Multiple linear Regression with Regression Diagnostics in R

Mohammad Abdul Wahed

2023-03-03

We'll use the marketing data set [datarium package], which contains the impact of the amount of money spent(in dollars) on three advertising medias (youtube, facebook and newspaper) on sales(units).

```
install.packages('datarium', repos = "http://cran.us.r-project.org")
```

```
## Installing package into 'C:/Users/abdul/AppData/Local/R/win-library/4.2'
## (as 'lib' is unspecified)
```

```
## package 'datarium' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\abdul\AppData\Local\Temp\Rtmpkpuo2L\downloaded_packages
```

```
require(datarium)
```

```
## Loading required package: datarium
```

```
## Warning: package 'datarium' was built under R version 4.2.2
```

Let's load the marketing dataset

```
d=marketing
d
```

```
##     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
## 7     69.00    39.36     28.20 14.16
## 8    144.24    23.52     13.92 15.84
## 9     10.32     2.52      1.20  5.76
## 10   239.76     3.12     25.44 12.72
## 11    79.32     6.96     29.04 10.32
## 12   257.64    28.80      4.80 20.88
## 13    28.56    42.12     79.08 11.04
## 14   117.00     9.12      8.64 11.64
## 15   244.92    39.48     55.20 22.80
## 16   234.48    57.24     63.48 26.88
## 17    81.36    43.92    136.80 15.00
## 18   337.68    47.52     66.96 29.28
## 19    83.04    24.60     21.96 13.56
## 20   176.76    28.68     22.92 17.52
## 21   262.08    33.24     64.08 21.60
## 22   284.88     6.12     28.20 15.00
## 23    15.84    19.08     59.52  6.72
## 24   273.96    20.28     31.44 18.60
## 25    74.76    15.12     21.96 11.64
## 26   315.48     4.20     23.40 14.40
## 27   171.48    35.16     15.12 18.00
## 28   288.12    20.04     27.48 19.08
## 29   298.56    32.52     27.48 22.68
## 30    84.72    19.20     48.96 12.60
## 31   351.48    33.96     51.84 25.68
## 32   135.48    20.88     46.32 14.28
## 33   116.64     1.80     36.00 11.52
## 34   318.72    24.00      0.36 20.88
## 35   114.84     1.68      8.88 11.40
## 36   348.84     4.92     10.20 15.36
## 37   320.28    52.56      6.00 30.48
## 38    89.64    59.28     54.84 17.64
## 39    51.72    32.04     42.12 12.12
## 40   273.60    45.24     38.40 25.80
## 41   243.00    26.76     37.92 19.92
## 42   212.40    40.08     46.44 20.52
## 43   352.32    33.24      2.16 24.84
## 44   248.28    10.08     31.68 15.48
## 45    30.12    30.84     51.96 10.20
## 46   210.12    27.00     37.80 17.88
## 47   107.64    11.88     42.84 12.72
## 48   287.88    49.80     22.20 27.84
## 49   272.64    18.96     59.88 17.76
## 50    80.28    14.04     44.16 11.64
## 51   239.76     3.72     41.52 13.68
## 52   120.48    11.52      4.32 12.84
## 53   259.68    50.04     47.52 27.12
## 54   219.12    55.44     70.44 25.44
## 55   315.24    34.56     19.08 24.24
## 56   238.68    59.28     72.00 28.44
## 57     8.76    33.72     49.68  6.60
## 58   163.44    23.04     19.92 15.84
## 59   252.96    59.52     45.24 28.56
## 60   252.84    35.40     11.16 22.08
## 61    64.20     2.40     25.68  9.72
## 62   313.56    51.24     65.64 29.04
## 63   287.16    18.60     32.76 18.84
## 64   123.24    35.52     10.08 16.80
## 65   157.32    51.36     34.68 21.60
## 66    82.80    11.16      1.08 11.16
## 67    37.80    29.52      2.64 11.40
## 68   167.16    17.40     12.24 16.08
## 69   284.88    33.00      13.20 22.68
## 70   260.16    52.68      32.64 26.76
## 71   238.92    36.72      46.44 21.96
## 72   131.76    17.16      38.04 14.88
## 73    32.16    39.60      23.16 10.56
## 74   155.28     6.84      37.56 13.20
## 75   256.08    29.52      15.72 20.40
## 76    20.28    52.44     107.28 10.44
## 77    33.00     1.92      24.84  8.28
## 78   144.60    34.20      17.04 17.04
## 79     6.48    35.88      11.28  6.36
## 80   139.20     9.24      27.72 13.20
## 81    91.68    32.04      26.76 14.16
## 82   287.76     4.92      44.28 14.76
## 83    90.36    24.36      39.00 13.56
## 84    82.08    53.40      42.72 16.32
## 85   256.20    51.60      40.56 26.04
## 86   231.84    22.08      78.84 18.24
## 87    91.56    33.00      19.20 14.40
## 88   132.84    48.72      75.84 19.20
## 89   105.96    30.60      88.08 15.48
## 90   131.76    57.36      61.68 20.04
## 91   161.16     5.88      11.16 13.44
## 92    34.32     1.80      39.60  8.76
## 93   261.24    40.20      70.80 23.28
## 94   301.08    43.80      86.76 26.64
## 95   128.88    16.80      13.08 13.80
## 96   195.96    37.92      63.48 20.28
## 97   237.12     4.20       7.08 14.04
## 98   221.88    25.20      26.40 18.60
## 99   347.64    50.76      61.44 30.48
## 100  162.24    50.04      55.08 20.64
## 101  266.88     5.16      59.76 14.04
## 102  355.68    43.56     121.08 28.56
## 103  336.24    12.12      25.68 17.76
## 104  225.48    20.64      21.48 17.64
## 105  285.84    41.16       6.36 24.84
## 106  165.48    55.68      70.80 23.04
## 107   30.00    13.20      35.64  8.64
## 108  108.48     0.36      27.84 10.44
## 109   15.72     0.48      30.72  6.36
## 110  306.48    32.28       6.60 23.76
## 111  270.96     9.84      67.80 16.08
## 112  290.04    45.60      27.84 26.16
## 113  210.84    18.48       2.88 16.92
## 114  251.52    24.72      12.84 19.08
## 115   93.84    56.16      41.40 17.52
## 116   90.12    42.00      63.24 15.12
## 117  167.04    17.16      30.72 14.64
## 118   91.68     0.96      17.76 11.28
## 119  150.84    44.28      95.04 19.08
## 120   23.28    19.20      26.76  7.92
## 121  169.56    32.16      55.44 18.60
## 122   22.56    26.04      60.48  8.40
## 123  268.80     2.88      18.72 13.92
## 124  147.72    41.52      14.88 18.24
## 125  275.40    38.76      89.04 23.64
## 126  104.64    14.16      31.08 12.72
## 127    9.36    46.68      60.72  7.92
## 128   96.24     0.00      11.04 10.56
## 129  264.36    58.80       3.84 29.64
## 130   71.52    14.40      51.72 11.64
## 131    0.84    47.52      10.44  1.92
## 132  318.24     3.48      51.60 15.24
## 133   10.08    32.64       2.52  6.84
## 134  263.76    40.20      54.12 23.52
## 135   44.28    46.32      78.72 12.96
## 136   57.96    56.40      10.20 13.92
## 137   30.72    46.80      11.16 11.40
## 138  328.44    34.68      71.64 24.96
## 139   51.60    31.08      24.60 11.52
## 140  221.88    52.68       2.04 24.84
## 141   88.08    20.40      15.48 13.08
## 142  232.44    42.48      90.72 23.04
## 143  264.60    39.84      45.48 24.12
## 144  125.52     6.84      41.28 12.48
## 145  115.44    17.76      46.68 13.68
## 146  168.36     2.28      10.80 12.36
## 147  288.12     8.76      10.44 15.84
## 148  291.84    58.80      53.16 30.48
## 149   45.60    48.36      14.28 13.08
## 150   53.64    30.96      24.72 12.12
## 151  336.84    16.68      44.40 19.32
## 152  145.20    10.08      58.44 13.92
## 153  237.12    27.96      17.04 19.92
## 154  205.56    47.64      45.24 22.80
## 155  225.36    25.32      11.40 18.72
## 156    4.92    13.92       6.84  3.84
## 157  112.68    52.20      60.60 18.36
## 158  179.76     1.56      29.16 12.12
## 159   14.04    44.28      54.24  8.76
## 160  158.04    22.08      41.52 15.48
## 161  207.00    21.72      36.84 17.28
## 162  102.84    42.96      59.16 15.96
## 163  226.08    21.72      30.72 17.88
## 164  196.20    44.16       8.88 21.60
## 165  140.64    17.64       6.48 14.28
## 166  281.40     4.08     101.76 14.28
## 167   21.48    45.12      25.92  9.60
## 168  248.16     6.24      23.28 14.64
## 169  258.48    28.32      69.12 20.52
## 170  341.16    12.72       7.68 18.00
## 171   60.00    13.92      22.08 10.08
## 172  197.40    25.08      56.88 17.40
## 173   23.52    24.12      20.40  9.12
## 174  202.08     8.52      15.36 14.04
## 175  266.88     4.08      15.72 13.80
## 176  332.28    58.68      50.16 32.40
## 177  298.08    36.24     24.36 24.24
## 178  204.24     9.36     42.24 14.04
## 179  332.04     2.76     28.44 14.16
## 180  198.72    12.00     21.12 15.12
## 181  187.92     3.12      9.96 12.60
## 182  262.20     6.48     32.88 14.64
## 183   67.44     6.84     35.64 10.44
## 184  345.12    51.60     86.16 31.44
## 185  304.56    25.56     36.00 21.12
## 186  246.00    54.12     23.52 27.12
## 187  167.40     2.52     31.92 12.36
## 188  229.32    34.44     21.84 20.76
## 189  343.20    16.68      4.44 19.08
## 190   22.44    14.52     28.08  8.04
## 191   47.40    49.32      6.96 12.96
## 192   90.60    12.96      7.20 11.88
## 193   20.64     4.92     37.92  7.08
## 194  200.16    50.40      4.32 23.52
## 195  179.64    42.72      7.20 20.76
## 196   45.84     4.44     16.56  9.12
## 197  113.04     5.88      9.72 11.64
## 198  212.40    11.16      7.68 15.36
## 199  340.32    50.40     79.44 30.60
## 200  278.52    10.32     10.44 16.08
```

