Linear Regression Analysis

B.M Njuguna

2022-09-01

Contents

1. Regression Analysis

Regression analysis is a set of statistical methods used to identify or estimate the relationship(s) between the **dependent** and **independent** variable(s). It can also be utilized to assess the strength of the relationship and also to model the future relationship between the variables. The dependent variable which is also known as **response variable** is the variable being tested or measured in an experiment, while the independent or the **explanatory or predictor variable** is the variable which is included to the model to explain changes in the dependent variable. In most cases, the dependent variable is denoted by y while the independent variable is usually denoted by x.

1.1 Types of Regression Analysis

There are several types of regression analysis depending on what you want to achieve, or depending on the nature of the study or the nature of the variables. They include;[1]

1. Linear Regression
2. Logistic Regression
3. Polynomial Regression
4. Ridge Regression
5. Quantile Regression
6. Bayesian Linear Regression
7. Principal Component Regression
8. Partial Least Square Regression amongst other types.

2. Linear Regression

A linear regression is a regression model that estimates the relationship between the dependent and the independent variables using a straight line. A linear regression model is as follows;

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \epsilon_i$$

Where

- y_i is the response variable.
- β_k is the k^{th} coefficient, where β_0 is the constant term in the model.
- X_{ij} is the i^{th} observation on the j^{th} predictor variable, $j = 1, ..., p$.
- ϵ_i is the i^{th} noise term, that is, random error.

If the model includes one predictor variable, that is p=1, then the model is known as a simple linear regression model.

2.1 Simple Linear Regression model

A simple linear regression model is of the form;

$$y = \beta_o + \beta_1 x$$

where;

y - is the response variable

β_0 - is the intercept. It refers to the value of y when $x = 0$.

[1]This paper was compiled by Brian Mwangi Njuguna on 22-08-2022, for acadameic purposes

β_1 - is the regression coefficient or the slope. It represents the change in variable y caused by a unit change in the explanatory variable x.

It is used to model the relationship between two continuous variables. The assumptions are;

1. Linearity- The variables x and y must have a linear relationship
2. The error terms ϵ_i are independent and that the error is normally distributed with mean 0 and variance σ^2. That is; $\epsilon_i \sim N(0, \sigma^2)$

For example, We may wish to determine whether advertisement and sales have a linear relationship. Below is a data set containing the budget of advertisement in various platforms including TV,Radios and Newspaper as well as the sales, in 1000$.

```
> ## Importing the data set from my library
> library(readxl)
> AdvertisingBudgetandSales <- read_excel("AdvertisingBudgetandSales.xlsx", col_
    types = c("skip",
+     "numeric", "numeric", "numeric", "numeric"))
> ## view first rows of the data set
> head(AdvertisingBudgetandSales)

## # A tibble: 6 x 4
##   `TV Ad Budget ($)` `Radio Ad Budget ($)` `Newspaper Ad Budget ($)` `Sales ($)`
##                <dbl>                 <dbl>                      <dbl>       <dbl>
## 1              230.                   37.8                       69.2        22.1
## 2               44.5                  39.3                       45.1        10.4
## 3               17.2                  45.9                       69.3         9.3
## 4              152.                   41.3                       58.5        18.5
## 5              181.                   10.8                       58.4        12.9
## 6                8.7                  48.9                       75           7.2
```

I am going to fit a simple linear regression model where sales is my response variable and advertisement budget in TV is my predictor variable. In r, we use the function *lm()* to fit simple linear regression model as follows;

```
> ## slr implying simple linear regression
> slrTV <- lm(formula = `Sales ($)` ~ `TV Ad Budget ($)`, data =
    AdvertisingBudgetandSales)
> slrTV

##
## Call:
## lm(formula = `Sales ($)` ~ `TV Ad Budget ($)`, data = AdvertisingBudgetandSales)
##
## Coefficients:
##       (Intercept)  `TV Ad Budget ($)`
##           7.03259             0.04754
```

From the results above, the model can be written as;

$$\hat{y} = 7.03259 + 0.04754\hat{x}$$

or specifically;

