# MULTIPLE LINEAR REGRESSION: MODEL CHECKING AND DIAGNOSTICS

By Bidhisha Ghosh

'21

BETHUNE COLLEGE
Affiliated to University of Calcutta

# Table of Contents

# Abstract

Multiple linear Regression, also simply known as multiple regression, is a statistical technique that uses several explanatory or concomitant or independent variables to predict the outcome of a response or dependent variable. In this project, our aim is to get a best fit multiple linear regression model for our data by detection and removal of influential values. When we have a data, which constitutes of one dependent or response variable and multiple independent or regressor variables and we want to fit a linear model to it, it is most important to first check if the linear model is suitable for the data. The data might not show a linear relationship or there may be several inadequacies present in it. Also, the presence of outliers or influential points may affect the fitted model and fail to show the true nature of the model. The outliers also degrade the overall precision of estimation. Thus, before delving deep into fitting the linear model and working with it, it is of utmost importance to check for such deficiencies and remove them or take necessary actions as and when required. This project deals with the checking of the validity of the assumptions of a linear model and detection and removal of serious model deficiencies.

# 1. Introduction

A linear model is usually fit to a data set where the response (*y*) and the regressor ($x_i{}'s$) variables have a clear linear relationship. Then our linear model is of the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{p-1} x_{p-1} + e$$

Where the $\beta_i' s$ are the parameters and *e* is the error.

However, this model is valid only under the following assumptions:

- There is a linear relationship between the response and each of the regressor variables.
- The mean of the errors must be equal to 0.
- The errors are independently distributed.
- The errors follow a normal distribution with mean 0 and constant variance.

However, every data set may not always follow these assumptions. Therefore, in this project, the above assumptions for our model is diagnosed and checked for validity. Any departure from the above assumptions may be caused due to inadequacy of the model or presence of outliers. Hence, the presence of such inadequacies or outliers in our data is also checked and necessary measures are taken if such outliers are found.

# 2. Methodology

## 2.1 The Data

In this project, the Life Expectancy Data set is used which shows the variation of Life Expectancy in age (response variable) of the population of India during the years 2000-2015, depending on several factors (regressor variables).

***"The data was collected from WHO and United Nations website with the help of Deeksha Russel and Duan Wang."*** *This Data was then published as an open source data by Kaggle, from where I collected it.*

The original data set consisted of Life expectancy in age along with 18 other factors of which 6 factors are taken into consideration. Also the data set consisted of data values for 193 countries of which only the values for India are considered.

## Variables in the Data.

I have used the following variables in the data for this project:

$y$ = Average Life Expectancy in age.

$x_1$ = Number of Infant Deaths per 10000 population.

$x_2$ = Average Body Mass Index (BMI) of the entire population.

$x_3$ = Number of Under five-deaths per 10000 population.

$x_4$ = Percentage of Polio (Pol3) immunization coverage among 1-year-olds

$x_5$ = Percentage of Diphtheria Tetanus Toxoid and Pertussis (DTP3) immunization coverage among 1-year-olds.

$x_6$ = Human Development Index in terms of Income composition of resources (ICOR) (index ranging from 0 to 1).

Here $y$ is the response variable or the dependent variable and $x_1$, $x_2$, $x_3$, $x_4$, $x_5$ and $x_6$ are the regressor or the independent variables.

Also, the total number of observations is denoted by $n$ and the number of independent variables+1 is denoted by $p$.

The data set is given in *Table 1 in Appendix*.

***The R code to read the data is:***

```
library(readxl)

file="C:/Users/Bidisha/Desktop/project sem 6/project/data.xlsx"

data1=read_excel(file)
```

## 2.2 Check for Linearity

After getting the data in hand, our first job is to check if the linear model is suitable for the data. To do this, we calculate the correlation coefficients of each of the regressor variables ($x_1$, $x_2$, $x_3$, $x_4$, $x_5$, $x_6$) with the response variable ($y$). If the correlation coefficients are all close to -1 or 1, it is said that the data is highly linear and a linear model is suitable for the data.

***The R code for calculating the correlation coefficients is:***

```
correlation=cor(data1[,3],data1[4:(p+2)])
correlation
```

## 2.3 Check if the model fits the data

Once a linear model is fitted to the data, next we check if the model fits the data well. This is done by calculating the F-Statistic and comparing it with the tabulated F-Statistic. If the calculated F-Statistic is greater than the tabulated F-Statistic, we reject the null hypothesis $H_0: \beta_0 = \beta_1 = \cdots = \beta_6 = 0$. That is, at least one $\beta_i$, $i = 0(1)6$ is non-zero which implies that our linear model fits the data well.

***The R code to fit a linear model and find the F-Statistic is:***

```
data1.model=lm(LifeExpectancy~InfantDeaths+BMI+UnderFiveDeaths+Polio
+Diphtheria+IncomeCompositionOfResources,data=data1)
data1.model
summary(data1.model)
```

## 2.4 Check for possible model inadequacies

When we have a data in hand, and we have already fitted a linear model to it, it is not always necessary that our fitted model will be adequate for the data. For this purpose, we use residual analysis. Residuals are defined as $e_i = y_i - \hat{y}_i, i = 1,2, \ldots, n$, where $y_i$ is an observation and $\hat{y}_i$ is the corresponding fitted value. So residual may be viewed as the **deviation** between the **data** and the **fit** and it is convenient to think of the residuals as the realized or observed values of the model errors. Thus, any departures from the assumptions on the errors should show up in the residuals. Analysis of the residuals is an effective way to discover several types of model inadequacies. **Plotting residuals** is a very effective way to investigate how well the regression model fits the data and to check the assumptions. Since the variance of residuals are not same, we may use the scaled residuals like *R-Student residuals* or *externally studentised residuals* for our linear model as it is often more useful to work with scaled residuals than the unscaled ones. If the R-Student residuals lie within the range from -3 to 3, it is said that the data is free from any inadequacy. To check for such defects, the R-Student residuals are plotted against the fitted values ($\hat{y}$) and any prominent pattern in the plot was looked for. If the plots have a random occurrence around a straight horizontal line, then it is considered that the model is proper. However, if the plots show a particular pattern then it indicates the presence of some model inadequacies such as heteroscedasticity or lack of fit. For example, if the plots show a funnel pattern, we conclude that the data is heteroscedastic or has a non-constant residual variance or if the plots show a curved pattern we conclude that the data is nonlinear. Plots of the R-Student residuals against the regressor variables are also often used for this purpose.

***The R code to plot the R-Student residuals versus fitted values graph is:***

```
windows()
```

```
plot(yhat,r,ylab="R-Student residuals",xlab="Fitted values", main="Graph of R-student residuals against fitted values")

windows()

plot(x[,2],r,main="graph of Number of Infant Deaths against R-student residuals",ylab="R-student residuals",xlab="Number of Infant Deaths per 10000 population")

windows()

plot(x[,3],r,main="graph of Body Mass Index(BMI) against R-student residuals",ylab="R-student residuals",xlab="Average Body Mass Index(BMI) of the entire population")

windows()

plot(x[,4],r,main="graph of Under 5 Deaths against R-student residuals",ylab="R-student residuals",xlab="Number of Under 5 Deaths per 10000 population")

windows()

plot(x[,5],r,main="graph of Percentage of Polio immunisation against R-student residuals",ylab="R-student residuals",xlab="Number of Polio (Pol3) immunisation coverage among 1 year olds")

windows()

plot(x[,6],r,main="graph of Percentage DTP3 immunisation against R-student residuals",ylab="R-student residuals",xlab="Percentage of DTP3 immunisation coverage among 1 year olds")

windows()

plot(x[,7],r,main="Income composition of resources against R-student residuals",ylab="R-student residuals",xlab="Income composition of resources ranging from 0 to 1")
```

## 2.5 Check for normality of residuals

While fitting a linear model, it is assume that the residuals follow a normal distribution with 0 mean and constant variance. However, a small departure from this normality assumption do not affect the model too greatly, but as the t and F statistic and the confidence intervals are dependent on the normality assumption, too much deviation from normality is more serious.

A simple way to check the normality assumption of residuals is to construct the ***normal probability plot.*** For this, the R-Student residuals are first sorted in ascending order and this sorted R-Student residuals are plotted against the cumulative probability which is given by

$$\boldsymbol{P_i} = (\boldsymbol{i} - \tfrac{1}{2})/\boldsymbol{n}, \quad i = 1,2,....,n$$

This plot is called the *normal probability plot*.

The resulting points on the plots must lie approximately on a straight line. The straight line is usually determined visually, with more emphasis on the central values rather than the extremes. Any departure from a straight line indicates that the distribution of the residuals is not normal.