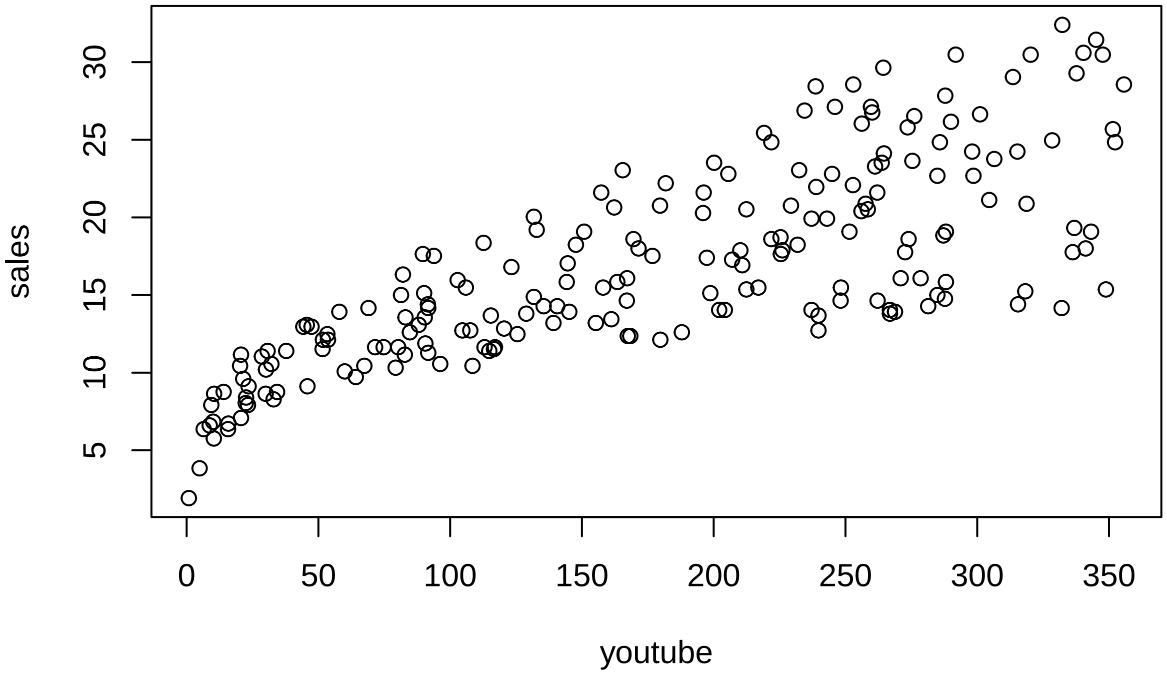
```
attach(d)
```

EXPLORATORY DATA ANALYSIS
Let's perform EDA before we start to model

Let's see how sales varies across amount of money spent on three advertising medias(youtube, facebook and newspaper)