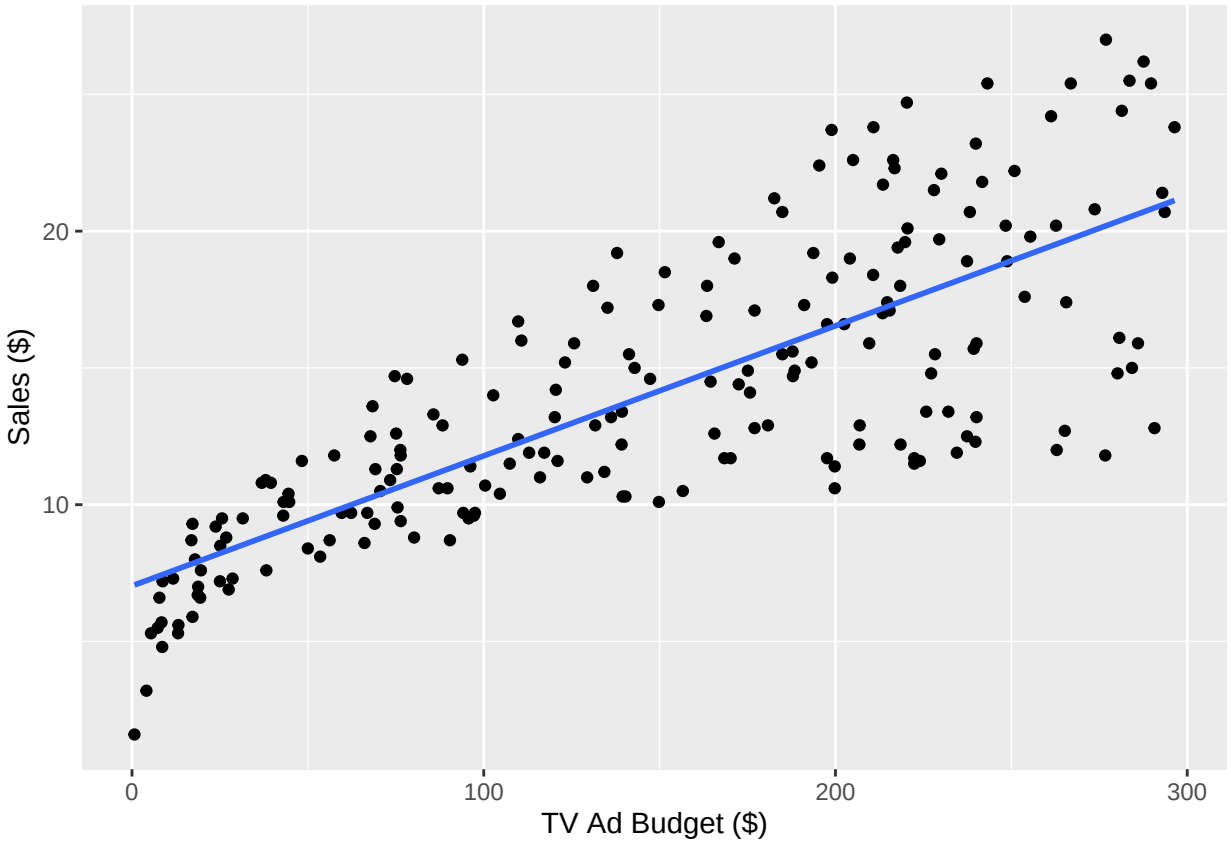
$$sales = 7.03259 + 0.04754 TV Advert$$

This implies that if there is no budget on TV advertisement, then the sales will stand at \$7032.59, that is (7.03259*1000, since the cost or sales were in 1000 dollars). Then a $\hat{\beta_1}$ of 0.04754 implies that for a TV advertisement budget equal to 1000 dollars, we expect an increase of 47.54 (0.047541000) units in sales. This implies that;

$$sales = 7.03259 + 0.04754 * 1000 = 54.57259 units$$

Since we are operating in units of thousand dollars, this represents a total sale of 54572.59 dollars. The fitted regression line is shown below;

```
> library(ggplot2)
> PLOT1 <- ggplot(data = AdvertisingBudgetandSales, aes(`TV Ad Budget ($)`, `Sales
    ($)`)) +
+     geom_point() + stat_smooth(method = lm, se = FALSE)
>
> PLOT1
```



2.1.1 Model Assessment

Before using the model to predict future values, we need to check whether;

1. There is a statistically significant relationship between the predictor variable (TV advert) and the response variable (sales)
2. The model fits well with the data.

Using the *summary()* function, we will get more insight about the model.

```
> summary(slrTV)
```

```
##
## Call:
## lm(formula = `Sales ($)` ~ `TV Ad Budget ($)`, data = AdvertisingBudgetandSales)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3860 -1.9545 -0.1913  2.0671  7.2124
##
```

```
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         7.032594   0.457843   15.36   <2e-16 ***
## `TV Ad Budget ($)` 0.047537   0.002691   17.67   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.259 on 198 degrees of freedom
## Multiple R-squared:  0.6119, Adjusted R-squared:  0.6099
## F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

We use the p-value or the t-statistic to check whether there is a statistically significant relationship between the given predictor variable and the response variable. That is, we check whether or not the β coefficient of the predictor variable is significantly different from zero. The hypothesis is formulated as;

$$H_0 : \hat{\beta_1} = 0 Vs H_1 : \beta_1 \hat{} \neq 0$$

In this case, the p-value is less than $0.05(\alpha)$ hence we reject the null hypothesis and conclude that there is a statistically significant relationship between sales and TV advertisement. We rarely test $\hat{\beta_0}$. The t-statistic is calculated as;

$$t = \frac{\hat{\beta_1} - 0}{SE(\hat{\beta_1})}$$

,

where SE is the standard error of the coefficient $\hat{\beta_1}$.

It is worthy to note that a high t-statistic and a low p-value indicates that the specific predictor variable should be retained in the model,like in our case.

The **standard error** represented by *Std.Error* in the r output above, measures the variability or the accuracy of the β coefficients. The standard error is used to calculate the confidence interval of the coefficients. For example, a 95% confidence interval of $\hat{\beta_1}$ is calculated as;

$$\hat{\beta_1} \pm 2SE(\hat{\beta_1})$$

The lower limit;

$$\hat{\beta_1} - 2SE(\hat{\beta_1})$$

$$0.047537 - 2 * 0.002691 = 0.042155$$

The upper limit;

$$\hat{\beta_1} + 2SE(\hat{\beta_1})$$

$$0.047537 + 2 * 0.002691 = 0.042155 = 0.052919$$

Therefore, there is a 95% chance that the interval (0.042155,0.052919) will contain the true value of $\hat{\beta_1}$. Alternatively, it can be done using the *confint()* function in r,

```
> confint(slrTV)

##                         2.5 %      97.5 %
## (Intercept)         6.12971927 7.93546783
## `TV Ad Budget ($)` 0.04223072 0.05284256
```

2.1.2 Model Accuracy

The overall quality of the linear regression can be assessed using the following three quantities.

1. RSE (Residual Standard Error, also known as the model sigma) - it is the standard deviation of the residuals. It represents the average variation of observations points around the regression line. When comparing two model, the model with the lower RSE is the better one. In this case, the RSE is 3.259 which is relatively low.
2. R-Squared (R^2) - It represents the proportion or variation in the data that can be explained by the model, where $0 < R^2 < 1$, but is mostly outlined as a percentage for easier interpretation. The higher the R^2, the better the model. In this case, the $R^2 = 0.6119$ which is equivalent to 61.19%, implies that 61.19% of the total variation in sales, is explained by the model. As you add more predictor variables, R^2 tend to increase, therefore in multiple linear regression, we use the $adjusted R^2$, To check the accuracy of the model. In simple linear regression, R^2 is the square of the Pearson's correlation coefficient r.

```
> cor(AdvertisingBudgetandSales$`TV Ad Budget ($)`, AdvertisingBudgetandSales$`
    Sales ($)`,
+     method = c("pearson"))

## [1] 0.7822244

> 0.7822244^2

## [1] 0.611875
```

3. The F-statistic gives the overall significance of the model. Notice that the F-statistic is used to test the overall significance of the model while the t-statistic is used to test the significance of the individual predictor variables. However, in simple linear regression, it has no much use since we only have one predictor variable. It becomes useful while dealing with multiple linear regression. In fact, for any simple linear regression model with 1 degree of freedom, the F-statistic is approximately equal to the square of the t-statistic(of $\hat{\beta}_1$).

