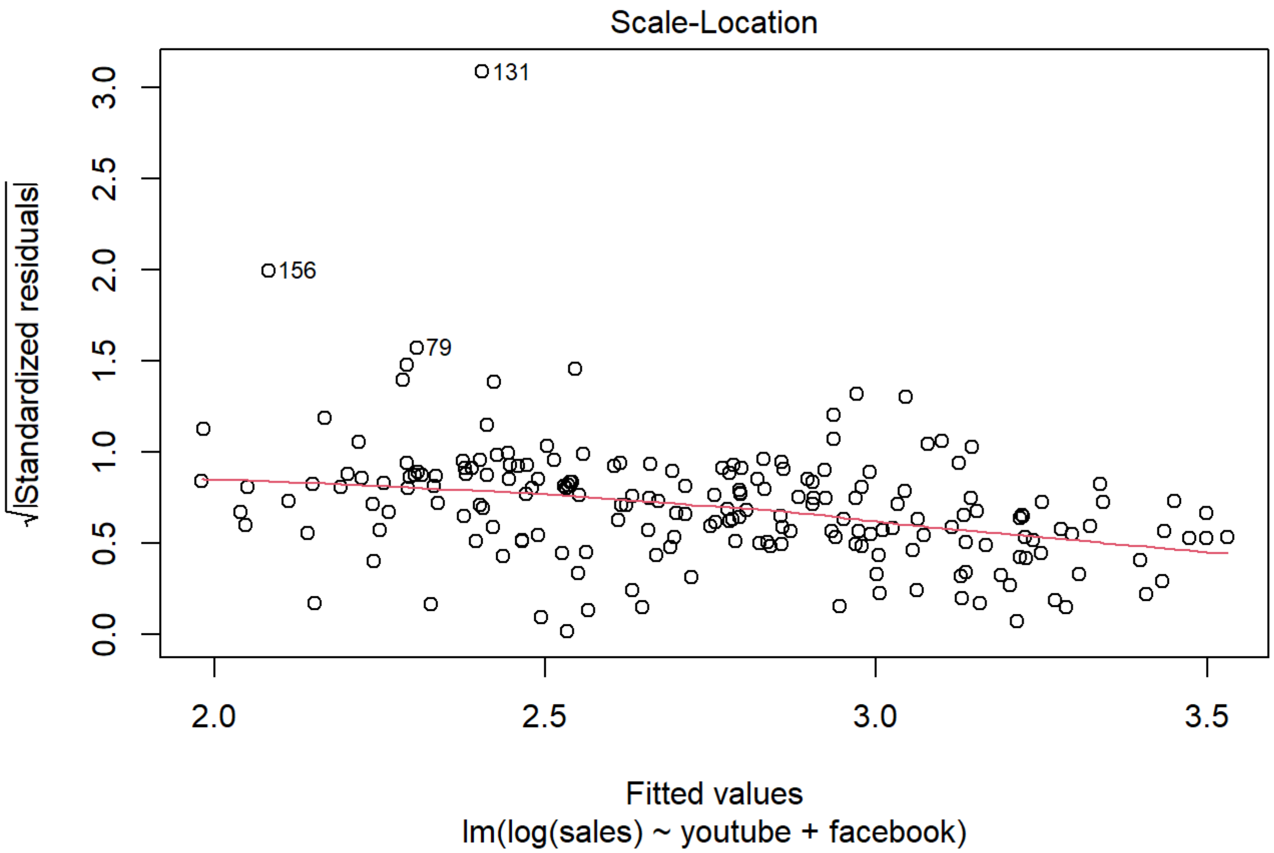
Now before checking others assumptions we reconstruct our model using logarithmic transformation of response variable.

```
new_model=lm(log(sales)~youtube+facebook,data=data)
summary(new_model)
```

```
##
## Call:
## lm(formula = log(sales) ~ youtube + facebook, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.75225 -0.05628  0.04626  0.10554  0.20542
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.9273997  0.0326640   59.01   <2e-16 ***
## youtube     0.0030609  0.0001285   23.82   <2e-16 ***
## facebook    0.0099874  0.0007431   13.44   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1865 on 197 degrees of freedom
## Multiple R-squared:  0.7995, Adjusted R-squared:  0.7974
## F-statistic: 392.7 on 2 and 197 DF,  p-value: < 2.2e-16
```