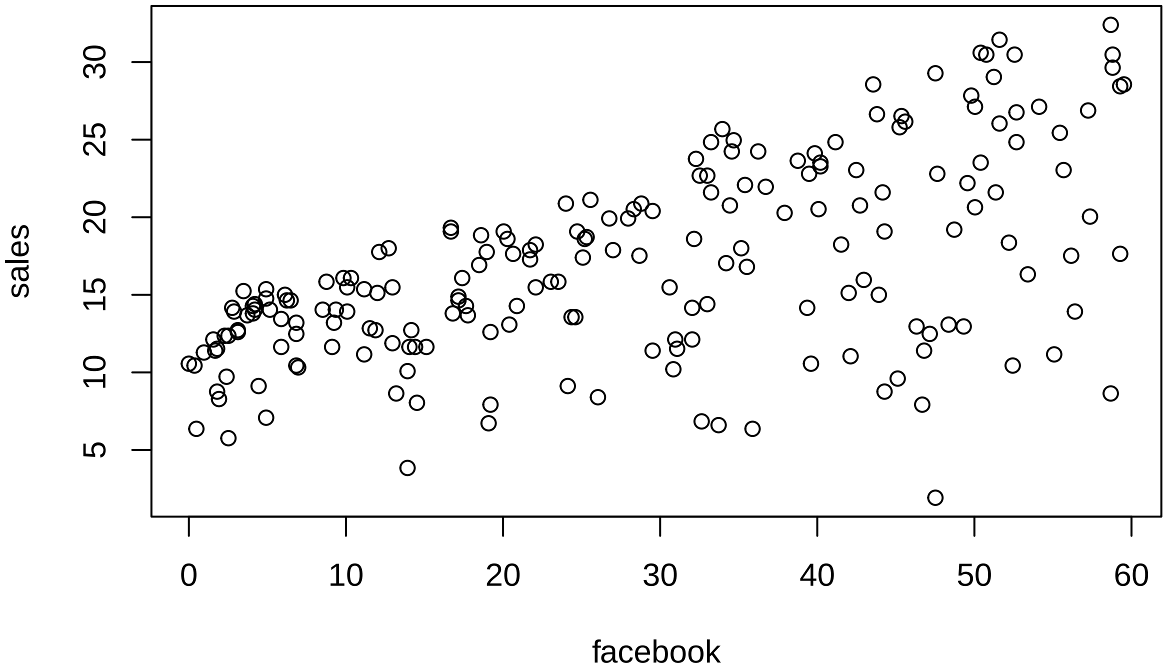
For that we'll use a scatterplot (sales vs money spent on different advertising medias)

```
plot(sales~youtube)
```



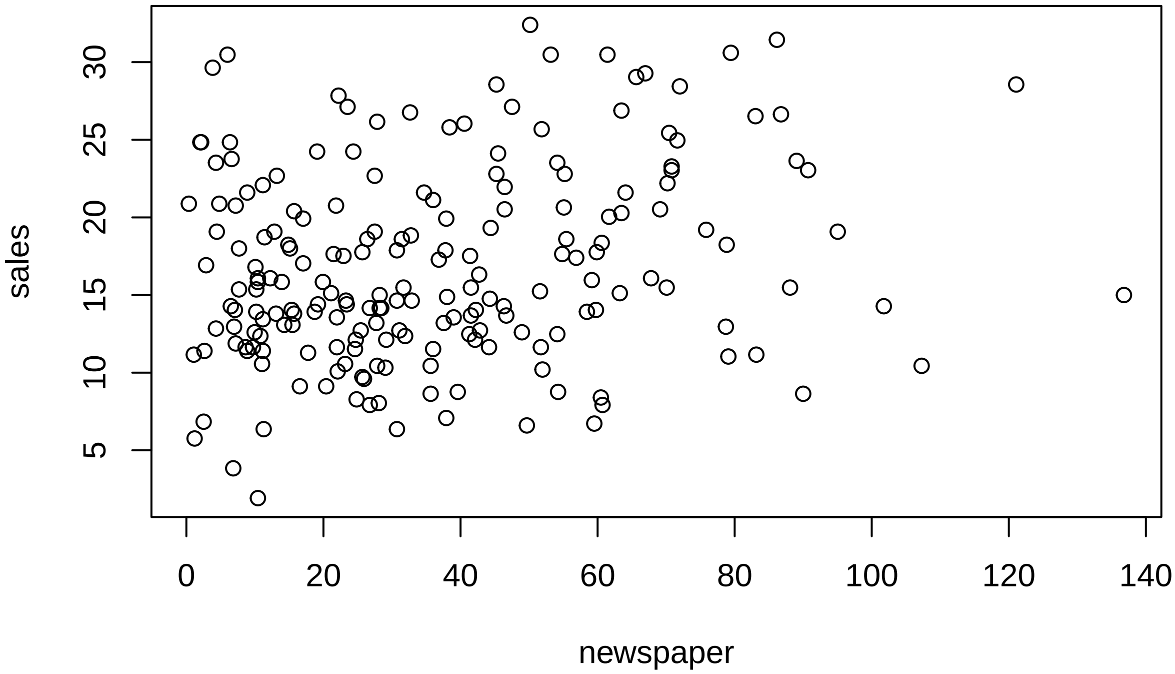
From the plot, we can observe that as the amount of dollars spent on youtube advertising increases the units sold also increases

```
plot(sales~facebook)
```



From the plot, we can again observe that as the amount of dollars spent on facebook advertising increases the units sold also increases

```
plot(sales~newspaper)
```