```
> 17.67^2

## [1] 312.2289
```

2.2. Multiple Linear Regression

Multiple linear regression is an extension of simple linear regression, whereby several predictor variables are used to predict the outcome of the response variable. Assuming that there are three predictor variables, the model can be written as;

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

For easier understanding, let's build a model for estimating sales based on advertisement budged invested on TV, Radio and newspaper. This model can be written as;

$$sales = \beta_0 + \beta_1 TV + \beta_2 Radio + \beta_3 Newspaper$$

In r,it is done as follows;

```
> attach(AdvertisingBudgetandSales)
> ## entering the model;, mlr implying multiple linear regression
>
> mlr <- lm(`Sales ($)` ~ `TV Ad Budget ($)` + `Radio Ad Budget ($)` + `Newspaper
    Ad Budget ($)`,
+     data = AdvertisingBudgetandSales)
> mlr
```

```
##
## Call:
## lm(formula = `Sales ($)` ~ `TV Ad Budget ($)` + `Radio Ad Budget ($)` +
##     `Newspaper Ad Budget ($)`, data = AdvertisingBudgetandSales)
##
## Coefficients:
##                 (Intercept)          `TV Ad Budget ($)`
##                    2.938889                    0.045765
##     `Radio Ad Budget ($)`  `Newspaper Ad Budget ($)`
##                    0.188530                   -0.001037
```

Therefore, the model can be written as;

$$sales = 2.939 + 0.046TV + 0.189Radio - 0.001Newspaper$$

For more analysis of the model, we use the *summary ()* function.

```
> summary(mlr)
```

```
##
## Call:
## lm(formula = `Sales ($)` ~ `TV Ad Budget ($)` + `Radio Ad Budget ($)` +
##     `Newspaper Ad Budget ($)`, data = AdvertisingBudgetandSales)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8277 -0.8908  0.2418  1.1893  2.8292
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  2.938889   0.311908   9.422   <2e-16 ***
## `TV Ad Budget ($)`           0.045765   0.001395  32.809   <2e-16 ***
## `Radio Ad Budget ($)`        0.188530   0.008611  21.893   <2e-16 ***
## `Newspaper Ad Budget ($)` -0.001037   0.005871  -0.177     0.86
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.686 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
```

For multiple linear regression, the first step is to check whether the model is significant using the F-statistic or the corresponding p-value. The hypothesis is formulated as;

$$H_0 : \hat{\beta_1} = \hat{\beta_2} = \hat{\beta_3} = 0 vs H_1 : \hat{\beta_i} \neq 0 for i = 1, \cdots, 4$$

That is, at least one coefficient is not equal to zero or it is significant. In this case the $p - value = 2.2e - 16 < 0.05$, hence we reject the null hypothesis and conclude that the model is significant.

To test the individual significance of the predictor models, we use the t-statistic. If a predictor variable is not statistically significant, then the variable should be dropped.

```
> summary(mlr)$coefficients
```

```
##                                Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept)                2.938889369 0.311908236  9.4222884 1.267295e-17
## `TV Ad Budget ($)`         0.045764645 0.001394897 32.8086244 1.509960e-81
```

```
## `Radio Ad Budget ($)`       0.188530017 0.008611234 21.8934961 1.505339e-54
## `Newspaper Ad Budget ($)` -0.001037493 0.005871010 -0.1767146 8.599151e-01
```

From the output above, TV and Radio predictor variables are statistically significant, but Newspaper is not, since its p-value is greater than 0.05.

The coefficients are interpreted as follows, for a fixed amount of Radio and Newspaper advertisement budget, spending an additional $1000 on TV advertisement leads to an increase in sales by approximately 0.045764645*1000= 45.76465 sale units on average. For the Radio advertisement it can be interpreted through the same way. However, for the Newspaper advertisement, it implies that for a fixed amount of TV and Radio advertisement budget, changes in the advertisement budget will not significantly change the sales unit, hence we should remove it from the model, to increase the adjusted R squared.

```
> mlr2 <- lm(`Sales ($)` ~ `TV Ad Budget ($)` + `Radio Ad Budget ($)`, data =
    AdvertisingBudgetandSales)
> summary(mlr2)

##
## Call:
## lm(formula = `Sales ($)` ~ `TV Ad Budget ($)` + `Radio Ad Budget ($)`,
##     data = AdvertisingBudgetandSales)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.7977 -0.8752  0.2422  1.1708  2.8328
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            2.92110    0.29449   9.919   <2e-16 ***
## `TV Ad Budget ($)`     0.04575    0.00139  32.909   <2e-16 ***
## `Radio Ad Budget ($)`  0.18799    0.00804  23.382   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.681 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
```

The model can be written as;

$$sales = 2.921 + 0.046TV + 0.188Radio$$

The confidence interval is;

```
> confint(mlr2)

##                           2.5 %     97.5 %
## (Intercept)            2.34034299 3.50185683
## `TV Ad Budget ($)`     0.04301292 0.04849671
## `Radio Ad Budget ($)` 0.17213877 0.20384969
```

In multiple linear regression, R^2 is the correlation between the observed values of the response variable and the fitted (or predicted) values of the response variable, hence we use the $adjustedR^2$ to measure the accuracy of the model. In this case, the $adjustedR^2 = 0.8962$, which implies that 89.62% of the total variation in the sales, is explained by the model.

2.2.1 Sums of Squares

Sums of Squares in regression is a technique used to determine dispersion of data points. They are divided into two.

1. Sums of Squares due to Regression (SSR)- It is the sum of the differences between the fitted values and the mean of the response variable.

$$\sum_{n=1}^{n}(\hat{y}-\bar{y})^2$$

2. Sums of Squares Error (SSE). It is the sum of the differences between the observed values and the predicted or fitted values.

Total sums of squares is the sum of error and regression sums of squares.

$$SST = SSR + SSE$$

Note that, if $SSR = SSE$, then it implies that the regression model captures all the observed variability and is perfect.

3. Residual Sums of Squares (RSS)- It used to measure the amount of variance in a data set that is not explained by a regression model. It measures the overall difference between the observed data, and the values predicted (or fitted) by the estimation model. The lower the value, the better the model

$$\sum_{n=1}^{n} e_i^2$$

In r, we get the above information using the Analysis of Variance function *anova()* as follows;

```
> anova(mlr2)

## Analysis of Variance Table
##
## Response: Sales ($)
##                          Df Sum Sq Mean Sq F value    Pr(>F)
## `TV Ad Budget ($)`        1 3314.6  3314.6 1172.50 < 2.2e-16 ***
## `Radio Ad Budget ($)`     1 1545.6  1545.6  546.74 < 2.2e-16 ***
## Residuals               197  556.9     2.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The package *qpcR* in r also have very important functions that are useful in regression analysis.