***The R code to find the normal probability plot is:***

```
sorted_r=sort(r)

prob=(seq(1:16)-0.5)/n

plot(sorted_r,prob,main="Normal Probablity Plot",xlab="R-Student
residuals",ylab="Probability")
```

## 2.6 Check for leverage values

**Leverage point**: In a data where the response and the regressor variables are approximately linear, we may find some points which is which is remote in x space from the rest of the sample. These points are called Leverage points or the points has large Leverage value. These points may or may not be influential point. If these point lies on or close to the regression line passing through the rest of the data points, then they are not influential.

**Influential point**: If we have an outlier which lies away from the line of regression and has a tendency to pull the regression line towards itself, we call it an influential point. These points may have an unusual $x$ value and $y$ value or unusual y values. These points affect the estimated parameters greatly and need to be removed

To check for outliers or leverage values, the Hat matrix **H** calculated. The hat matrix is given by

$$\mathbf{H=X(X'X)^{-1}X'}$$

The diagonals of this Hat matrix or the *hat values* are then calculated. The hat matrix diagonals is a standardized measure of the distance of the $i^{th}$ observation from the center (or centroid) of the *x* space. Thus, large hat diagonals reveal observations that are potentially a leverage point because they are remote in *x* space from the rest of the sample. Traditionally, it is said that a point whose corresponding hat matrix diagonal or hat value is greater than the cut off $2p/n$ is remote enough from the rest of the data to be considered as a leverage point. Here, where p = number of regressor variables + 1 and n is the total number of observations in our data.

***The R code to find the hat values and the leverage points is:***

```
h=hatvalues(data1.model)

h

outlier=h[c(h>(2*p/n))]

outlier
```

## 2.7 Check for influential points through Cook's Distance

Cook (1977-1979) has suggested a way to detect influential points in a data, using a measure of the squared distance between the least-squares estimate $\widehat{\boldsymbol{\beta}}$ based on all *n* points and the estimate obtained by deleting the $i^{th}$ point, say $\widehat{\boldsymbol{\beta}}_{(i)}$. This measure of distance can be expressed as

$$D_i = \frac{(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}}\,)'\boldsymbol{X}'\boldsymbol{X}(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}}\,)}{pMS_{res}}, \quad \boldsymbol{i} = 1, 2, \ldots, n$$

Also, The $D_i$ statistic may be rewritten as

$$D_i = \frac{r_i^2\, Var(\hat{y}_i)}{p\, Var(e_i)} = \frac{r_i^2\, h_{ii}}{p\,(1-h_{ii})}, \quad \boldsymbol{i} = 1, 2, \ldots, n$$

Points with large values of $D_i$ have a considerable influence on the least squares estimates $\widehat{\boldsymbol{\beta}}$.

To check the presence of any influential points which significantly affects our estimated parameters ($\widehat{\boldsymbol{\beta}}$), the Cook's Distance values for all the observations are calculated and compared with the cut off $4/(n-p)$. The values greater than this cut off is considered to be influential.

It must be noted that, originally, Cook suggested that if for any observation, the value of the above Cook's Distance is greater than 1, then it is considered to be an influential point. However, in many studies $4/n$ or $4/(n-p)$ are also commonly used as the cut off value for Cook's Distance.

***The R code to find the Cook's Distance values and the influential points and plot the Cook's Distance versus response variable graph is:***

```
cd=cooks.distance(data1.model)

cd

influence=cd[c(cd>(4/(n-p)))]

influence

with(data1,plot(LifeExpectancy,cooks.distance(data1.model), xlab="Life Expectancy", ylab="Cook's Distance", main="Graph of Cook's Distance against Life Expectancy"))

identify(data1$LifeExpectancy,cd)
```

## 2.8 Check for points which have a significant influence on the various estimated parameters (based on DFBETAS Method)

Belsley, Kuh and Welsch (1980) introduced two other measures of deletion influence. The first method is to find a statistic that indicates how much the regression coefficient $\hat{\beta}_j$ changes, in standard deviation units, if the $i^{th}$ observation was deleted. This statistic is called $DFBETAS_{j,i}$ and is given by

$$DFBETAS_{j,i} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{S_{(i)}^2 C_{jj}}}$$

Where $C_{jj}$ is the $j^{th}$ diagonal element of $(\boldsymbol{X'X})^{-1}$ and $\hat{\beta}_{j(i)}$ is the $j^{th}$

Regression coefficient when the $i^{th}$ observation is deleted. And $S_{(i)}^2$ is calculated as

$$S_{(i)}^2 = \frac{(n-p)MS_{res} - {\hat{e}_i^2}\big/{(1-h_{ii})}}{n-p-1}$$

For the computation of $DFBETAS_{j,i}$ we first define a p x n matrix

$$\boldsymbol{R} = (\boldsymbol{X'X})^{-\mathbf{1}}\boldsymbol{X'}$$

If we let $\boldsymbol{r'}_{\boldsymbol{j}}$ denote the $j^{th}$ row of **R**, then

$$DFBETAS_{j,i} = \frac{r_{j,i}\, t_i}{\sqrt{\boldsymbol{r'}_{\boldsymbol{j}}\boldsymbol{r}_{\boldsymbol{j}}}\sqrt{1-h_{ii}}}$$

Belsley, Kuh and Welsch (1980) suggest that if for the $i^{th}$ observation, the value of $|DFBETAS_{j,i}| > 2/\sqrt{n}$ then that observation is considered to have a significant influence on the $j^{th}$ estimated parameter.

Thus the values of $DFBETAS_{j,i}$ are calculated and compared with the cut off $2/\sqrt{n}$ and any value exceeding this cut off is said to have a significant influence on the $j^{th}$ estimated parameter.

***The R code to find the DFBETAS values and the influential points is:***

```
dfb=dfbetas(data1.model)
dfb
num=seq(1:n)
beta0=num[c(abs(dfb[,1])>2/sqrt(n))]
beta1=num[c(abs(dfb[,2])>2/sqrt(n))]
beta2=num[c(abs(dfb[,3])>2/sqrt(n))]
beta3=num[c(abs(dfb[,4])>2/sqrt(n))]
beta4=num[c(abs(dfb[,5])>2/sqrt(n))]
beta5=num[c(abs(dfb[,6])>2/sqrt(n))]
beta6=num[c(abs(dfb[,7])>2/sqrt(n))]
beta0
beta1
beta2
beta3
beta4
beta5
beta6
```

## 2.9 Check for points which have a significant influence on the fitted values (based on DFFITS Method)

The second diagnostic proposed by Belsley, Kuh and Welsch is the DFFITS measures. The $DFFITS_i$ is the amount of change that the fitted value $\widehat{y_i}$ experiences, in standard deviation units, when the $i^{th}$ observation is removed. It is given by

$$DFFITS_i = \sqrt{\frac{h_{ii}}{1-h_{ii}}}\frac{e_i}{S_{(i)}\sqrt{1-h_{ii}}} = \sqrt{\frac{h_{ii}}{1-h_{ii}}}t_i$$

Where $t_i$ denotes the R-Student residuals, $h_{ii}$ denotes the hat values and $S_{(i)}^2 = \frac{(n-p)MS_{res}-\hat{e}_i/(1-h_{ii})}{n-p-1}$.

Belsley, Kuh and Welsch suggests that if for the $i^{th}$ observation, the value of $|DFFITS_i|$ > $2\sqrt{p/n}$ then the observation needs to be examined.

Thus the values of $DFFITS_i$ are calculated and compared with the cut off $2\sqrt{p/n}$ and any value exceeding this cut off is claimed to be an influential point.

***The R code to find the DFFITS values and the influential points is:***

```
dff=dffits(data1.model)

dff

outfits=dff[c(abs(dff)>2*sqrt(p/n))]

outfits
```

## 2.10 Check for observations that affect the general precision of estimation significantly (based on COVRATIO method)

To find how the various points affects the overall precision of estimation, we use the COVRATIO measures. The $COVRATIO_i$ denotes the role of the $i^{th}$ observation on the precision of estimation. It is defined as

$$COVRATIO_i = \frac{|(\boldsymbol{X'}_{(i)}\boldsymbol{X}_{(i)})^{-1}\, S_{(i)}^2|}{|(\boldsymbol{X'X})^{-1} MS_{res}|}$$

Where $\boldsymbol{X}_{(i)}$ is the design matrix when the $i^{th}$ observation is removed.

It can be computationally rewritten as

$$COVRATIO_i = \frac{(S_{(i)}^2)^p}{MS_{res}^p}\left(\frac{1}{1-h_{ii}}\right)$$

Belsley, Kuh and Welsch (1980) also suggested that if for the $i^{th}$ observation the value of $|COVRATIO_i| > 1 \pm 3\,p/n$ then the point should be considered influential.

Note that the lower bound is only valid if $n > 3p$. If we have $n < 3p$, we take the lower bound as 0.

If $COVRATIO_i > 1$, then we say that the $i^{th}$ observation ***improves*** the precision of estimation and if $COVRATIO_i < 1$ then we say that the $i^{th}$ observation ***degrades*** the precision of estimation.