From the plot, we can observe that as the amount of dollars spent on newspaper advertising has no such relation to units sold

BUILDING A MODEL

```
model=lm(sales~youtube+facebook+newspaper)
model
```

```
##
## Call:
## lm(formula = sales ~ youtube + facebook + newspaper)
##
## Coefficients:
## (Intercept)      youtube     facebook    newspaper
##    3.526667     0.045765     0.188530    -0.001037
```

sales = 3.527 + 0.046 * youtube + 0.188 * facebook - 0.001 * newspaper

```
summary(model)
```

```
##
## Call:
## lm(formula = sales ~ youtube + facebook + newspaper)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.5932  -1.0690   0.2902   1.4272   3.3951
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.526667   0.374290   9.422   <2e-16 ***
## youtube      0.045765   0.001395  32.809   <2e-16 ***
## facebook     0.188530   0.008611  21.893   <2e-16 ***
## newspaper   -0.001037   0.005871  -0.177     0.86
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.023 on 196 degrees of freedom
## Multiple R-squared:  0.8972, Adjusted R-squared:  0.8956
## F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

Interpretation The first step in interpreting the multiple regression analysis is to examine the F-statistic and the associated p-value, at the bottom of model summary. In our example, it can be seen that p-value of the F-statistic is < 2.2e-16, which is highly significant. This means that, at least, one of the predictor variables is significantly related to the outcome variable. To see which predictor variables are significant, you can examine the coefficients table, which shows the estimate of regression beta coefficients and the associated t-statistic p-values:

```
summary(model)$coefficient
```

```
##                 Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept)  3.526667243 0.374289884  9.4222884 1.267295e-17
## youtube      0.045764645 0.001394897 32.8086244 1.509960e-81
## facebook     0.188530017 0.008611234 21.8934961 1.505339e-54
## newspaper   -0.001037493 0.005871010 -0.1767146 8.599151e-01
```

For a given the predictor, the t-statistic evaluates whether or not there is significant association between the predictor and the outcome variable, that is whether the beta coefficient of the predictor is significantly different from zero.

It can be seen that, changing in youtube and facebook advertising budget are significantly associated to changes in sales while changes in newspaper budget is not significantly associated with sales.

For a given predictor variable, the coefficient (b) can be interpreted as the average effect on y of a one unit increase in predictor, holding all other predictors fixed.

For example, for a fixed amount of youtube and newspaper advertising budget, spending an additional 1000 dollars on facebook advertising leads to an increase in sales by approximately 0.1885 * 1000 = 189 sale units, on average.

The youtube coefficient suggests that for every 1000 dollars increase in youtube advertising budget, holding all other predictors constant, we can expect an increase of 0.045 * 1000 = 45 sales units, on average.

We found that newspaper is not significant in the multiple regression model. This means that, for a fixed amount of youtube and newspaper advertising budget, changes in the newspaper advertising budget will not significantly affect sales units.

As the newspaper variable is not significant, it is possible to remove it from the model:

```
model<-lm(sales~youtube+facebook)
summary(model)
```

```
##
## Call:
## lm(formula = sales ~ youtube + facebook)
##
## 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
```

Our model equation can be written as follow:

sales = 3.5 + 0.045 * youtube + 0.187 * facebook

MODEL ACCURACY ASSESSMENT

The overall quality of the model can be assessed by examining the R-squared or Adjusted-R-Squared

In multiple linear regression, the R-Squared represents the correlation coefficient between the observed values of the outcome variable (y) and the fitted (i.e., predicted) values of y. For this reason, the value of R will always be positive and will range from zero to one.

R-Squared represents the proportion of variance, in the outcome variable y, that may be predicted by knowing the value of the x variables. An R-Squared value close to 1 indicates that the model explains a large portion of the variance in the outcome variable.

A problem with the R-Squared, is that, it will always increase when more variables are added to the model, even if those variables are only weakly associated with the response (James et al. 2014). A solution is to adjust the R-Squared by taking into account the number of predictor variables. The adjustment in the "Adjusted R Square" value in the summary output is a correction for the number of x variables included in the prediction model.

VERIFYING IF ASSUMPTIONS OF ORDINARY LEAST SQUARES REGRESSION ARE MET

The assumptions are:

1.) Linearity: The specified model must represent a linear relationship

2.) Homogeneity of variance(homoscedasticity): The error variance should be constant

3.) No autocorrelation: No identifiable relationship should exist between the values of the error term.

4.) Normality: The errors should be normally distributed. Technically normality is necessary only for hypothesis tests to be valid.

5.) No multicollinearity: No predictor variable should be perfectly (or almost perfectly) explained by the other predictors.

6.) No outliers: A single observation that is substantially different from all other observations can make a large difference in the results of your regression analysis.

7.) No endogenity: The independent variables shouldn't be correlated with the error term.

We verify if the data have met the underlying ordinary least squares regression assumptions

1.) Linearity

Checking Linearity To check linearity residuals should be plotted against the fit as well as other predictors. If any of these plots show systematic shapes, then the linear model is not appropriate and some nonlinear terms may need to be added. In package car, function residualPlots() produces those plots.

Installing package car

```
install.packages('car', repos = "http://cran.us.r-project.org")
```

```
## Installing package into 'C:/Users/abdul/AppData/Local/R/win-library/4.2'
## (as 'lib' is unspecified)
```

```
## package 'car' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\abdul\AppData\Local\Temp\Rtmpkpuo2L\downloaded_packages
```

#residual vs. fitted value and all predictors