3. Regression Model Diagonistics

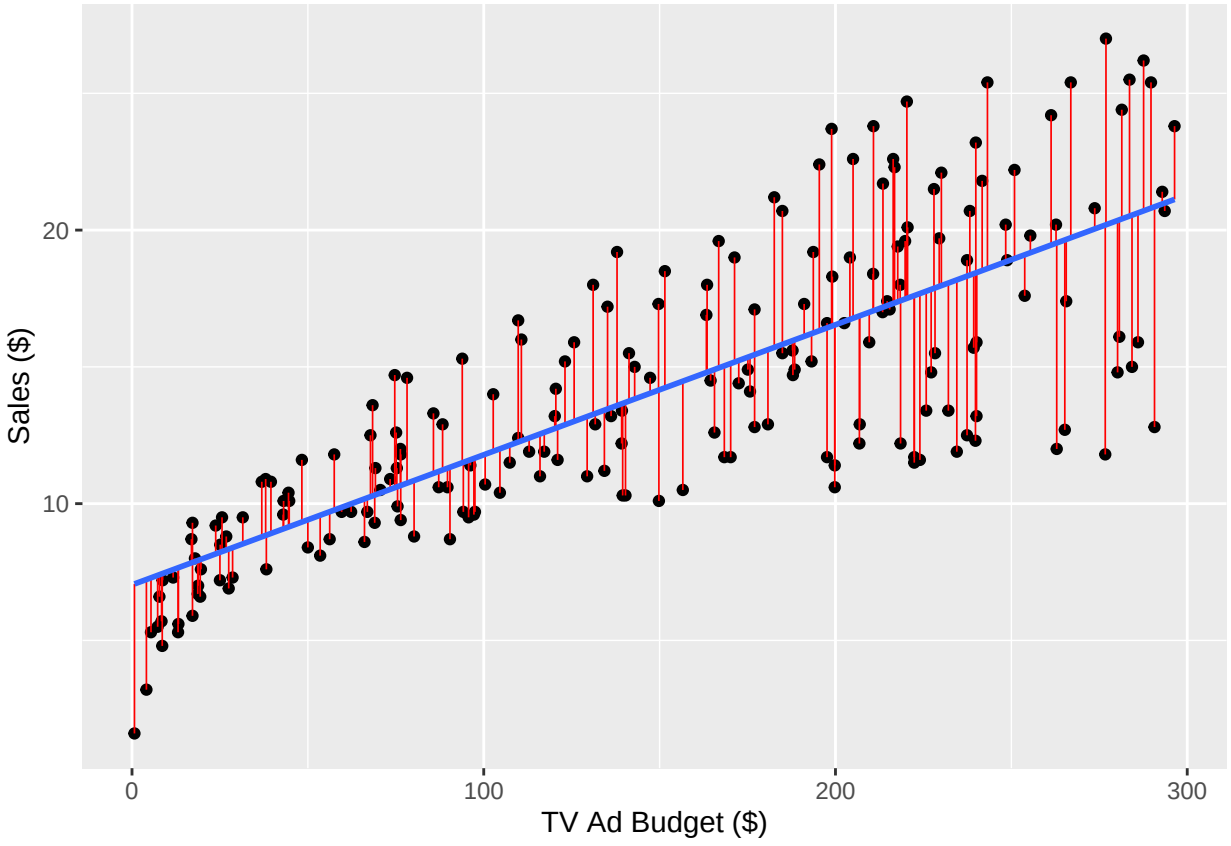
After performing regression analysis, it is important to check whether the model works well for the data in hand. This chapter will explore different ways to check the accuracy of the model. It is important to evaluate how well the model fits the data because it helps you check whether the linear regression assumptions have been met or not. For instance, linear regression assumes that there is a linear relationship between the predictor variable and the response variable which might not be the case. The relationship might be polynomial or logarithmic. In addition, data might contain outliers or extreme values which may affect the regression. This is achieved by checking the distribution of the residual errors. Note that the predicted or the fitted values are the response variable values that you would expect for the given predictor variable values, according to the built regression model. From the scatter plot below, you can see that not all points fall exactly on the regression line. This means that for a given TV or Radio advertisement budget, the observed or the measured values can be different from the predicted or fitted values. The difference is known as **residual errors**, represented by the vertical red lines. The *augment* function from *broom* package gives several metrics useful in regression diagnostic. For easier explanation, I'll use the simple linear regression model.

```
> library(tidyverse)
> library(broom)
> library(ggplot2)
> slrTVdiag <- augment(slrTV)
> head(slrTVdiag)
```

```
## # A tibble: 6 x 8
##   `Sales ($)` `TV Ad Budget (~` .fitted .resid    .hat .sigma .cooksd .std.resid
##         <dbl>             <dbl>   <dbl>  <dbl>   <dbl>  <dbl>   <dbl>      <dbl>
## 1        22.1              230.    18.0   4.13 0.00970   3.25 7.94e-3       1.27
## 2        10.4               44.5    9.15  1.25 0.0122    3.27 9.20e-4       0.387
## 3         9.3               17.2    7.85  1.45 0.0165    3.27 1.69e-3       0.449
## 4        18.5              152.    14.2   4.27 0.00501   3.25 4.34e-3       1.31
## 5        12.9              181.    15.6  -2.73 0.00578   3.26 2.05e-3      -0.839
## 6         7.2                8.7    7.45 -0.246 0.0180    3.27 5.34e-5      -0.0762
```

The plot is as follows;

```
> PLOT2 <- ggplot(slrTVdiag, aes(`TV Ad Budget ($)`, `Sales ($)`)) + geom_point() +
+     geom_segment(aes(xend = `TV Ad Budget ($)`, yend = .fitted), col = "red",
    size = 0.3) +
+     stat_smooth(method = lm, se = FALSE)
> PLOT2
```