***The R code to find the COVRATIO values and the influential points is:***

```
covr=covratio(data1.model)

covr

outcov=covr[c(abs(covr-1)>3*p/n)]

outcov


improve=covr[c(covr>1)]

degrade=covr[c(covr<1)]

improve

degrade
```

## 2.11 Removing the influential points and checking the change in residual variation.

Next, the influential points are removed one at a time, two at a time and so on and a linear model is fitted in each case and the value of $MS_{res}$ is calculated for each case and compared with the true value of variance of pure error. The combination of influential points, on the removal of which, the model $MS_{res}$ is minimised and is closest to the true value of variance of pure error, is our best model.

# 3. Results and Analysis

## *3.1 The relationship between the response variable and the regressor variables are highly linear and hence our linear model fits well to the data.*

In order to get an idea of the linearity of the data we calculate the correlation coefficients between the response and the regressor variables. The correlation coefficients between the response and the different regressor variables are tabulated in *Table 1* below.

***Table 1: Correlation coefficients between Life Expectancy (response variable) and the other regressor variables***

| | *Infant Deaths* | *BMI* | *Under-five deaths* | *Polio immune (%)* | *DTP3 immune (%)* | *ICOR (ranging from 0 to 1)* |
|---|---|---|---|---|---|---|
| *Life Expectancy* | -0.9946 | 0.9983 | -0.9980 | 0.9841 | 0.5263 | 0.9979 |

Thus, we see that all the correlation coefficients are either close to 1 or -1 except for the correlation coefficient between Life Expectancy and the percentage of DTP3 immunisation coverage among 1-year olds. Hence, it can be said that our response variable has a highly linear relationship with the regressor variables. The comparatively less value of the correlation coefficient between Life Expectancy and the percentage of DTP3 immunisation coverage might be due to the presence of outliers in the data.

## 3.2 *The linear model fits well to our data*

First, a linear model is fitted to the data and the values for the estimated parameters are found as in *Table 2* below.

***Table 2: Estimated values of the parameters of our fitted linear model***

| Intercept ($\hat{\beta}_0$) | Infant Deaths ($\hat{\beta}_1$) | BMI ($\hat{\beta}_2$) | Under-five Deaths ($\hat{\beta}_3$) | Polio immune (%) ($\hat{\beta}_4$) | DTP3 immune (%) ($\hat{\beta}_5$) | ICOR (ranging from 0 to 1) ($\hat{\beta}_6$) |
|---|---|---|---|---|---|---|
| 53.9113 | 0.0002 | -0.0247 | -0.0019 | -0.0043 | 0.0019 | 27.8391 |

Then the F-Statistic calculated for our linear model with 0.05 level of significance. The degrees of freedom for the F-Statistic are $p-1$ and $n-p$ respectively. Our calculated F-Statistic $F_{cal} = F'_{6,9} = 985.2$ while the tabulated F- Statistic $F_{tab} = F_{6,9} = 4.10$ at 0.05 level of significance. Clearly, we see that the calculated F-Statistic is greater than the tabulated F-Statistic and thus our null hypothesis $H_0: \beta_0 = \beta_1 = \cdots = \beta_6 = 0$ is rejected. That is, at least one $\beta_i$, $i = 0(1)6$ is non-zero, which implies that our linear model fits the data well.

## 3.3 *Our model has no inadequacies present in it.*

In our data, first the R-Student residuals are calculated as a measure of scaled residuals. The R-Student residuals, so found, are listed in *Table 3* below. Then, the R-Student residuals are plotted against the fitted values ($\hat{y}$) as shown in *Figure 1* below. Here, it is seen that the plot shows no such prominently visible pattern. The points are randomly located around a straight horizontal line. This is a clear indication that ***<u>our model has no inadequacies present in it.</u>*** The R-Student residuals are also plotted against the regressor variables as shown in *Figure 2, Figure 3, Figure 4, Figure 5, Figure 6* and *Figure 7*.

***Table 3: The residuals and the R-Student residuals of the model***

| *Observation No.* | *Residuals* | *R-Student residuals* |
|---|---|---|
| 1 | -0.108025815 | -1.9327527 |
| 2 | -0.019447668 | -0.2250493 |
| 3 | -0.029999239 | -0.3352621 |
| 4 | 0.071222184 | 0.8383731 |
| 5 | 0.030567935 | 0.4068826 |
| 6 | 0.091168615 | 1.0842758 |
| 7 | 0.016275136 | 0.1980729 |
| 8 | 0.001430613 | 0.1440260 |
| 9 | 0.027440501 | 0.3630813 |
| 10 | 0.061364666 | 0.7455980 |
| 11 | -0.074400824 | -1.3488948 |
| 12 | -0.116297532 | -2.4737812 |
| 13 | 0.109163938 | 1.7829772 |
| 14 | 0.070073837 | 0.9854124 |
| 15 | -0.026595898 | -0.3947257 |
| 16 | -0.103940448 | -1.9014273 |