```
require(car)
```

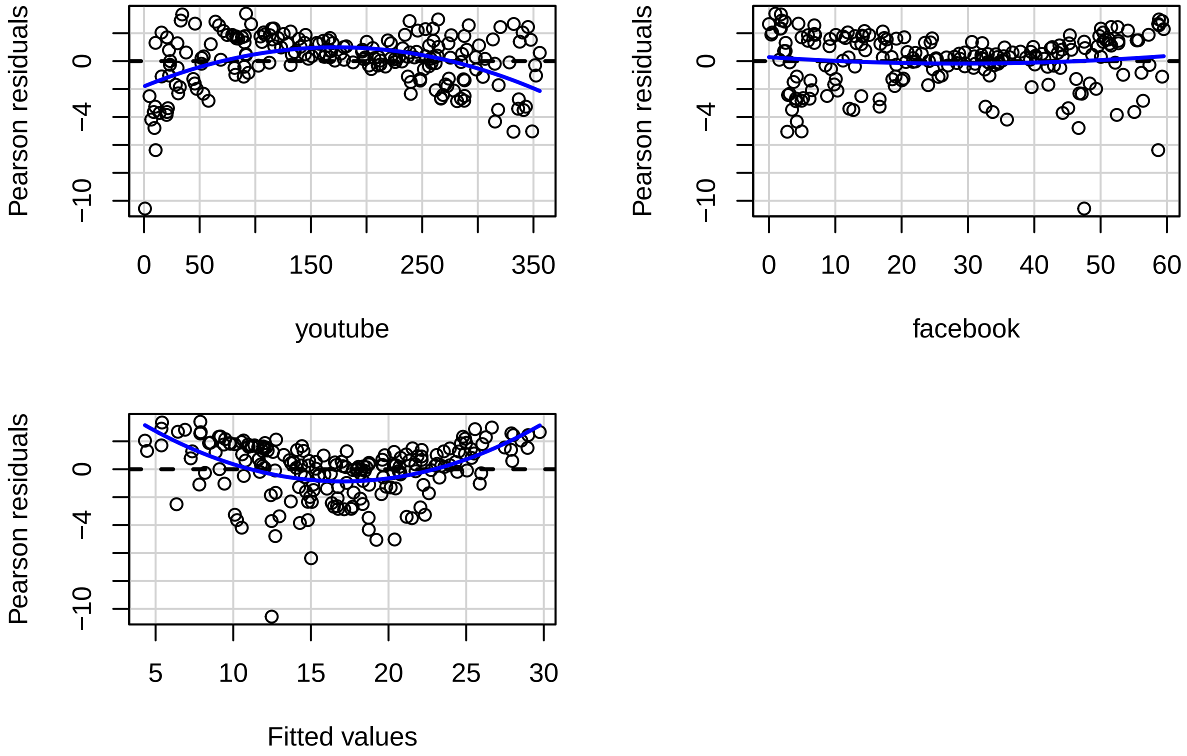
```
## Loading required package: car
```

```
## Warning: package 'car' was built under R version 4.2.2
```

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

```
## Warning: package 'carData' was built under R version 4.2.2
```

```
residualPlots(model)
```



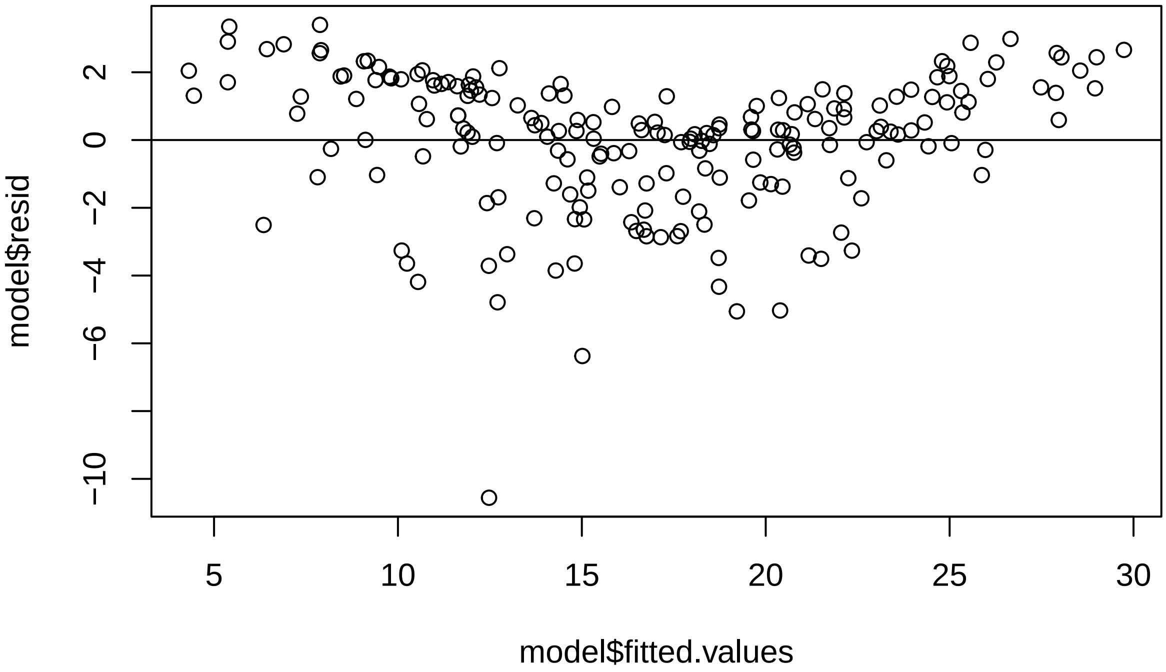
```
##             Test stat Pr(>|Test stat|)
## youtube      -6.7745        1.423e-10 ***
## facebook      1.0543           0.2931
## Tukey test    7.6351        2.256e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

To find out whether the 1-on-1 relationships are linear, you need to judge whether the data points are more or less on or around a straight line. We observe not much deviation from linearity.

2.) Homoscedasticity

One of the main assumptions for the ordinary least squares regression is the homogeneity of variance of the residuals. If the model is well-fitted, there should be no pattern to the residuals plotted against the fitted values. If the variance of the residuals is non-constant then the residual variance is said to be "heteroscedastic." There are graphical and non-graphical methods for detecting heteroscedasticity. A commonly used graphical method is to plot the residuals versus fitted (predicted) values

```
plot(model$resid~model$fitted.values)
abline(h=0)
```