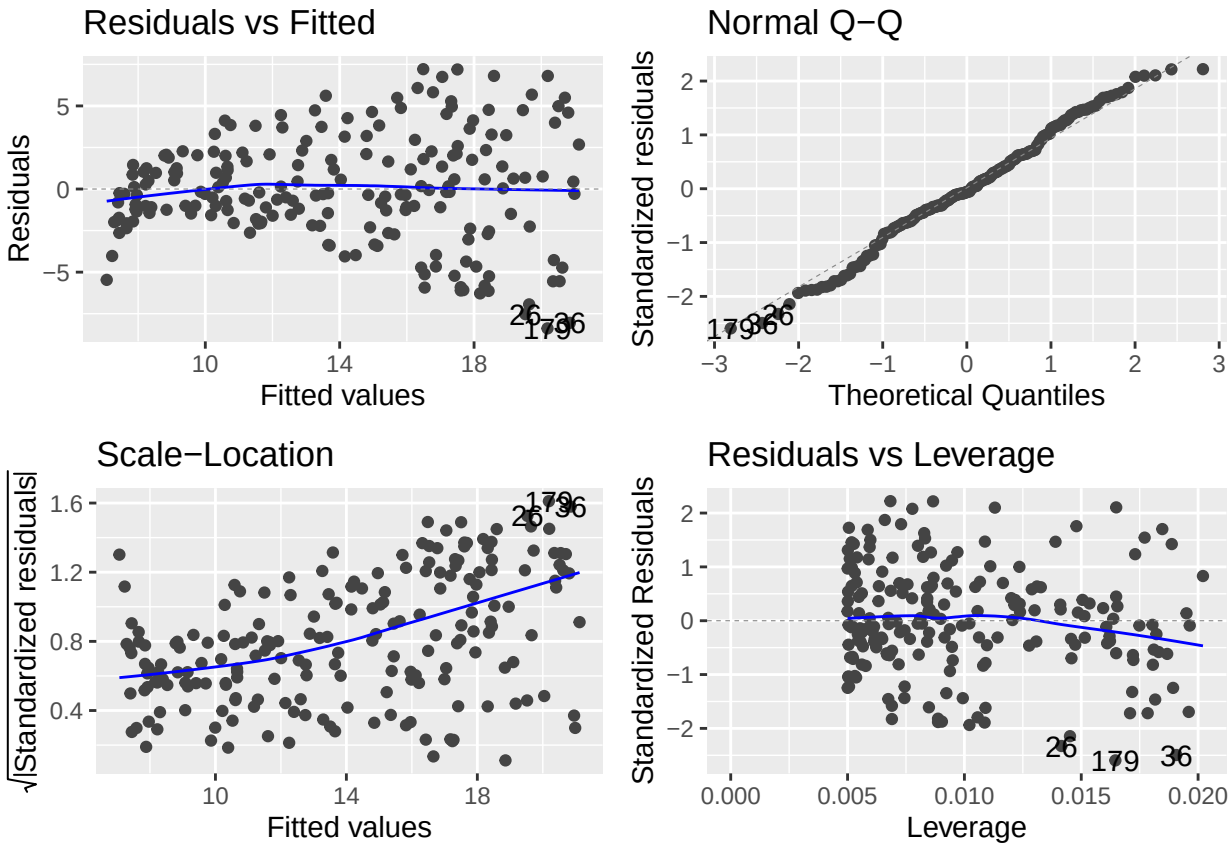


As mentioned earlier, the linear regression assumption are linearity, normality of residuals, Homogeneity of residual variance and the independence of the residual error terms.

3.1 Diagnostic Plot.

The base function *plot()* or the *autoplot()* function from *ggfortify* package can be used to plot regression diagnostic plots as follows;

```
> library(ggfortify)
> autoplot(slrTV)
```



1.The Residual vs Fitted plot- It used to check linear relationship assumption. An approximate horizontal line without distinct pattern is a good indication of linear relationship.

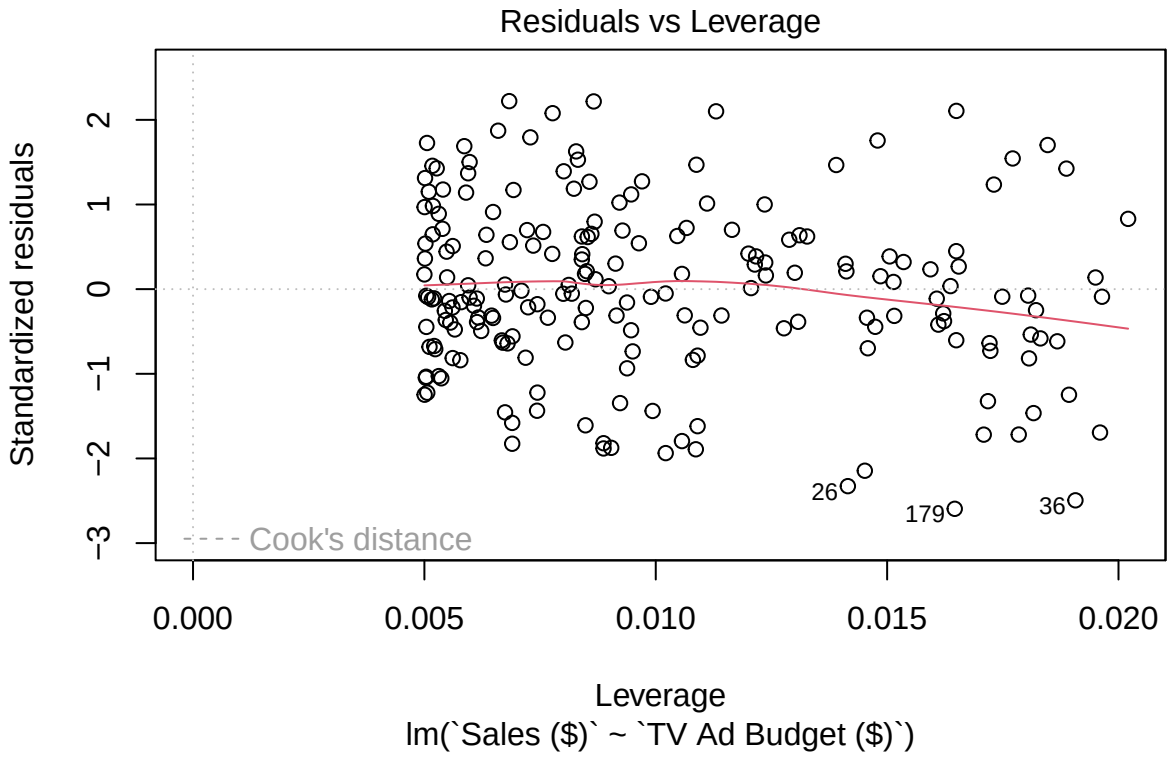
2. The Normal Q-Q Plot- This plot is used to check whether the residuals are normally distributed. The residual terms should follow the straight dashed line to satisfy the assumption.

***3. The Scale-Location Plot*- It used to check whether the residuals have a constant variance(homoscedasticity). A horizontal line with equally spread points is an indication of a constant variance which is not the case in our plot. The plot indicates that the variance of the residuals is heteroscedastic, which should be dealt with.

4. Residuals vs Leverage Plot- It is used to check extreme values that may affect the regression.Outliers may affect the interpretation of the model since they increase the RSE of the model.

