Life Expectancy Prediction through Model building using R software

Siddhi Gosavi

INTRODUCTION

Life expectancy is a statistical measure of the average time an organism is expected to live, based on the year of its birth, its current age and other demographic factors including gender. It is a very important factor for development of a country. It is also useful in the insurance companies to plan their medical insurances.

The data-set aims to answer the following key questions:

• Do various predicting factors really affect the Life expectancy?
• What are the predicting variables actually affecting the life expectancy?
• Should a country having a lower life expectancy value (<65) increase its healthcare expenditure in order to improve its average lifespan?
• By how much does the Adult mortality rates affect life expectancy?
• Does Life Expectancy have positive or negative correlation with eating habits, lifestyle, exercise, smoking, drinking alcohol etc.
• What is the impact of schooling on the lifespan of humans?
• Does Life Expectancy have positive or negative relationship with drinking alcohol?
• Do densely populated countries tend to have lower life expectancy?
• What is the impact of Immunization coverage on life Expectancy?

DATA DESCRIPTION

The data was collected from WHO and United Nations website. In this dataset there are 22 variables and 549 data points. The description of the variables is given below:

1. **Country:** Name of Country.
2. **Status:** Developed or developing status
3. **Life expectancy:** Expected number of years the person will survive
4. **Adult Mortality:** Death rate for people between the ages 15 to 60
5. **Infant Death:** Number of Infant Deaths per 1000 population
6. **Alcohol:** Alcohol, recorded per capita consumption
7. **Percentage expenditure:** Expenditure on health as a percentage of GDP per capita
8. **Hepatitis B:** Hepatitis B immunization coverage among 1-year-olds
9. **Measles:** Number of reported cases of measles per 1000 population
10. **BMI:** Average Body Mass Index of entire population
11. **Under 5 death:** Number of under-five deaths per 1000 population
12. **Polio:** Polio immunization coverage among 1-year-olds
13. **Total Expenditure:** Government expenditure on health as a percentage of total government expenditure.
14. **Diphtheria:** Diphtheria immunization coverage among 1-year-olds
15. **HIV/AIDS:** Deaths per 1 000 live births HIV/AIDS
16. **GDP:** Gross Domestic Product per capita
17. **Population:** Population of the country
18. **Thinness 10-19 years:** Prevalence of thinness among children and adolescents for Age 10 to 19
19. **Thinness 5-9 years:** Prevalence of thinness among children for Age 5 to 9
20. **Income composition of resources:** Human Development Index in terms of income composition of resources
21. **Schooling:** Number of years of Schooling

OBJECTIVE

- To fit a linear model of life expectancy data
- To find the variables that significantly affect the response variable which is Life Expectancy
- To predict the values of the response variable- Life Expectancy

Deep Dive into objective

Response variable is Life Expectancy. Using the variables of Life Expectancy data, predictions on response variables are. The model is built using Multiple Linear Regression. and find the effect of variables on the response variable. Assumptions are checked pre and post fitting of the model. The model adequacy is examined at the closing stages.

Process Flow

1. Pre-Model Assumptions
    - Checking linearity and normality of independent variables

2. Data Preparation
    - Numeric Conversion of variable
    - Missing Value Treatment
    - Scaling
    - Outlier Treatment
3. Model Building- Train Data