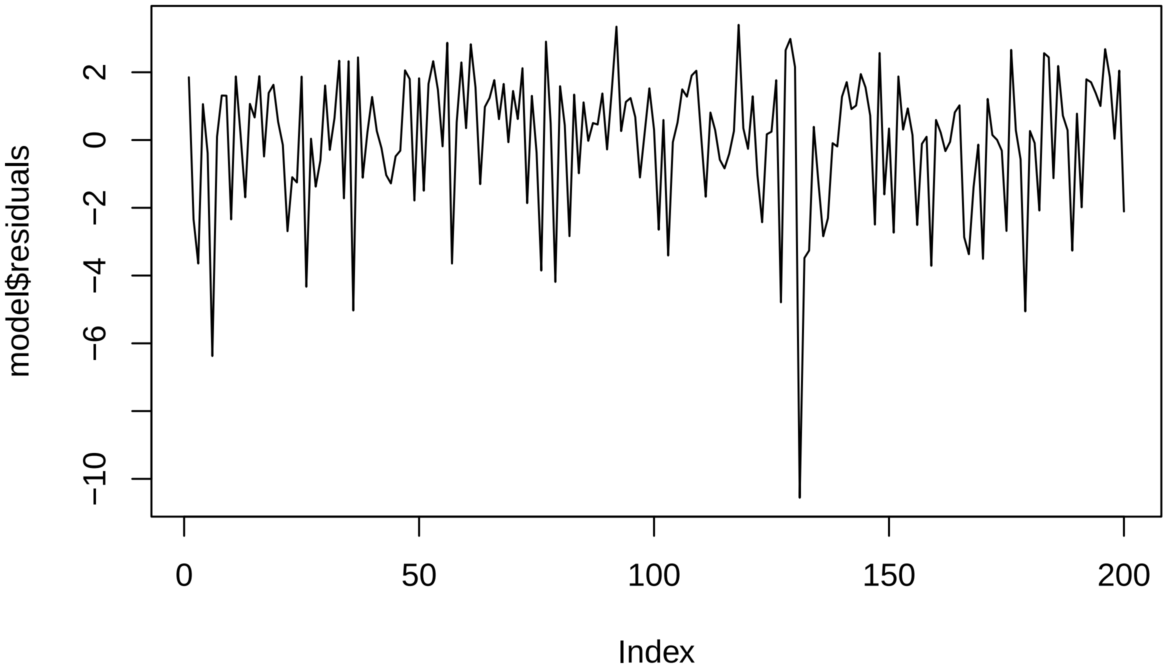
Homoscedasticity means a constant error, you are looking for a constant deviation of the points from the zero-line. Except one outlier at the bottom left, the deviation of points from the zero line is constant

3.) No autocorrelation

Observations of the error term are uncorrelated with each other

You can do a visual check by plotting the residuals against the order of the residuals. The following code snippet allow you to do this:

```
plot(model$residuals, type = 'l')
```

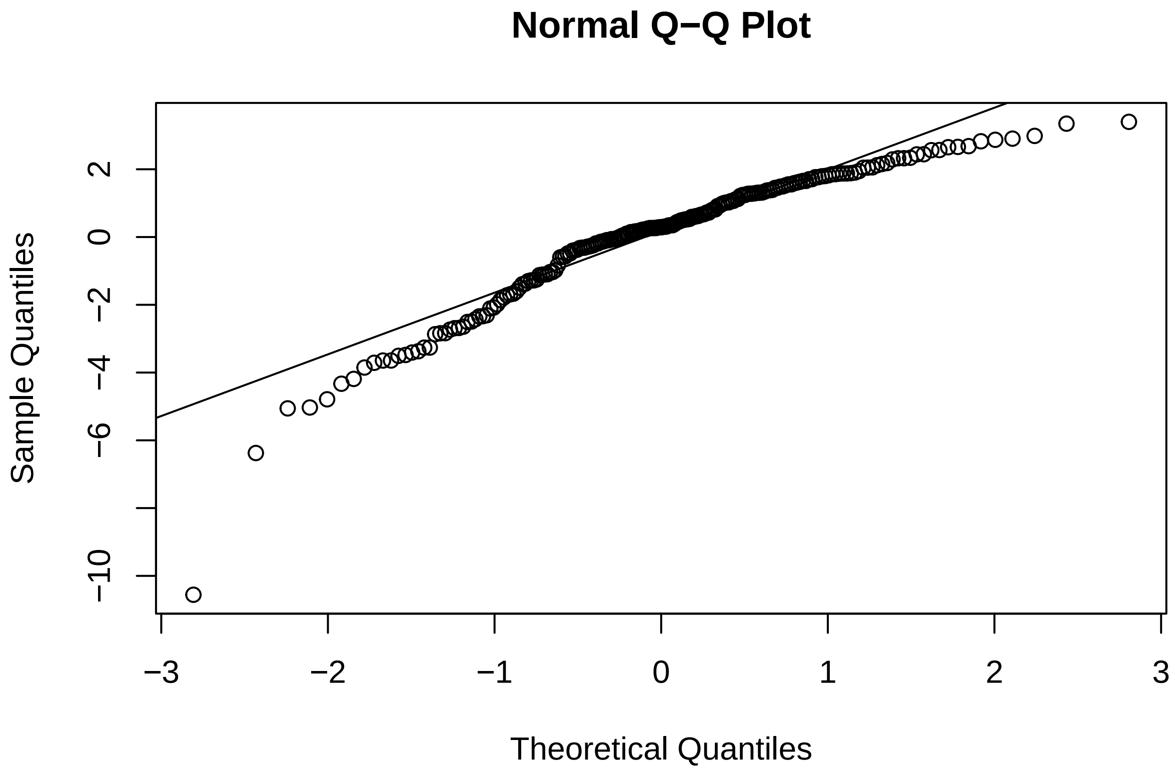


If a pattern occurs, it is likely that you have a case of a misspecified model.

4.) Normality

Once you obtain the residuals from your model, this is relatively easy to test using either a QQ Plot. Let's see how to make QQ Plots of the residuals using R

```
qqnorm(model$residuals)
abline(qqline(model$residuals))
```



What you need to look at in QQ Plots is whether the points are on the straight line going from bottom left to top right. We observe not much deviance from normality

5.) No multicollinearity

Multicollinearity is the phenomenon when a number of the explanatory variables are strongly correlated. You can test for multicollinearity problems using the Variance Inflation Factor, or VIF in short. The VIF indicates for an independent variable how much it is correlated to the other independent variables. You can compute VIF in R with the following code.

```
library(car)
car::vif(model)
```