For our case, the plot shows that there is linear relationship and that the residuals are normally distributed. Let us check the high leverage values and influential values.

```
> plot(slrTV, 5)
```



A data point has a high leverage if it has an extreme predictor variable values. A data point above the statistic

$$\frac{2(p+1)}{n}$$

(where p is the number of predictors and n is the number of observations) indicates an observation with high leverage. In our case, the statistic is

$$\frac{2*2}{200} = 0.02$$

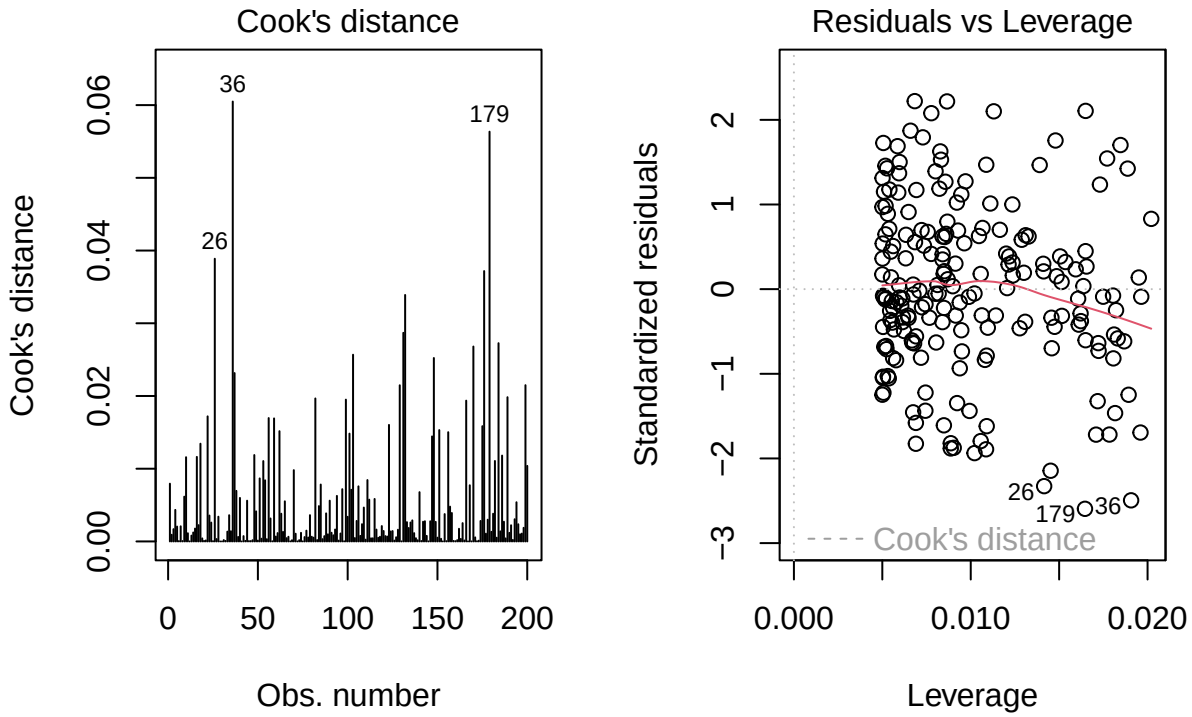
The plot above indicates outliers on the 26, 36 and 179 which have a standardized error below -2, however none exceed a standard deviation of 3. All the observations are below 0.02, hence there are no observations with high leverage.

An influential value is a value which may alter the regression if it is included or excluded in the building of model. Note that not all ouliers are influential values. An observation has influence if its Cook's distance (P. Bruce and Bruce 2017) exceeds;

$$\frac{4}{n-p-1}$$

.

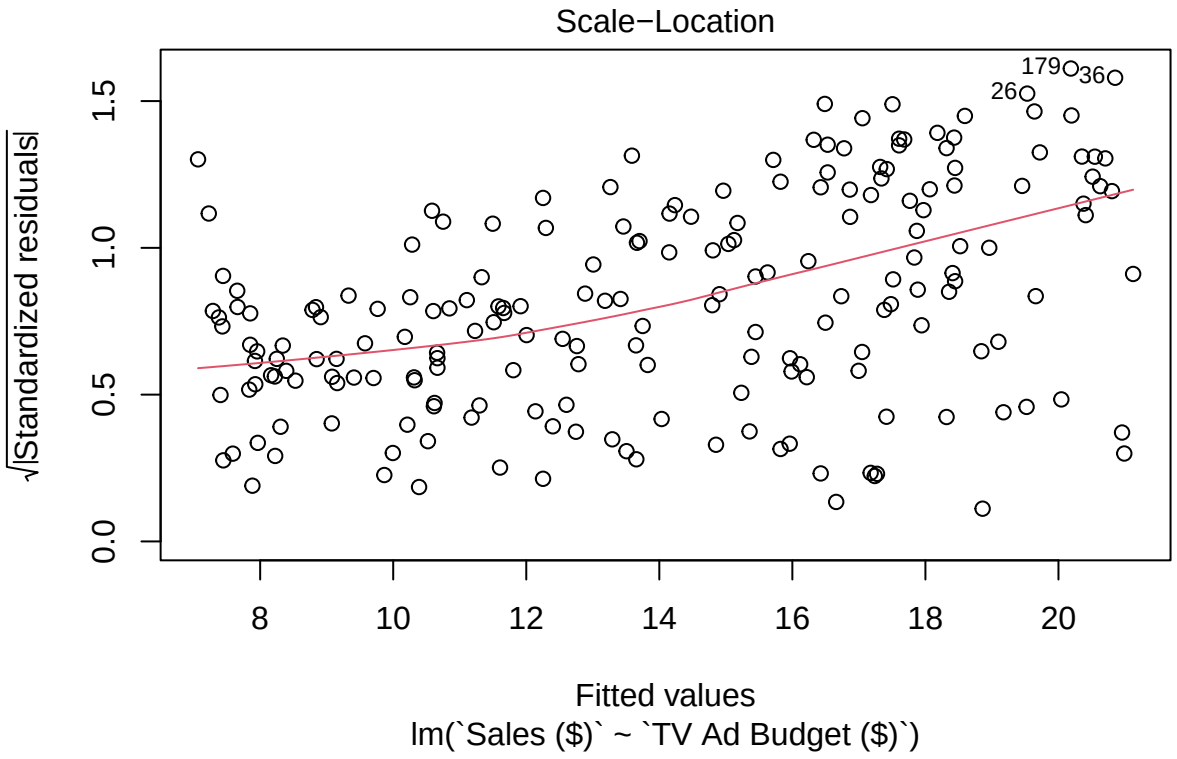
```
> par(mfrow = c(1, 2))
> plot(slrTV, 4)
> plot(slrTV, 5)
```