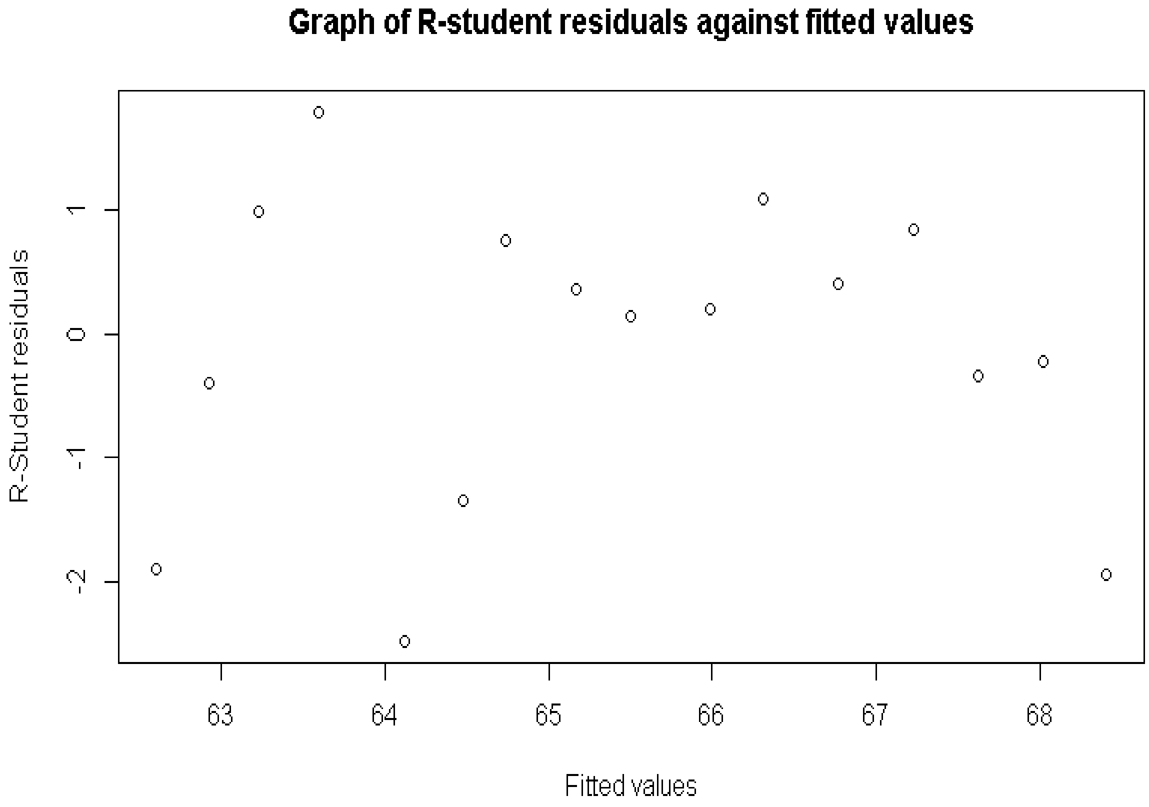


***Figure 1**: Plot of R-Student residuals against the fitted values*

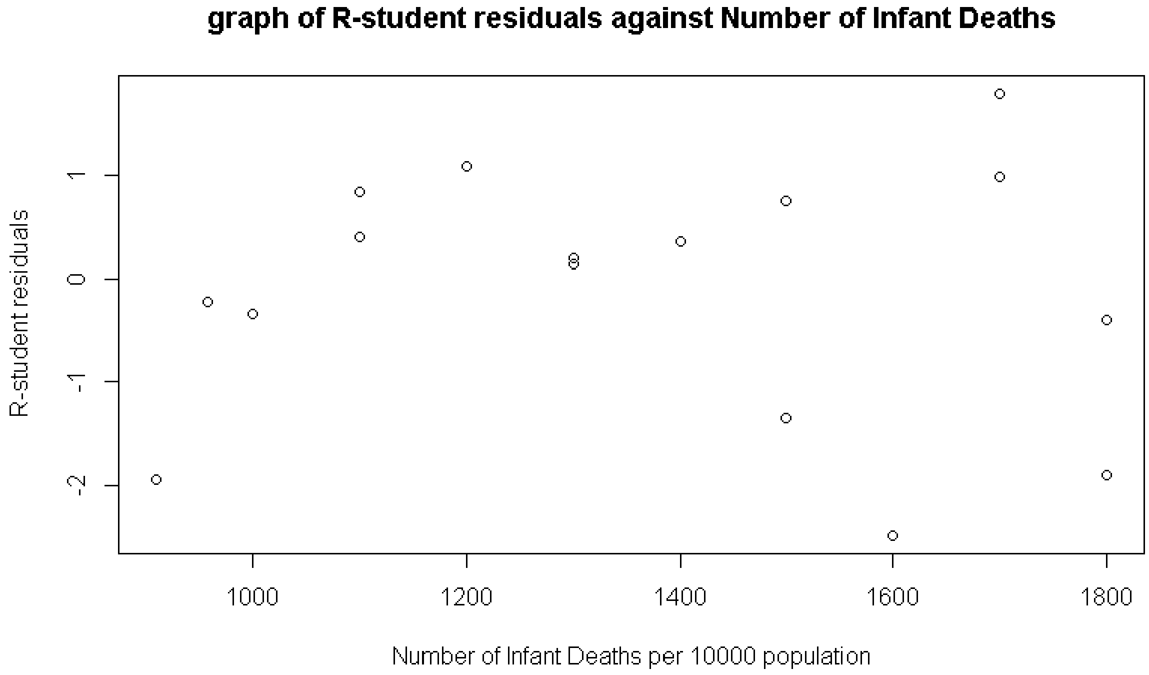


***Figure 2**: Plot of R-Student residuals against the Number of infant deaths per 10000 population*

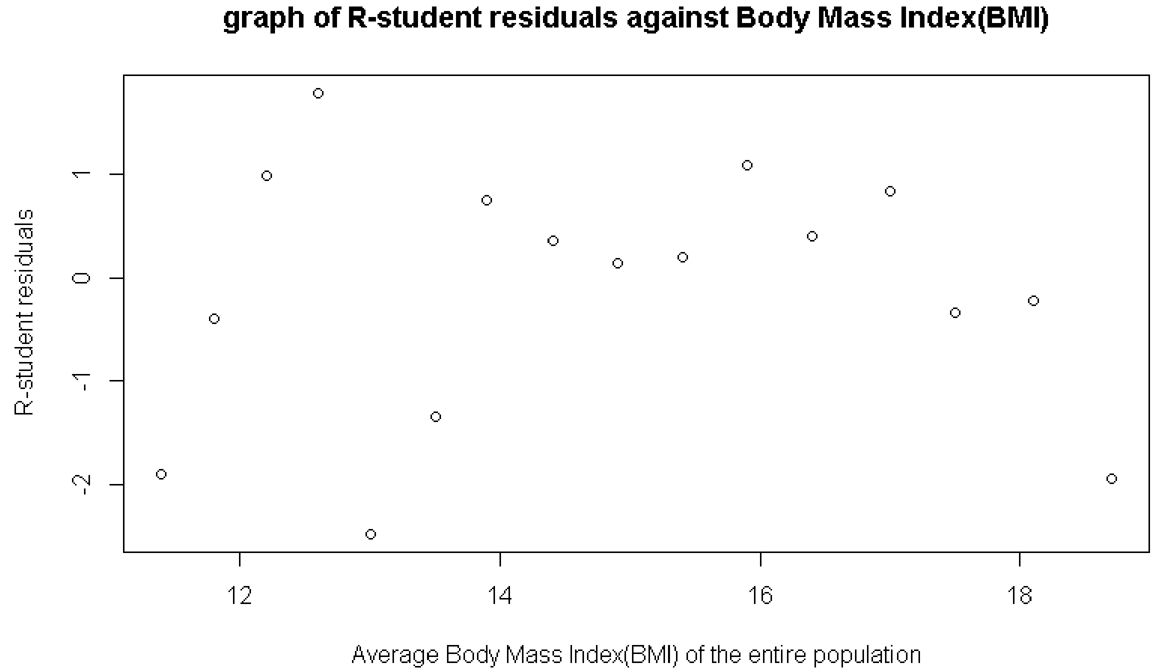


***Figure 3**: Plot of R-Student residuals against the average Body Mass Index (BMI) of the entire population*

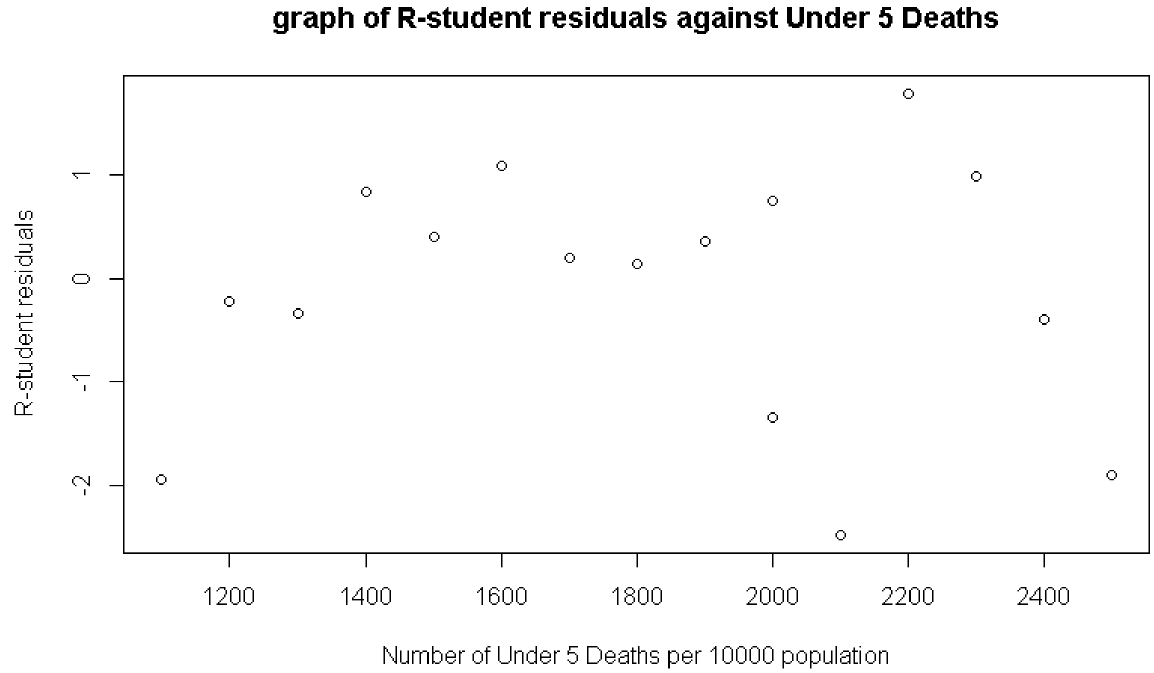
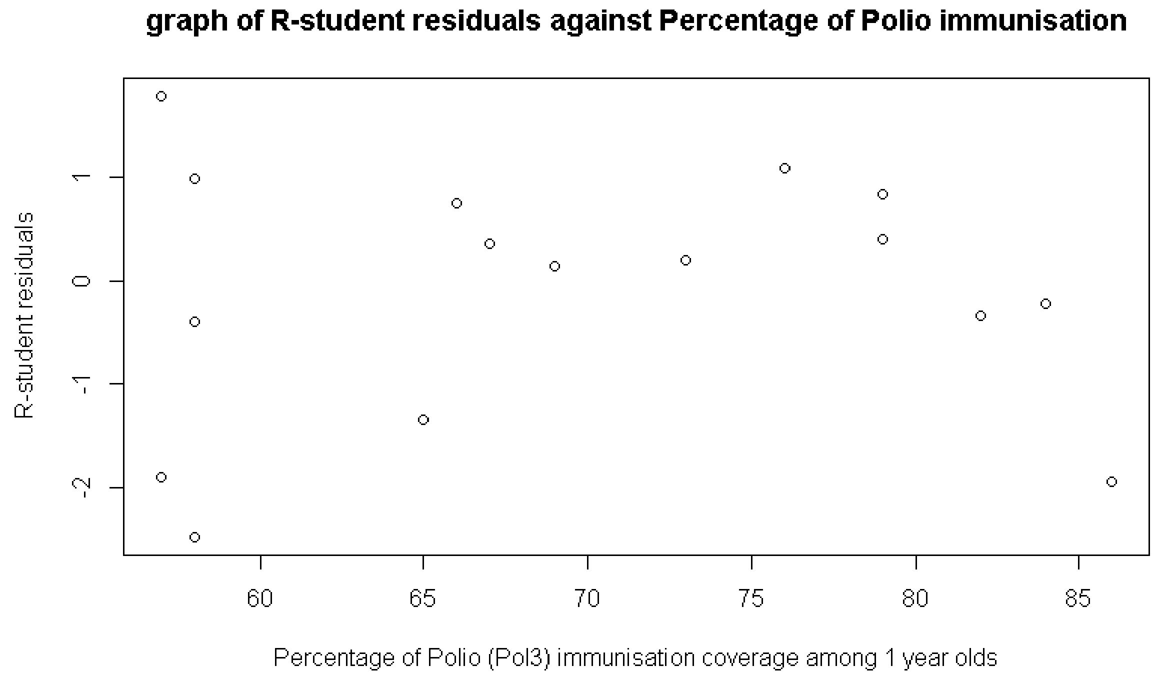
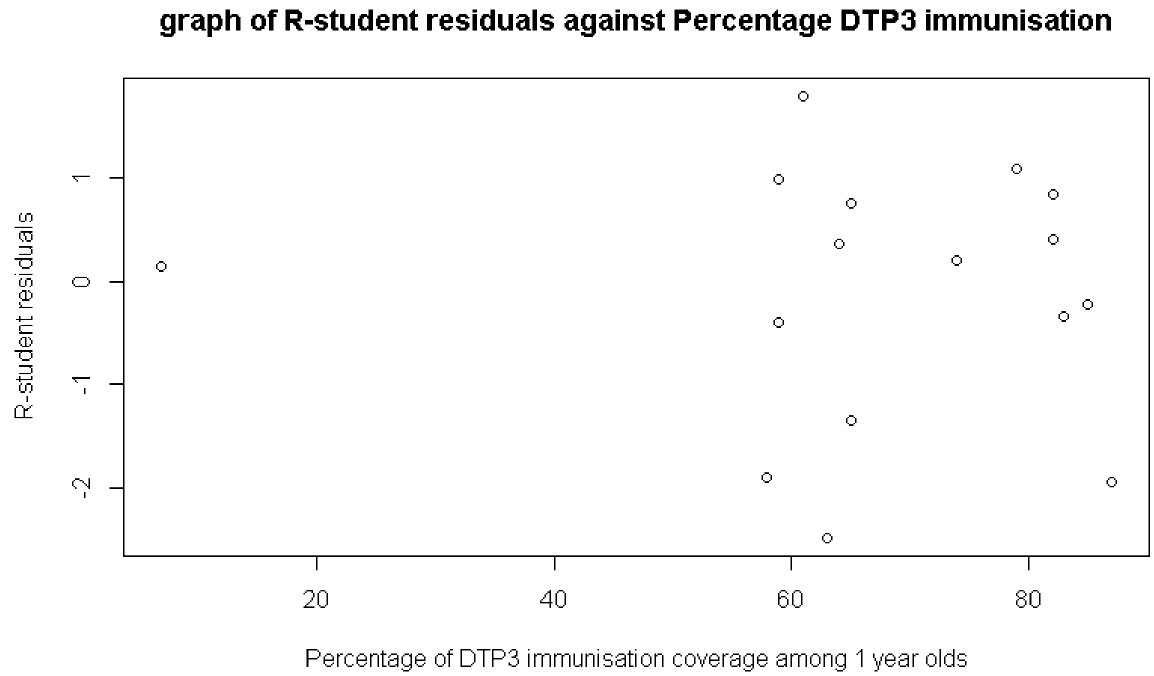