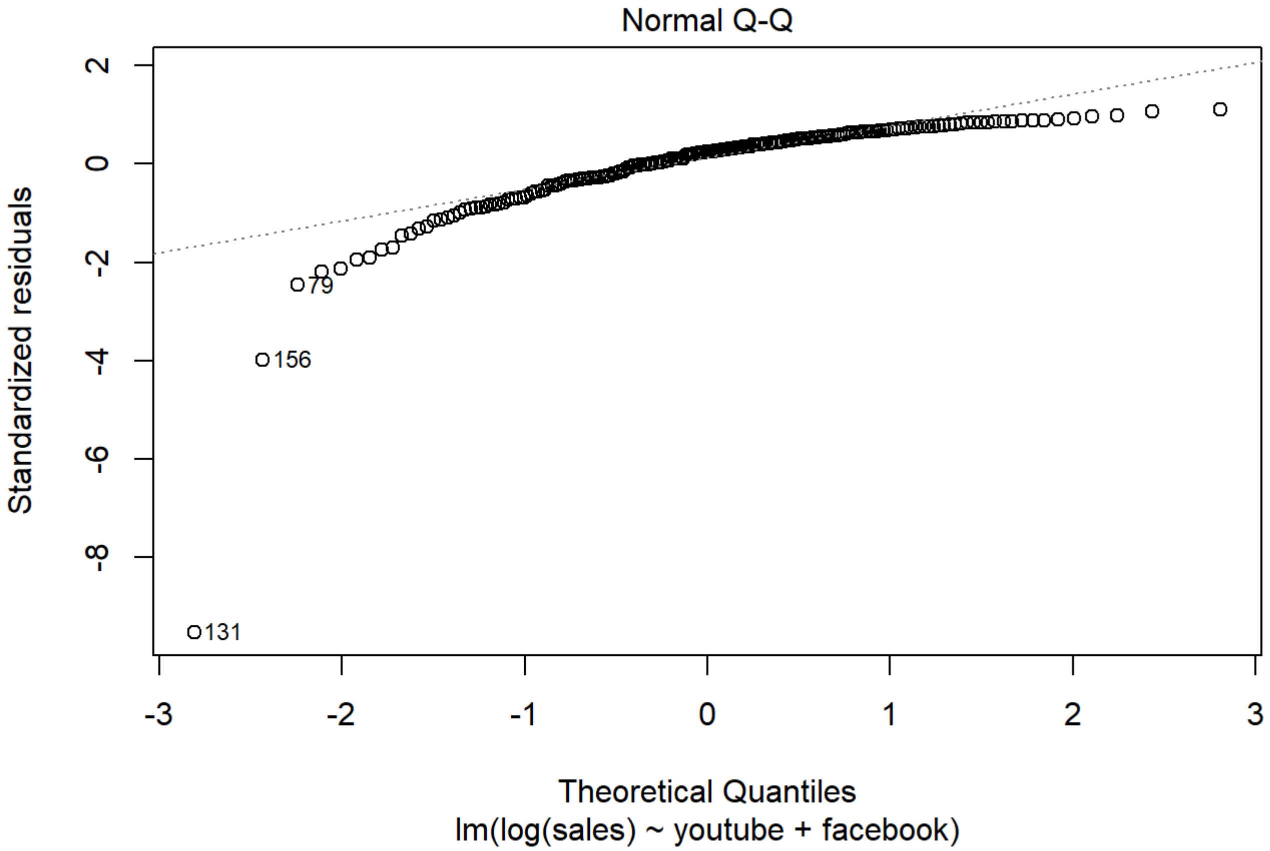
Assumptions checking on new model:

```
plot(new_model,3)
```



This plot ensure that our new model contains the homoscedasticity in residuals.

```
plot(new_model,2)
```



This Q-Q plot ensure the normality in residuals.

No Autocorrelation in Residuals:

```
library(car)
```

```
## Loading required package: carData
```

```
durbinWatsonTest(new_model)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1     -0.03011608      2.058211   0.726
##  Alternative hypothesis: rho != 0
```

The output of the test shows that p value is greater than 0.05 and hence accept the null hypothesis that is, rho=0 and there is no autocorrelation in residuals.

#Multicolinearity in predictors:

```
vif(new_model)
```

```
##  youtube facebook
## 1.003013 1.003013
```

The variance influence factors(VIF) of the predictors are less than 5 and very less and hence no multicolinearity presents in the predictors.