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

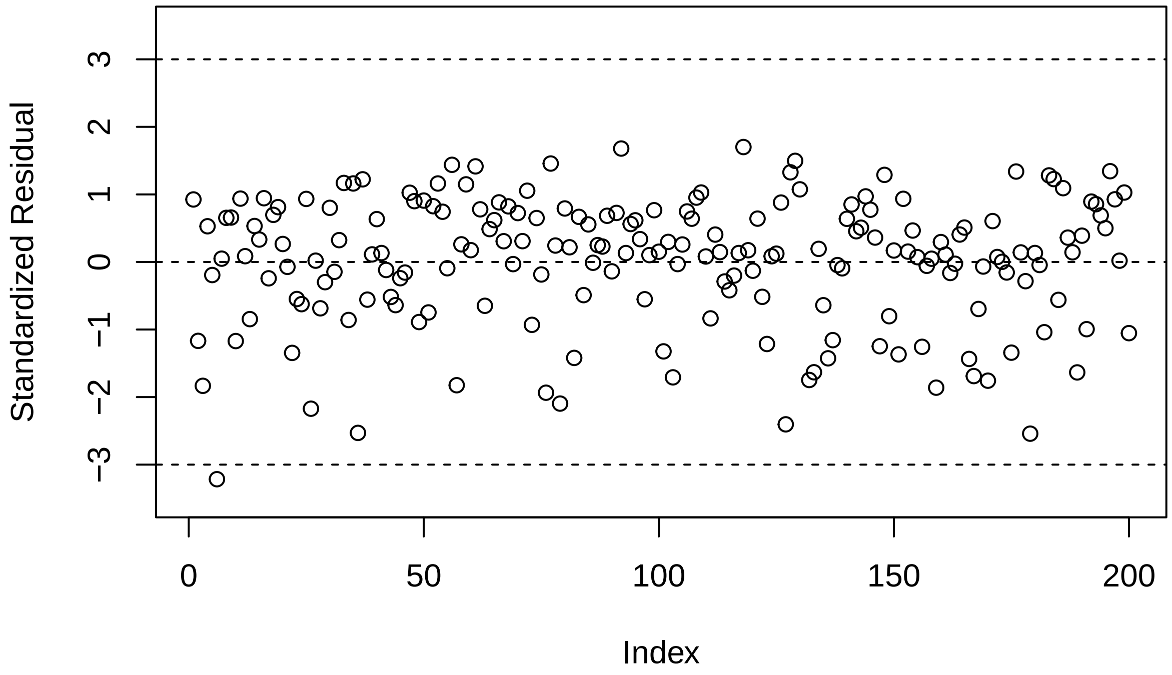
VIF starts from 1 and has no upper limit. A VIF of 1 is the best you can have as this indicates that there is no multicollinearity for this variable. A VIF of higher than 5 or 10 indicates that there is a problem with the independent variables in your model.

In the current model, there seems to be no multicollinearity.

6.) No outliers

Studentized residuals can be used to identify outliers. In R we use rstandard() function to compute Studentized residuals.

```
res.std <- rstandard(model) #studentized residuals stored in vector res.std
#plot Standardized residual in y axis. X axis will be the index or row names
plot(res.std, ylab="Standardized Residual", ylim=c(-3.5,3.5))
#add horizontal lines 3 and -3 to identify extreme values
abline(h =c(-3,0,3), lty = 2)
```



We should pay attention to studentized residuals that exceed +2 or -2, and get even more concerned about residuals that exceed +2.5 or -2.5 and even yet more concerned about residuals that exceed +3 or -3.

```
#find out which data point is outside of 3 standard deviation cut-off
#index is row numbers of those point
index <- which(res.std > 3 | res.std < -3)
```

```
#print row number of values that are out of bounds
print(index)
```

```
##   6 131
##   6 131
```

Row number 6 and 131 are to be deleted for a more robust model

```
d = d[-c(6, 131),]
```

Plotting studentized residuals again

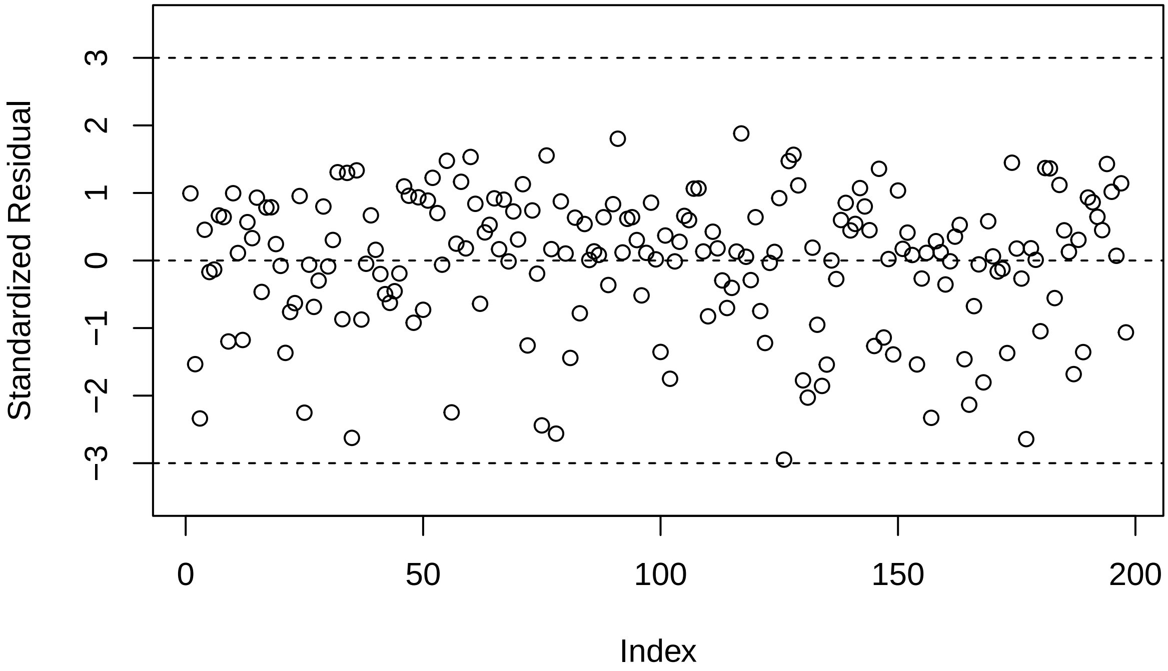
```r
attach(d)
```

```
## The following objects are masked from d (pos = 5):
##
##     facebook, newspaper, sales, youtube
```

```r
model=lm(sales~youtube+facebook)
model
```

```
##
## Call:
## lm(formula = sales ~ youtube + facebook)
##
## Coefficients:
## (Intercept)      youtube     facebook
##     3.66229      0.04422      0.19529
```

```r
res.std <- rstandard(model) #studentized residuals stored in vector res.std
#plot Standardized residual in y axis. X axis will be the index or row names
plot(res.std, ylab="Standardized Residual", ylim=c(-3.5,3.5))
#add horizontal lines 3 and -3 to identify extreme values
abline(h =c(-3,0,3), lty = 2)
```