***Figure 4**: Plot of R-Student residuals against the Number of under 5 deaths per 10000 population*



***<u>Figure 5</u>: Plot of R-Student residuals against the percentage of Polio (Pol3) immunisation coverage among 1 year olds***



***<u>Figure 6</u>: Plot of R-Student residuals against the percentage of DTP3 immunisation coverage among 1 year olds***

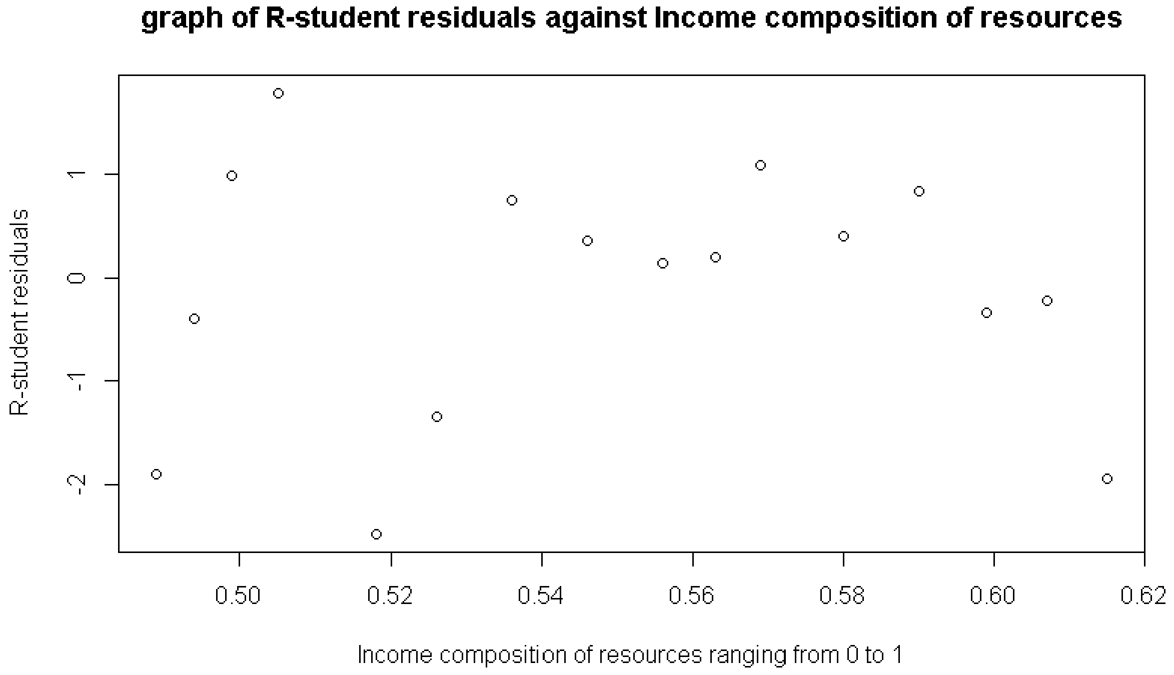


***Figure 7*: Plot of R-Student residuals against the Income composition of Resources ranging from 0 to 1**

## 3.4 The errors of our linear model approximately follow a normal distribution.

For our linear model, the normal probability plot is as shown in *Figure 3* below.

In *Figure 3*, we see that the distribution of the residuals of our model is approximately linear with few points showing a departure from linearity. This might be due to the presence of outliers in the data. Thus it can be said that the errors of our model approximately follows a normal distribution with slight deviations due to the presence of outliers.