4. Post Model Assumptions
    - Multicollinearity

5. Variable Selection
    - Stepwise Regression

6. Variable transformation
    - Box Tidwell
    - Boxcox Transformation
7. Model Adequacy checking

8. Checking Influential and leverage points
    - Cooks Distance Plots
    - Removing insignificant variables

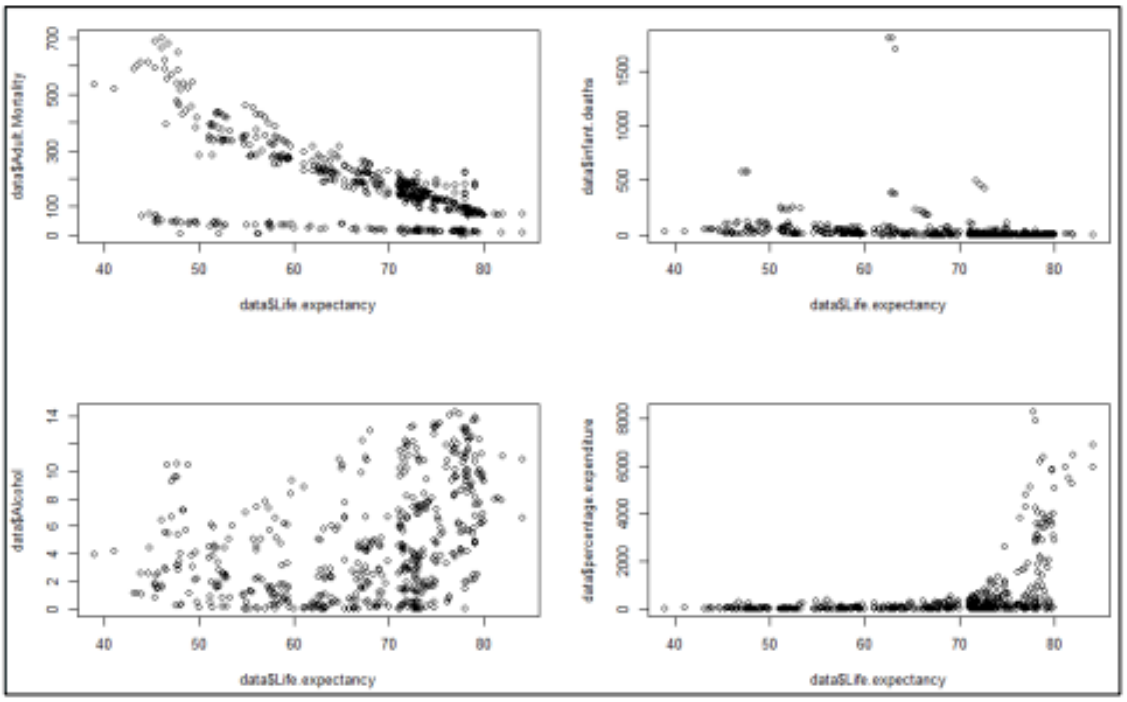
9. Fitting-Test Data

Analysis

1. Pre-Model Assumptions

- **Linearity of variables**

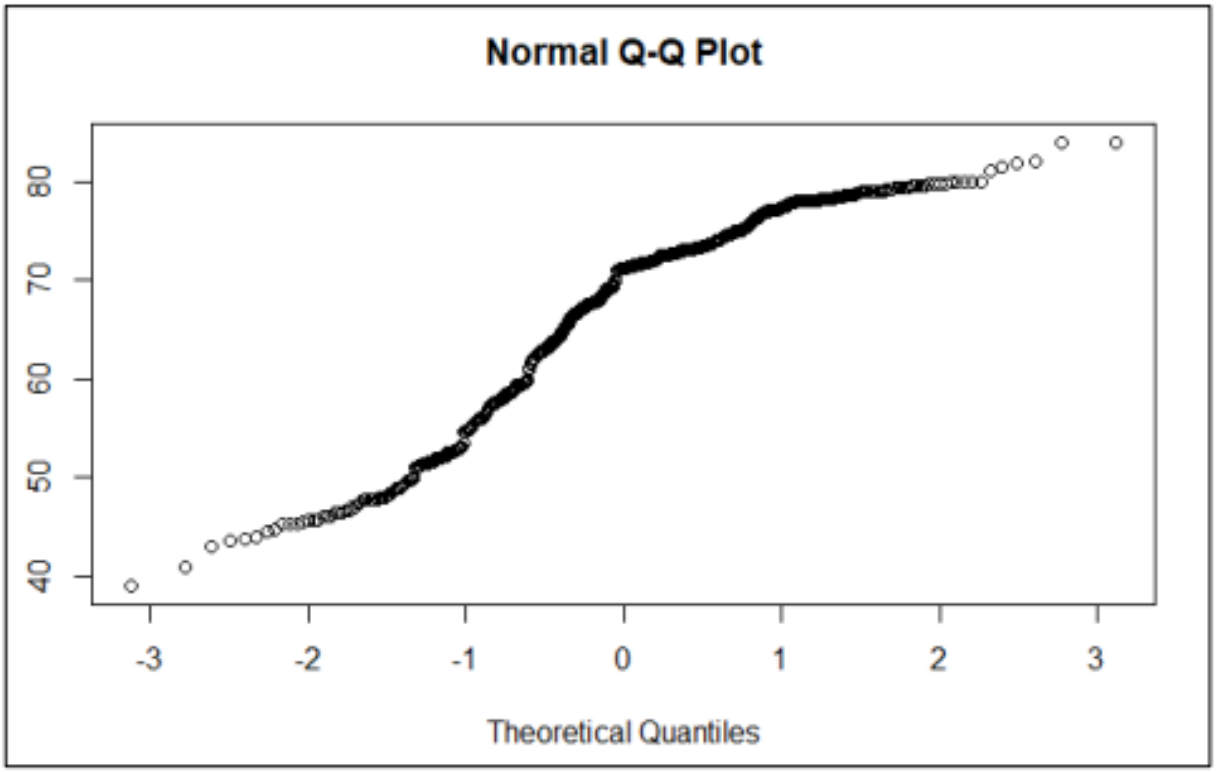
  At the initial stage, linear relationship of response variables and explanatory variables were examined. Simple scatter plot was used for this