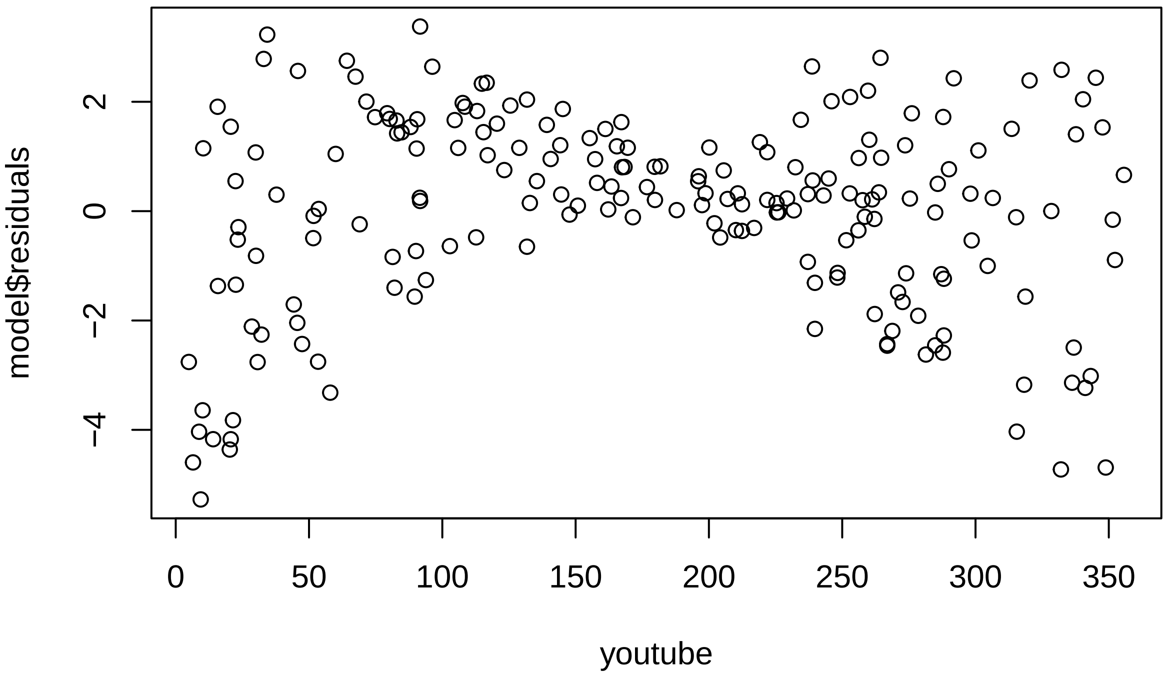
As we see no residuals are exceeding +3 or -3 in the studentized residuals plot We have successfully removed outliers from our data

7.) No endogenity

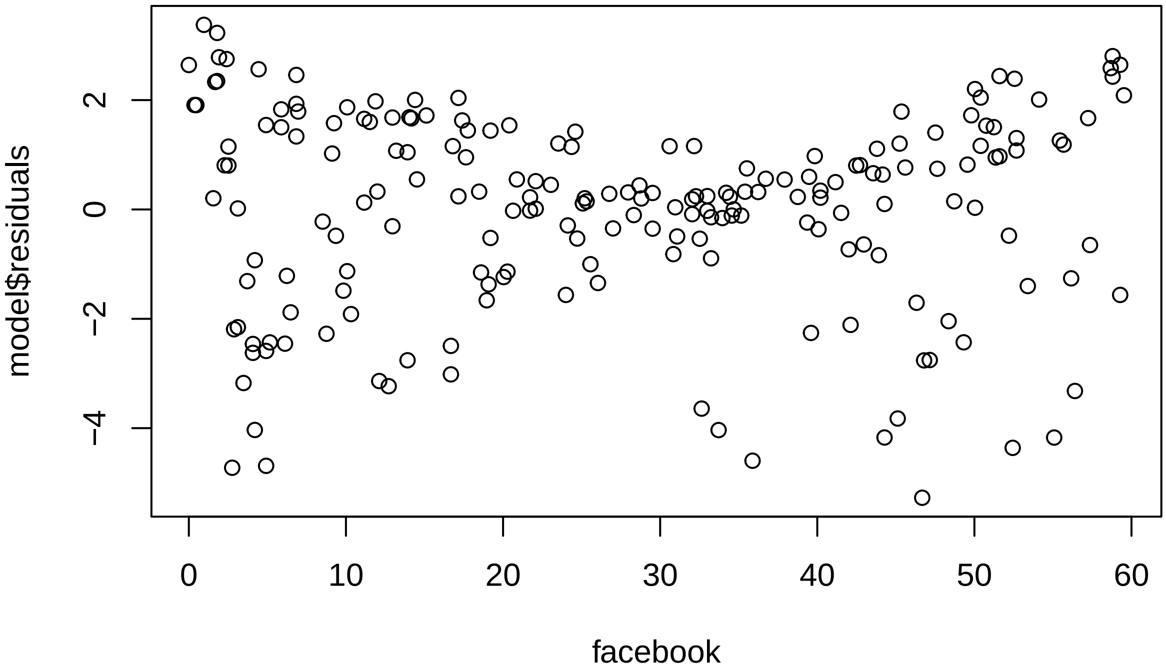
All independent variables are uncorrelated with the error term The seventh diagnostical check of our linear regression model serves to check whether there is correlation between any of the independent variables and the error term. If this happens, it is likely that you have a case of a misspecified model. You may have forgotten an important explanatory variable.

You can obtain the scatter plots using the following R code:

```
plot(youtube, model$residuals)
```



```
plot(facebook, model$residuals)
```



In these scatter plots, we do not see any clear correlation.

CONCLUSIONS

Finally, our model equation can be written as follows:

sales = 3.66 + 0.044 * youtube + 0.195 * facebook