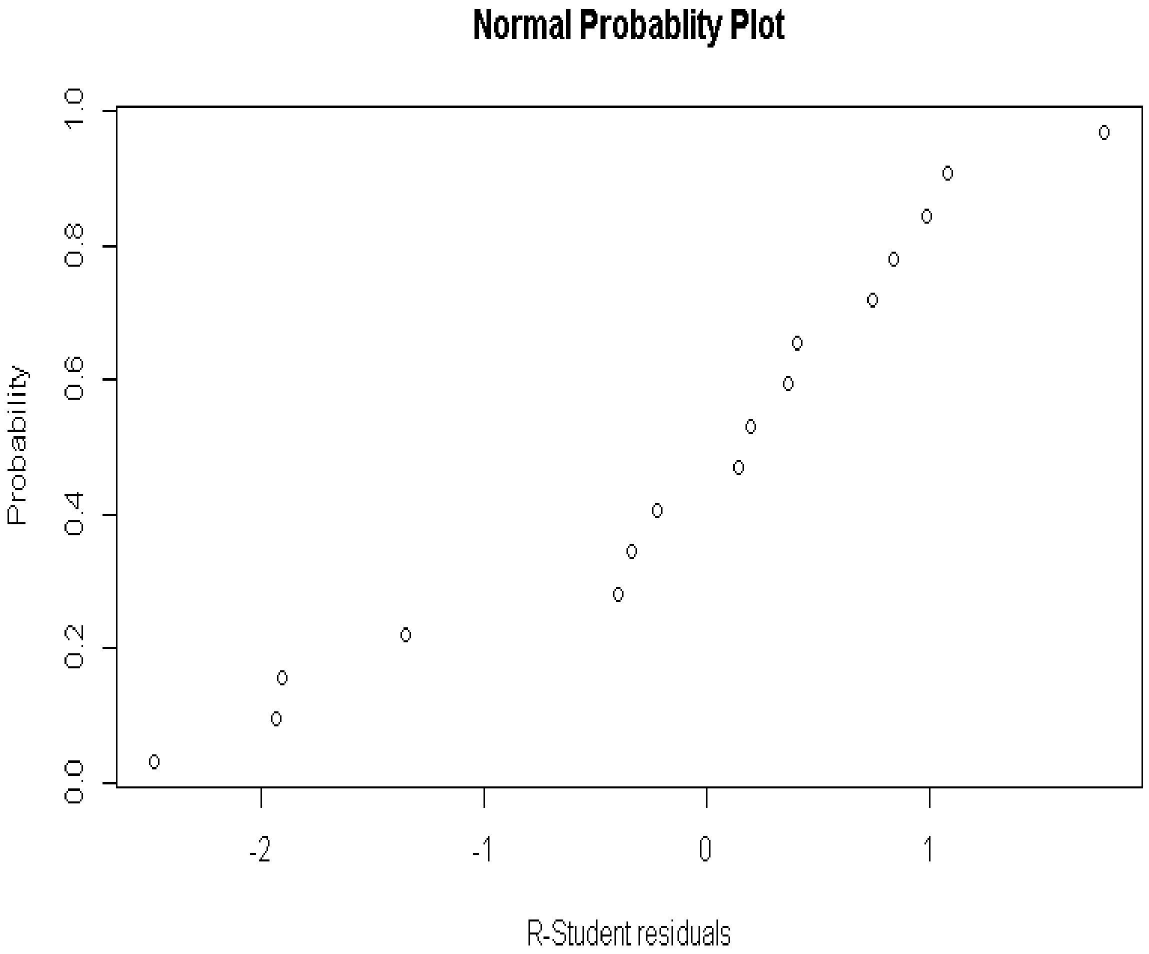


***Figure 8: The Normal Probability Plot***

## *3.5 The 8th observation is a leverage point and the 11th observation also shows mild leverage behaviour in our data.*

For our data, the hat values of all the observations are listed in *Table 4*.

From *Table 4*, we see that the 8th observation has a hat value equal to 0.990395, which is too high and it exceeds our cut off value which is $2p/n = 2 * 7/16 = 0.875$. ***Thus, the 8th observation is our leverage point.*** Also, from the hat values in *Table 4*, we notice that the hat value for the 11th observation is 0.623994, which is quite large compared to the remaining observations, though it does not exceed the mathematical cut off *2p/n*. This indicates that ***the 11th point might also show some leverage behaviour in our data.***

## ***3.6 Thus, the 1st, 12th and the 16th observations are significantly influential and the 8th, 11th and 13th observations have little influence on the values of the least square estimates $\widehat{\boldsymbol{\beta}}$ (according to Cook's Distance values).***

In my study, the cut off $4/(n-p)$ is used and any observation whose value of Cook's Distance is more than $4/(n-p) = 4/(16-7) = 4/9 = 0.4444$, is declared as an influential point. The Cook's Distance values are listed in *Table 4* below.

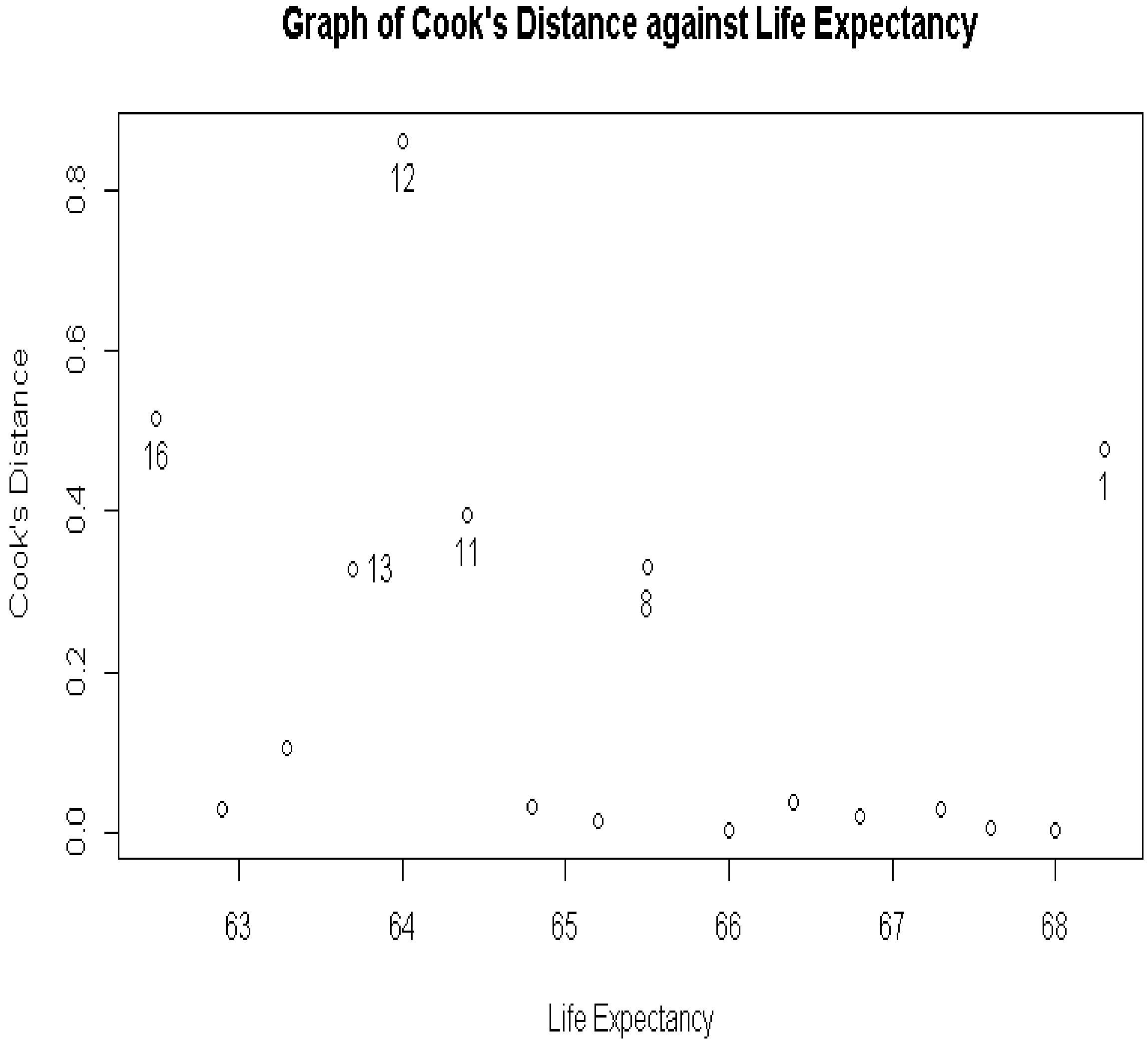
From *Table 4*, it is seen that the 1st, 12th and 16th observations have their Cook's Distance values 0.477668700, 0.861493081 and 0.515862284 respectively, which are all greater that the theoretical threshold 0.4444. ***Thus, the 1st, 12th and the 16th observations are significantly influential.*** It is also noticed that, the 8th, 11th and 13th observations also have a moderately high value of Cook's Distance, though not more than the theoretical threshold. Therefore, it can be said that ***the 8th, 11th and 13th observations have little influence on the values of the least square estimates $\widehat{\boldsymbol{\beta}}$.***

The Cook's Distance values are also plotted against the response variable *y* (Life Expectancy) in the *Figure 4*.

***Table 4: Hat values and Cook's Distance values***

| ***Observation No.*** | ***Hat values ($h_{ii}$)*** | ***Cook's Distance ($D_i$)*** |
|---|---|---|
| 1 | 0.5385678 | 0.477668700 |
| 2 | 0.2433184 | 0.002600942 |
| 3 | 0.1824704 | 0.003976061 |
| 4 | 0.2094654 | 0.027513637 |
| 5 | 0.4199278 | 0.018870770 |
| 6 | 0.1835102 | 0.037025122 |
| 7 | 0.3168494 | 0.002910155 |
| 8 | 0.9900395 | 0.330509393 |
| 9 | 0.4153888 | 0.014809881 |
| 10 | 0.2705486 | 0.030983924 |
| 11 | 0.6239964 | 0.395367196 |
| 12 | 0.6072256 | 0.861493081 |
| 13 | 0.4725606 | 0.327580210 |
| 14 | 0.4290199 | 0.104566897 |
| 15 | 0.5339763 | 0.028143779 |
| 16 | 0.5631349 | 0.515862284 |

From *Figure 4*, it is clear that the 1st, 12th and the 16th observations have a very high value of Cook's Distance while the 8th, 11th and 13th observations have a moderately high value of Cook's Distance compared to the rest of the observations. Thus***, the 1st, 12th and 16th observations are highly influential and need to be examined and removed if required while the 8th, 11th and 13th observations show less influence on the estimated parameters $\widehat{\boldsymbol{\beta}}$ and these observations may not be removed.***