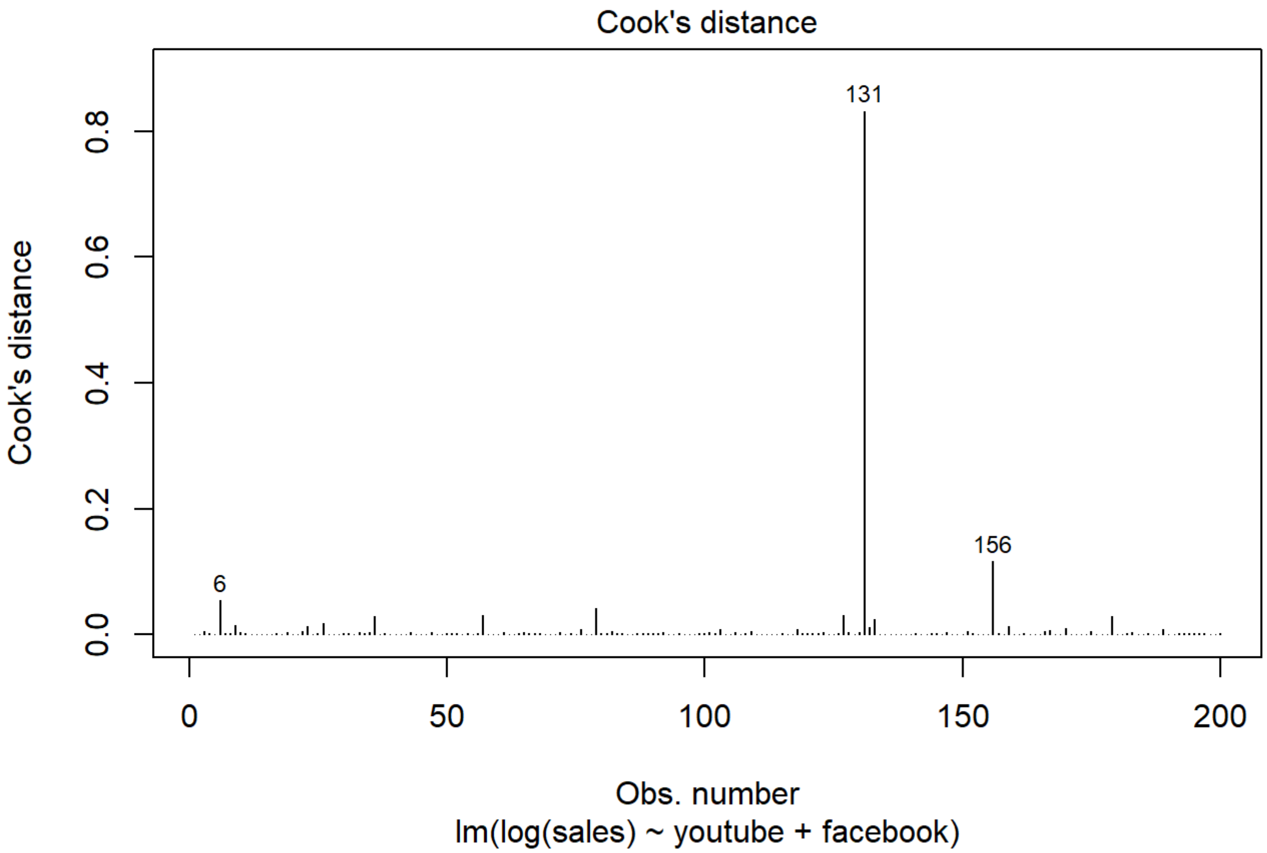
Outliers Test:

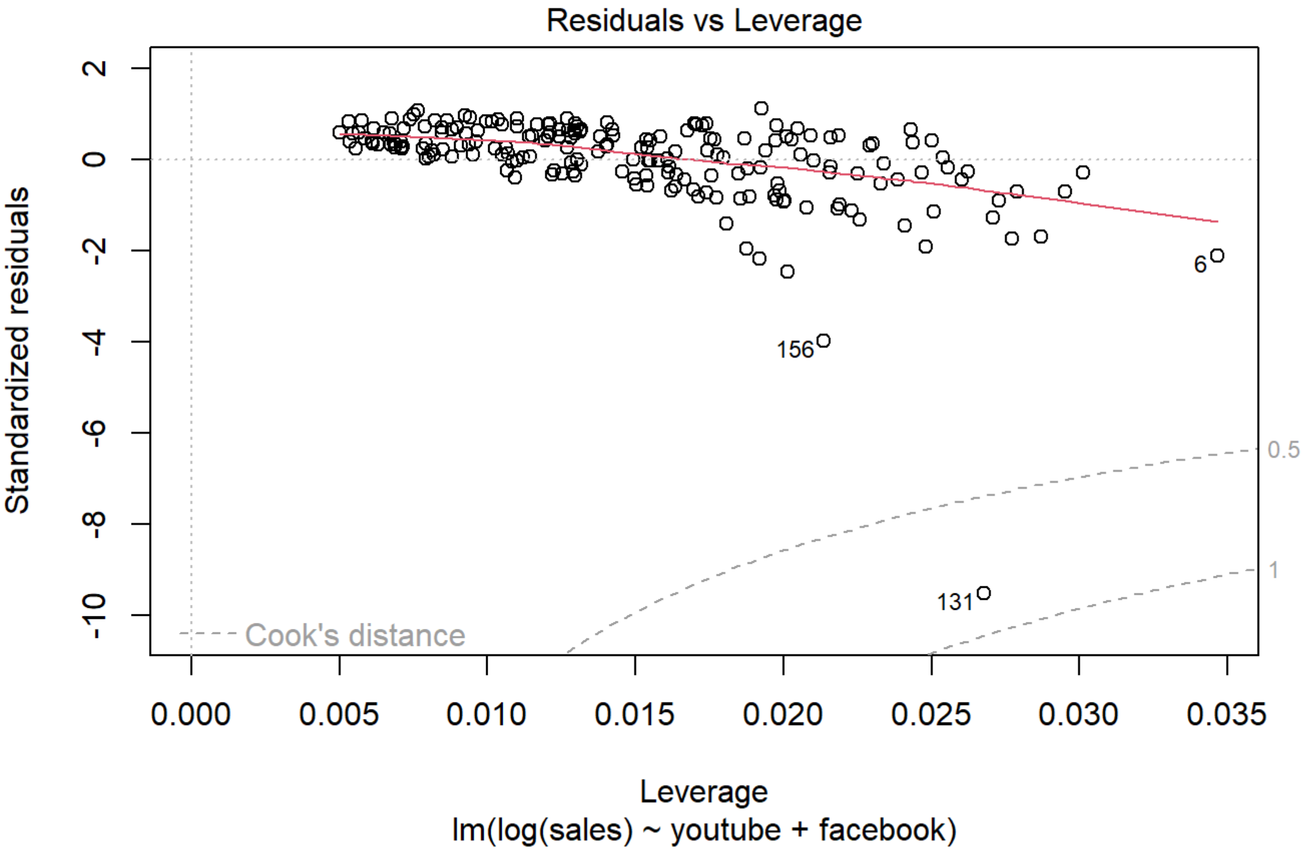
```
outlierTest(new_model)
```

```
##       rstudent unadjusted p-value Bonferroni p
## 131 -12.933454         4.5628e-28   9.1256e-26
## 156  -4.150502         4.9439e-05   9.8879e-03
```

The outliers test shows that 131th and 156 th observations results the larger error in our new model. Now we check which outliers are the influential observations. For this we consider the following plot.

```
plot(new_model,c(4,5))
```


Cook's distance
Cook's distance
0.0
0.2
0.4
0.6
0.8
131
156
6
0
50
100
150
200
Obs. number
lm(log(sales) ~ youtube + facebook)



we see that all the outliers has the cook’s distance inside 1 and hence we may consider that no influence observation there.

Prediction Model:

According to the new model(reconstruct), the prediction model is

log(sales^2)=1.9273997+(0.0030609)*youtube*+ *(0.0099874)*facebook.

Prediction Function:

```
predict=function(y,f){
  return(exp(1.9273997+(0.0030609)*y+ (0.0099874)*f))
}
```

Suppose we spent the money in youtube is $200 and facbook is $180 then the unit of sale is given by…

```
predict(200,180)
```

```
## [1] 76.5016
```

The unit of sale(in thousands of unit) is 76.5.

THANK YOU