Inference:
The analysis was taken forward with all variables as linearity assumption was satisfied.

• **Normality of variables:**

One of the assumptions for the regressors before fitting the model is that it tends to follow normal distribution. The method used for this is qqplot.



Inference:
Almost all regressors and response variable tend to follow Normality.

2. Train-Test Validation

The data is divided into two parts for the purpose of model building which are train and test. The 'train' data is trained in order to fit the model. In later stages this model is fitted on the 'test' data to test the accuracy of the model. The dataset is partitioned into 70-30, where 70% is 'train' and 30% is 'test' data.

```
set.seed(123)
idx = sample(1:nrow(data),ceiling(0.7*nrow(data)))
train = data[idx,]
test = data[-idx,]
```

3. Data Preparation

- **Missing Value Treatment**
  Missing values were observed in the dataset. They were treated by replacing the value by Median. Median was used since it doesn't change the distribution of the data.

```
train$Adult.Mortality= ifelse(is.na(train$Life.expectancy),
median(train$Adult.Mortality,na.rm = TRUE), train$Adult.Mortality)

train$infant.deaths=ifelse(is.na(train$infant.deaths),
median(train$infant.deaths, na.rm=TRUE),train$infant.deaths)

train$Alcohol=ifelse(is.na(train$Alcohol),median(train$Alcohol,na.rm=TRUE),train$Alcohol)

train$Measles=ifelse(is.na(train$Measles),median(train$Measles,na.rm=TRUE),train$Measles)
```

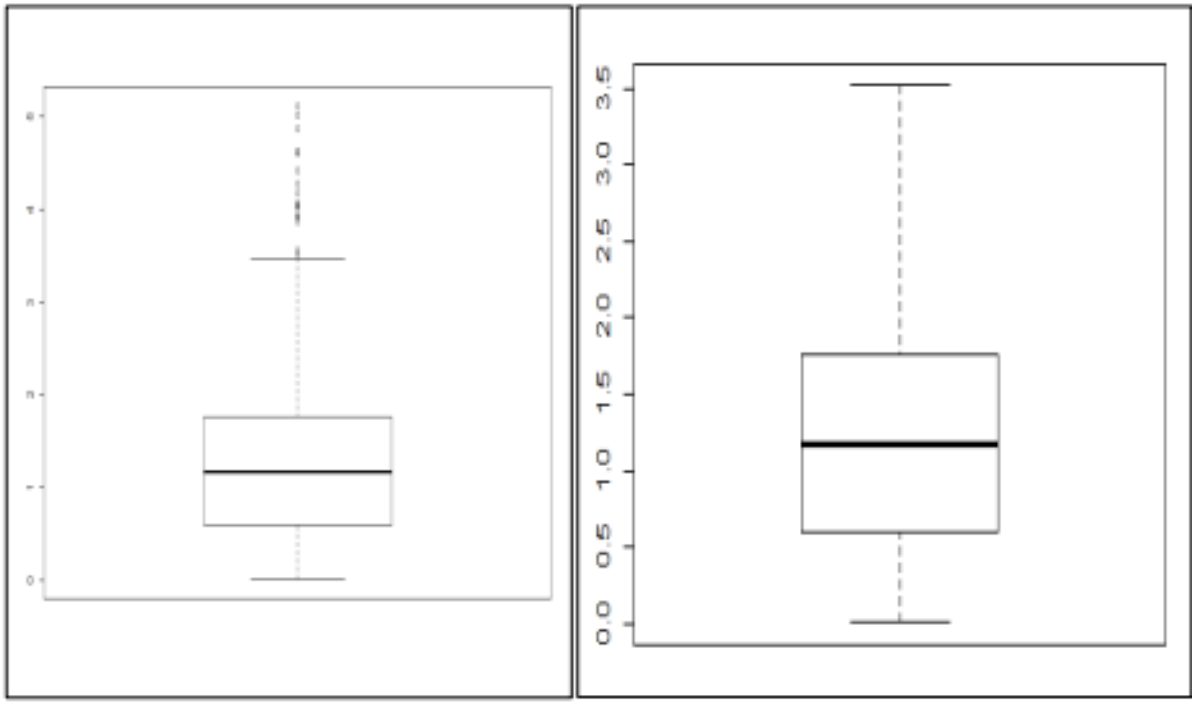
- **Scaling**
  There is lot of variation in the data in terms of scale and location, which might affect the final model. Hence, it becomes necessary to bring them in a certain range.

```
train$Adult.Mortality<-scale(train$Adult.Mortality, center=F,
scale=sd(train$Adult.Mortality))

train$infant.deaths<-scale(train$infant.deaths,center=F,
scale=sd(train$infant.deaths))

train$Alcohol<-scale(train$Alcohol,center=F,scale=sd(train$Alcohol))

train$percentage.expenditure<-
scale(train$percentage.expenditure,center=F,scale=sd(train$
percentage.expenditure)
```