***<u>Figure 9:</u> Graph of Cook's Distance against the Response variable***

## ***3.7 <u>Thus, the observations with maximum influence on the estimated parameters are the 1st, 12th and the 16th observations (based on DFBETAS Method)</u>***

For our data, the $DFBETAS_{j,i}$ values are listed in the following table:

***Table 5: Table of DFBETAS for the corresponding estimated parameters***

| ***Obs. No.*** | ***(Intercept) ($\beta_0$)*** | ***Infant Deaths ($\beta_1$)*** | ***BMI ($\beta_2$)*** | ***Under Five Deaths ($\beta_3$)*** | ***Polio immune (%) ($\beta_4$)*** | ***DTP3 immune (%) ($\beta_5$)*** | ***ICOR ($\beta_6$)*** |
|---|---|---|---|---|---|---|---|
| 1 | 0.0747 | -0.1164 | -1.1498 | -0.4465 | 0.1837 | 0.2707 | 0.9253 |
| 2 | 0.0018 | -0.0003 | -0.0435 | -0.0177 | 0.0079 | 0.0011 | 0.0349 |
| 3 | -0.0118 | 0.0358 | 0.0133 | -0.0072 | 0.0130 | -0.0432 | -0.0091 |
| 4 | -0.1967 | 0.1945 | -0.1060 | -0.0736 | 0.0206 | 0.0423 | 0.2311 |
| 5 | 0.1449 | -0.1838 | -0.0561 | 0.0182 | 0.0242 | 0.1109 | -0.0580 |
| 6 | 0.2435 | -0.1795 | -0.0365 | -0.0204 | 0.1182 | 0.0788 | -0.1739 |
| 7 | -0.0630 | 0.0957 | -0.0885 | -0.0633 | 0.0620 | -0.0143 | 0.1065 |
| 8 | 0.0652 | -0.0059 | -0.0045 | -0.0377 | 0.0773 | -1.0027 | -0.0378 |
| 9 | -0.1512 | -0.1381 | 0.0703 | 0.1882 | -0.2191 | 0.1841 | 0.1088 |
| 10 | -0.3484 | 0.2205 | -0.1207 | 0.0026 | 0.0461 | 0.0375 | 0.3413 |
| 11 | -1.1931 | -0.3790 | 1.2160 | 1.3814 | -1.0307 | 0.5220 | 0.0497 |
| 12 | -0.2115 | 0.4239 | 0.4461 | 0.2588 | 1.7630 | -1.1562 | -0.7721 |
| 13 | 0.45760 | 0.2130 | 0.4184 | -0.2473 | -0.3545 | -0.1962 | -0.6072 |
| 14 | 0.40036 | -0.4998 | 0.4958 | 0.3268 | -0.2496 | 0.1155 | -0.6640 |
| 15 | 0.10131 | -0.2783 | 0.0709 | 0.1245 | -0.2661 | 0.1692 | -0.0535 |
| 16 | 0.65118 | 0.7615 | -0.8992 | -1.3401 | 0.2466 | -0.6246 | 0.2987 |

Here, it can be seen that the 1st point shows significant effect on $\widehat{\boldsymbol{\beta}}_2$ and $\widehat{\boldsymbol{\beta}}_6$ while the 12th point shows significant effect on $\widehat{\boldsymbol{\beta}}_4$, $\widehat{\boldsymbol{\beta}}_5$ and $\widehat{\boldsymbol{\beta}}_6$ and the 16th point shows significant effect on $\widehat{\boldsymbol{\beta}}_0$, $\widehat{\boldsymbol{\beta}}_1$, $\widehat{\boldsymbol{\beta}}_2$, $\widehat{\boldsymbol{\beta}}_3$, and $\widehat{\boldsymbol{\beta}}_5$. This is because these 3 points are influential points and so their effect on the respective estimated parameters are quite remarkable.

Also, it is noted that the 11th observation also shows significant effect on $\hat{\beta}_0$, $\hat{\beta}_2$, $\hat{\beta}_3$ $\hat{\beta}_4$ and $\hat{\beta}_5$, as it has a tendency to show leverage behaviour though it is not a leverage value theoretically as its hat value does not exceed our theoretical cut off for a leverage point. Also, the 8th observation shows significant effect on $\hat{\beta}_5$. This might be because it is a leverage point with a high hat value of 0.9900395 and a moderately high Cook's Distance value.

It must also be noted that both 13th and 14th observations show significant effect on $\hat{\beta}_6$. These observations have their |DFBETAS| values exceeding the theoretical cut off given by $\frac{2}{\sqrt{n}} = \frac{2}{4} = 0.5$ but they exceed it by a very small extent. This might be as the Cook's Distance value for the 13th observation is moderately high and that of the 14th observation is slightly higher than the remaining observations. Hence their influence on the respective estimated parameters is not remarkable.

## ***3.8 <u>The 1st, 12th and 16th observations have a high influence on the corresponding fitted values and it might be necessary to remove them. However, the 8th, 11th and 13th observations have little influence on the fitted values and they need not be removed (based on DFFITS values).</u>***

For our data, the DFFITS values are given in *Table 6*.

From *Table 6*, it is seen that the values of $|DFFITS_i|$ for the 1st, 12th and 16th observations are too high and show high departure from the theoretical threshold $2\sqrt{p/n} = 2 \times \sqrt{7/16} = 1.3229$. However, the

8th, 11th and 13th observations are moderately high and show little departure from the theoretical threshold which is equal to 1.3229.

Therefore***, the 1st, 12th and 16th observations have a high influence on the corresponding fitted values and it might be necessary to remove them.*** However, ***the 8th, 11th and 13th observations have little influence on the fitted values and they need not be removed.***

## *3.9 The 1st, 12th, 13th and 16th observations degrade the precision of estimation while the other observations improve the overall precision of estimation(based on COVRATIO values).*

In our model, it is seen that the 2nd, 3rd, 5th, 7th, 8th, 9th and 15th observations exceed the cut off $1 \pm 3\,p/n = 1 \pm 3 \times 7/16$ which are equal to 2.3125 and 0 as $n < 3p$. However, all these observations *improve* the precision of estimation and so they are not removed. On the other hand the 1st, 12th, 13th and 16th observations degrade the precision of estimation. But, as the value of COVRATIO for the 13th observation is more closer to 1 than the remaining 3 observations and it is not a strong influential observation as given by Cook's Distance, it *is not* removed. The COVRATIO values for all the observations are tabulated in *Table 6*.

***Table 6: Table of DFFITS and COVRATIO values***

| **Observation number** | **DFFITS** | **COVRATIO** |
|---|---|---|
| 1 | -2.088 | 0.338 |
| 2 | -0.128 | 2.884 |
| 3 | -0.158 | 2.53 |
| 4 | 0.432 | 1.6 |
| 5 | 0.346 | 3.407 |
| 6 | 0.514 | 1.07 |
| 7 | 0.135 | 3.226 |
| 8 | 1.436 | 224.861 |
| 9 | 0.306 | 3.479 |
| 10 | 0.454 | 1.954 |
| 11 | -1.738 | 1.445 |
| 12 | -3.076 | 0.109 |
| 13 | 1.688 | 0.416 |
| 14 | 0.854 | 1.791 |
| 15 | -0.423 | 4.276 |
| 16 | -2.159 | 0.384 |

## *3.10 Our linear model is most improved when all the three influential points that are the 1st, 12th and 16th observations, are deleted.*

Thus, from all the above measures, it is seen that the 1st, 12th and 16th observations are the strong influence points and needs to be removed. So, these observations are removed one at a time, two at a time and then all at a time and the changes in the values of the estimated parameters are observed. On doing this, the following table is observed:

***Table 7: Table of estimated parameters when influential points are removed***

| | **Remove none obs** | **Remove only 1st obs** | **Remove only 12th obs** | **Remove only 16th obs** | **Remove 1st and 12th obs** | **Remove 1st and 16th obs** | **Remove 12th and 16th obs** | **Remove 1st, 12th and 16th obs** |
|---|---|---|---|---|---|---|---|---|
| $\hat{\beta}_0$ | 53.9113 | 53.5091 | 54.9495 | 50.3878 | 5.45E+01 | 50.4864 | 52.2034 | 51.9413 |
| $\hat{\beta}_1$ | 0.0002 | 0.0003 | -0.0002 | -0.0007 | -6.23E-05 | -0.0004 | -0.0008 | -0.0006 |
| $\hat{\beta}_2$ | -0.0246 | 0.2999 | -0.1395 | 0.2305 | 1.17E-01 | 0.4820 | 0.0661 | 0.2916 |
| $\hat{\beta}_3$ | -0.0019 | -0.0014 | -0.0022 | -0.0002 | -1.80E-03 | -2.9E-05 | -0.0009 | -0.0006 |
| $\hat{\beta}_4$ | -0.0042 | -0.0080 | -0.0374 | -0.0093 | -3.43E-02 | -0.0120 | -0.0354 | -0.0326 |
| $\hat{\beta}_5$ | 0.0018 | 0.0013 | 0.0038 | 0.0030 | 3.11E-03 | 0.0024 | 0.0043 | 0.0036 |
| $\hat{\beta}_6$ | 27.8390 | 18.2384 | 35.1428 | 24.7231 | 2.69E+01 | 16.7170 | 31.621 | 24.1323 |