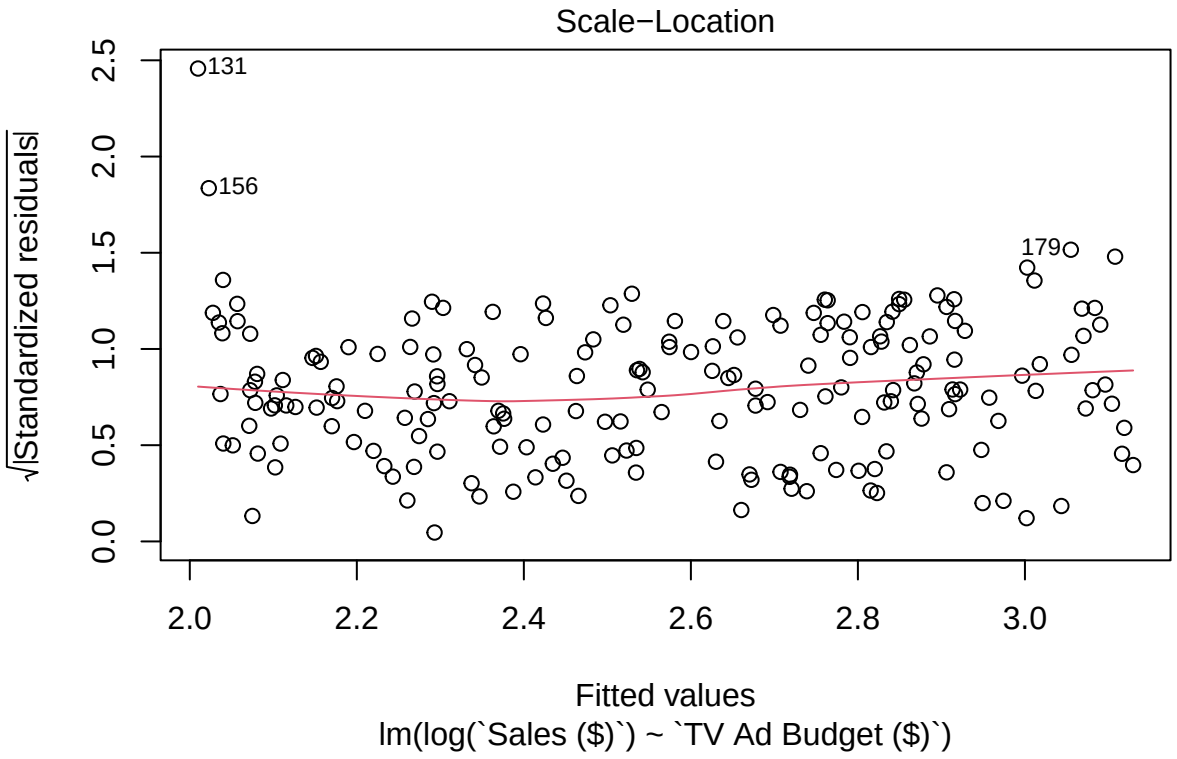
In our case, we do not have influential values, the cooks distance (represented by a red dotted line) is not drawn in the above plot because all the observations are well within the Cooks distance.

From the plot below, there is heteroscedasticity. This can be eliminated by transforming the data in various ways such as log transformation of the response variable.

```
> plot(slrTV, 3)
```




```
> slrTVlog <- lm(log(`Sales ($)`) ~ `TV Ad Budget ($)`, data =
    AdvertisingBudgetandSales)
> plot(slrTVlog, 3)
```



From the plot above, the variance of the residual is homoscedastic.

4 Interaction Effect in Multiple Linear Regression

The multiple linear regression $sales = 2.921 + 0.046TV + 0.188Radio$ is also known as an **additive model**. This model only investigates the main effects of the model, where the assumption is that the relationship between one predictor variable and the response variable is independent of the other predictor variables. For example, in the above model, the effect on sales due to TV advertisement is independent of Radio advertisement which might not be true. It can be the case that spending money on TV advertisement may also increase the Radio advertisement effectiveness. In business, this is known as **synergy** while in statistics it is known as *interaction effect*. Generally, the model is written as (assuming we have two independent variables);

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_1 x_2$$

In r, we can build an interaction model as follows;

```
> mlrinteract <- lm(`Sales ($)` ~ `TV Ad Budget ($)` + `Radio Ad Budget ($)` + `TV
    Ad Budget ($)`:`Radio Ad Budget ($)`,
+     data = AdvertisingBudgetandSales)
```

Or alternatively

```
> mlrinteract <- lm(`Sales ($)` ~ `TV Ad Budget ($)` * `Radio Ad Budget ($)`, data
    = AdvertisingBudgetandSales)
> ## Both ways will build the interaction model Let us have glimpse of the model
> ## metrics
>
> summary(mlrinteract)
```

```
##
## Call:
## lm(formula = `Sales ($)` ~ `TV Ad Budget ($)` * `Radio Ad Budget ($)`,
##     data = AdvertisingBudgetandSales)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3366 -0.4028  0.1831  0.5948  1.5246
##
## Coefficients:
##                                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                             6.750e+00  2.479e-01  27.233   <2e-16
## `TV Ad Budget ($)`                      1.910e-02  1.504e-03  12.699   <2e-16
## `Radio Ad Budget ($)`                   2.886e-02  8.905e-03   3.241   0.0014
## `TV Ad Budget ($)`:`Radio Ad Budget ($)` 1.086e-03  5.242e-05  20.727   <2e-16
##
## (Intercept)                             ***
## `TV Ad Budget ($)`                      ***
## `Radio Ad Budget ($)`                   **
## `TV Ad Budget ($)`:`Radio Ad Budget ($)` ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9435 on 196 degrees of freedom
## Multiple R-squared:  0.9678, Adjusted R-squared:  0.9673
## F-statistic:  1963 on 3 and 196 DF,  p-value: < 2.2e-16
```

From the output above it can be seen that all the coefficients including the interaction coefficient are statistically significant (Note: If the interaction effect is statistically significant, do not try to interpret the predictor variables independently). The model is;

$$sales = 6.750220 + 0.019101TV + 0.028860Radio + 0.001086TV * Radio$$

This interaction is known as **two way interaction** because it is interaction between two independent variables. High order interaction is possible also.

If the Radio advertisement budget is zero, then;

$$sales = 6.750 + 0.019TV$$

The above implies that if Radio budget is zero, then TV advertisement causes an average of 0.019*1000 dollars change in sales

However, if the Radio advertisement budget is one (or $1000 for these case), then;

$$sales = 6.750 + 0.019TV + 0.029Tv$$

Therefore;

$$sales = 6.750 + 0.048TV$$

The above implies that a budget of 1000 dollars in Radio advertisement, causes a 0.048*1000 dollars change in sales.