From the above table, it is seen that ***no observation produces a remarkable change on*** $\widehat{\boldsymbol{\beta}}_{\mathbf{0}}$. While, on removing the 16th observation the value of $\hat{\beta}_1$ undergoes a high change of 421.9923% of its value when no observation is removed. Also, when the 16th observation is removed along with the 1st, 12th or both other observations, it is seen that the percentage changes are 311.835865%, 484.17666%, 387.2413% respectively, which are close to the percentage change produced by the 16th observation alone. ***Thus, the 16th observation has a great effect on*** $\widehat{\boldsymbol{\beta}}_{\mathbf{1}}$***.***

Also, on removing the 1st observation, the value of $\hat{\beta}_2$ changes as much as 1315.086% of its value when no observation is removed. When the 1st observation is removed along with the 12th observation, it is seen that the value of $\hat{\beta}_2$ changes by 572.30843% which is much less than the change produced by removing only the

1st observation alone. Thus, the 12th observation also produces much change in the value of $\hat{\beta}_2$ but in the opposite direction as produced by 1st observation. The 16th observation also produces a change of 1033.8067% and when both 1st and 16th observations are removed the value of $\hat{\beta}_2$ experiences even larger change of 2052.851772% , and when 16th and 12th observations are deleted the change produced is 484.17666% which is much closer to the change produced by the 12th observation alone. Finally, when all the three observations are removed, the change experienced by $\hat{\beta}_2$ is 387.2413% which is somewhat close to the change produced by the 12th observation alone. ***Thus, on $\hat{\boldsymbol{\beta}}_2$, all the three influential points produce remarkable change with the influence of the 12th observation in one direction and the influence of the 1st and 16th observations in the other direction.***

While it is seen that ***no observation produces a remarkable influence on$\hat{\boldsymbol{\beta}}_3$***, the value of $\hat{\beta}_4$ is changed by 773.6079% on the removal of the 16th observation. Also, when the 12th observation is removed along with the 1st, 16th or both other observations, it is seen that the percentage changes are 700.405766%, 728.11307%, 662.7644% respectively, which are close to the percentage change produced by the 12th observation alone. ***Thus, the 12th observation has most influence on the value of $\hat{\boldsymbol{\beta}}_4$, followed by the 16th observation.***

It is also noticed that the 12th observation produces a change of 106.6564% on $\hat{\beta}_5$ which is however neutralised by the low offect of the 1st observation on $\hat{\beta}_5$ which is about 27.39055% when both 1st and the 12th observations are removed. When the 12th observation is removed along with the 16th observation and along with both 16th and 1st observations, it is seen that the changes are 134.69823% and

96.55822% respectively which are quite close to the percentage change produced by the 12th observation alone. ***Thus, the 12th observation has a significant effect on the value of $\widehat{\boldsymbol{\beta}}_5$.***

***However, it is seen that no observation produces any significant change on the value of $\widehat{\boldsymbol{\beta}}_6$.***

Also, on checking the Mean Squared Residuals when these outliers are removed and comparing the results with the true value of variance of pure error, the following table is obtained:

***Table 8: Table of change of mean squared errors when outliers are removed***

| | |
|---|---|
| ***Variance of pure error*** | 0.003975811 |
| $MS_{res}$ ***when no observation is removed*** | 0.008828 |
| $MS_{res}$ ***when only 1st observation is removed*** | 0.00677 |
| $MS_{res}$ ***when only 12th observation is removed*** | 0.005627 |
| $MS_{res}$ ***when only 16th observation is removed*** | 0.0068401 |
| $MS_{res}$ ***when both 1st and 12th observations are removed*** | 0.004681446 |
| $MS_{res}$ ***when both 1st and 16th observation observations are removed*** | 0.00510629 |
| $MS_{res}$ ***when both 12th and 16th observations are removed*** | 0.004698 |
| $MS_{res}$ ***when all the 1st, 12th and 16th observations are removed*** | 0.0037092 |

***Thus, it is seen that on removing all the 1st, 12th and 16th observations, the mean squared residuals ($MS_{res}$) is minimised and***

***is closest to the variance of pure error. Thus our linear model is most improved when all the three influential points are deleted.***

# Conclusion

From the above results, it is noticed that the 8th observation is not an influential point though it has a high hat value and is a strong leverage point. This is as the 8th point is located far away from the rest of the observations in the x space, but lies approximately close to the fitted regression line passing through the rest on the observations. Thus it does not affect the estimated parameters significantly. It only has a moderate influence on the estimated parameters and the fitted values.

On the other hand, the 1st, 12th and 16th observations are stronger influential points than the 8th observation, though they have a comparatively less hat value than the 8th observation. This is because these observations are not too far from the remaining observations in the x space, but they are located in such a position that they pull the regression line towards themselves and thus affect the estimated parameters more significantly.

On removing the major influential points, that are the 1st, 12th and 16th observations, the mean squared residuals is minimised, indicating an improvement in our fitted linear model.

# Reference

1) *Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining (fifth edition): Introduction To Linear Regression Analysis, Wiley Series of probability and statistics.*

2) *Alvin C. Rencher and G. Bruce Schaalje (second edition): Linear Models In Statistics, John Wiley*

3) *https://www.kaggle.com/kumarajarshi/life-expectancy-who*

# Acknowledgement

It is my utmost pleasure to present the project on “Multiple Linear Regression: Model Checking And Diagnostics”.

Any Achievement cannot be made without the advice and guidance of noble minds. My project also could be successfully presented only with the valuable suggestions of a mind with high experience. Thus, the project can never be complete without the mention of the driving force that enhanced my hard work, determination and dedication.

I want to express my sincere gratitude to my supervisor, Professor Tanusree Banerjee, for guiding me throughout this project. She motivated me to work on this project and learn the application of our *Core Course 12: Linear Models and Regression*.

I also thank all the faculty members of Bethune College Kolkata, for their support in the various stages of study for this project.

*Thank You*

***Bidhisha Ghosh***

# Appendix

The data used is as follows:

***Table 1: The Life Expectancy Data***

| *Sl. No.* | *Year* | *Life Expectancy* | *Infant Deaths (per 10000 population)* | *BMI* | *Under-five deaths (per 10000 population)* | *Polio immune (%)* | *DTP3 immune (%)* | *ICOR (ranging from 0 to 1)* |
|---|---|---|---|---|---|---|---|---|
| 1 | 2015 | 68.3 | 910 | 18.7 | 1100 | 86 | 87 | 0.615 |
| 2 | 2014 | 68 | 957 | 18.1 | 1200 | 84 | 85 | 0.607 |
| 3 | 2013 | 67.6 | 1000 | 17.5 | 1300 | 82 | 83 | 0.599 |
| 4 | 2012 | 67.3 | 1100 | 17 | 1400 | 79 | 82 | 0.59 |
| 5 | 2011 | 66.8 | 1100 | 16.4 | 1500 | 79 | 82 | 0.58 |
| 6 | 2010 | 66.4 | 1200 | 15.9 | 1600 | 76 | 79 | 0.569 |
| 7 | 2009 | 66 | 1300 | 15.4 | 1700 | 73 | 74 | 0.563 |
| 8 | 2008 | 65.5 | 1300 | 14.9 | 1800 | 69 | 7 | 0.556 |
| 9 | 2007 | 65.2 | 1400 | 14.4 | 1900 | 67 | 64 | 0.546 |
| 10 | 2006 | 64.8 | 1500 | 13.9 | 2000 | 66 | 65 | 0.536 |
| 11 | 2005 | 64.4 | 1500 | 13.5 | 2000 | 65 | 65 | 0.526 |
| 12 | 2004 | 64 | 1600 | 13 | 2100 | 58 | 63 | 0.518 |
| 13 | 2003 | 63.7 | 1700 | 12.6 | 2200 | 57 | 61 | 0.505 |
| 14 | 2002 | 63.3 | 1700 | 12.2 | 2300 | 58 | 59 | 0.499 |
| 15 | 2001 | 62.9 | 1800 | 11.8 | 2400 | 58 | 59 | 0.494 |
| 16 | 2000 | 62.5 | 1800 | 11.4 | 2500 | 57 | 58 | 0.489 |

https://www.kaggle.com/kumarajarshi/life-expectancy-who