A positive interaction in this case implies that the larger the Radio advertisement budget, the higher the effect of TV advertisement on the sales and similary, the larger the TV advertisement budget, the higher the effect of Radio advertisement on sales.

4.1 Additive and Interaction Models Comparison

The **Root Mean Square Error (RMSE)** of the additive model is;

```
> library(qpcR)
> RMSE(mlr2)

## [1] 1.668703
```

While the RMSE of the interaction model is;

```
> RMSE(mlrinteract)

## [1] 0.9340326
```

The lower the RMSE, the better the model. RMSE is the standard deviation of the residuals. It is a metric that tells us the average distance between the predicted or the fitted values and the observed or measured values. It is calculated as;

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} P_i - O_i}{n-1}}$$

Also the RMSE is the square root of the Mean Square Error(MSE) where MSE is the average of the squared differences between the observed values and the predicted values. For example;

```
> anova(mlrinteract)

## Analysis of Variance Table
##
## Response: Sales ($)
##                                          Df Sum Sq Mean Sq F value    Pr(>F)
## `TV Ad Budget ($)`                        1 3314.6  3314.6 3723.36 < 2.2e-16
## `Radio Ad Budget ($)`                     1 1545.6  1545.6 1736.22 < 2.2e-16
## `TV Ad Budget ($)`:`Radio Ad Budget ($)`  1  382.4   382.4  429.59 < 2.2e-16
## Residuals                               196  174.5     0.9
##
## `TV Ad Budget ($)`                       ***
## `Radio Ad Budget ($)`                    ***
## `TV Ad Budget ($)`:`Radio Ad Budget ($)` ***
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> ## MSE = 0.9 implying that RMSE=sqrt(0.9)
>
> sqrt(0.9)

## [1] 0.9486833

> RMSE(mlrinteract)

## [1] 0.9340326
```

The Residual Standard Error or the model sigma is the variant of RMSE adjusted for the number of predictors. Therefore, since the interaction model has lower RMSE, it is the best model.

Also, the $adjusted R^2$ of the interaction model is 0.9673 which is equivalent to 96.73%, while that of the additive model is 0.8962 or 89.62%, implying that the interactive model is better than the additive model, since 96.73% of the total variation in sales is explained by the interactive model.

5 Regression Model Validation.

The commonly used metrics for validation of a regression model are;

1. Root Mean Square Error (RMSE)
2. Residual Standard Error (RSE)
3. R-squared (R^2)
4. Mean Absolute Error (MAE)
5. Akaike Information Criterion (AIC) or AICc
6. Bayesian Information Criterion (BIC)

The first three metrics have already been discussed above.

5.1 Akaike Information Criterion (AIC)

AIC was developed by a Japanese statistician Hirotugu Akaike, in 1970. AIC penalizes the inclusion of additional variables in the model. It is used to compare various models of the same data and determine which model is the best. The best model is the model with the lowest AIC. Notice that adding more parameters increases the AIC, hence the model with fewer parameters will have the lower AIC. [2]

According to AIC, the best model is the model which explains the greatest amount of variation in the dependent variable using the fewest possible independent variables. It is calculated as;

$$AIC = 2k - 2\ln L$$

where;

-k is the number of independent variables.

-L is the log-likelihood estimate (Likelihood that your model would have produced the observed model)

The default number of independent variables is 2, so if you have one independent variable, then, $k = 3$ and so on.

To use the AIC, you need to build several models and then compare them. For example, we may build several models using the advertisement data set and see which best explains the variations in sale. In this case, I will build separate models for each independent variable, and then compare it with the model for the combined independent variables (TV, Radio and Newspaper). I had already done models for TV advert then TV, Radio and Newspaper advert, as well as TV, Radio advert, hence I'll just proceed to the remaining two separate models for Radio and Newspaper.

```
> slrRadio <- lm(`Sales ($)` ~ `Radio Ad Budget ($)`, data =
    AdvertisingBudgetandSales)
> slrNewspaper <- lm(`Sales ($)` ~ `Newspaper Ad Budget ($)`, data =
    AdvertisingBudgetandSales)
```

For clarification, I have named the models as follows;

- slrTV - TV advertisement
- slrRadio - Radio Advertisement
- slrNewspaper - Newspaper advertisement
- mlr - model for combined TV, Radio and Newspaper advert
- mlr2 - combined model for TV and Radio adverts without newspaper
- mlrinteract - interaction model

I'm going to use the function *aictab()* from the *AICcmodavg* package as follows;

[2] Cavanaugh, J. E., & Neath, A. A. (2019). The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements. Wiley Interdisciplinary Reviews: Computational Statistics, 11(3), e1460.

```
> ## listing the models in a list
> Models <- list(slrTV, slrRadio, slrNewspaper, mlr, mlr2, mlrinteract)
> ## naming the models
> Models.names = c("slrTV", "slrRadio", "slrNewspaper", "mlr", "mlr2", "mlrinteract
    ")
> library(AICcmodavg)
> ## Then use the function
> aictab(cand.set = Models, modnames = Models.names)

##
## Model selection based on AICc:
##
##                K    AICc Delta_AICc AICcWt Cum.Wt      LL
## mlrinteract    5  550.59       0.00      1      1 -270.14
## mlr2           4  780.60     230.01      0      1 -386.20
## mlr            5  782.67     232.08      0      1 -386.18
## slrTV          3 1044.21     493.63      0      1 -519.05
## slrRadio       3 1152.80     602.21      0      1 -573.34
## slrNewspaper   3 1222.79     672.21      0      1 -608.34
```

The best fit model is always listed first. The best model for this study is the Interaction model (Recall that it is the interaction between TV and Radio Adverts). The AICc in the output above contains the model information -the lower the value the better the model. The lowercase "c" implies that it is the AIC of small samples. The Delta_AICc (or Delta_AIC) is the difference between the AICc (or AIC) of the best model and the model being compared. LL is the log likelihood used to calculate the AIC.

5.2 Bayesian Information Criterion

It is a method for scoring and selecting a model. It is almost similar to AIC,only that while AIC penalizes the additional parameters, BIC penalizes the complexity.More complex models have higher BICs. The lower the value, the better the model. It is widely used in *logistic Regression* It is calculated as;

$$BIC = -2L + \ln N * K$$

Where; -k is the number of independent variables.

-L is the log-likelihood estimate (Likelihood that your model would have produced the observed model) - N is the number of observations.

Note that AIC and BIC are best used in models fit by Maximum Likelihood Estimation framework

5.3 Mean Absolute Error (MAE)

MAE is a loss function used for regression. The loss is the mean over the absolute differences of the observed values and the predicted values. The lower the value the better the model. It is calculated as;

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y - \hat{y}_i|$$