- **Outlier Treatment**

Outlier is an observation point that is distant from other points. Data modelling requires treatment of outliers as it may give misleading results. Here, Boxplot is used to detect the outliers.

```
boxplot(train$Adult.Mortality)
Adult.Mortality_LL=quantile(train$Adult.Mortality,0.25)- 1.5*IQR(train$Adult.Mortality)
Adult.Mortality_UL=quantile(train$Adult.Mortality, 0.75) +1.5*IQR(train$Adult.Mortality)
Adult.Mortality_LL=0
train$Adult.Mortality=ifelse(train$Adult.Mortality>Adult.Mortality_UL,
Adult.Mortality_UL, ifelse(train$Adult.Mortality<Adult.Mortality_LL,
Adult.Mortality_LL,train$Adult.Mortality))
```



Inference:
Outliers are detected and treated.
The variables having numerous outliers were removed as it may influence the model.
For the rest of the outliers are replaced by upper and lower limits according to the placement of the outliers.

4. Model Building

The model is built after checking the pre-assumptions and Data preparation.

```
mdl=lm(data=train,Life.expectancy~Status+Adult.Mortality+Alcohol+BMI+Polio+
       Total.expenditure+Diphtheria+thinness..1.19.years+thinness.5.9.years+
        Income.composition.of.resources+Schooling)
summary(mdl)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-22.1005 -2.8175  0.3946  3.3067 13.2783
Coefficients:
                              Estimate Std.  Error    t value  Pr(>|t|)
(Intercept)                     50.8178    1.8492  27.482    < 2e-16 ***
Status                          1.9986     1.0035   1.992     0.0471 *
Adult.Mortality                -3.2934     0.3561  -9.247    < 2e-16 ***
Alcohol                        -0.2570     0.3978  -0.646     0.5186
BMI                             1.9966     0.4193   4.762    2.75e-06 ***
Polio                           2.2606     0.5562   4.064    5.88e-05 ***
Total.expenditure              -0.5560     0.3197  -1.739     0.0828 .
Diphtheria                      0.9735     0.4892   1.990     0.0473 *
thinness..1.19.years           -1.3070     0.8762  -1.492     0.1366
thinness.5.9.years              0.7344     0.9073   0.809     0.4188
Income.composition.             1.0027     1.7285   0.580     0.5622
of.resources
Schooling                       3.0484     0.5154   5.914    7.55e-09 ***
---
Residual standard error: 5.508 on 373 degrees of freedom
Multiple R-squared:  0.7108, Adjusted R-squared:  0.7023
F-statistic: 83.36 on 11 and 373 DF,  p-value: < 2.2e-16
```

Inference:

The Value of adjusted R square is 0.7023. This value indicates that 70.23% variation in the response variable is explained by explanatory variables.

5. Multicollinearity

Multicollinearity occurs when independent variables in a regression model are correlated. It reduces the precision of the estimated coefficients which weakens the statistical power of the regression model. VIF(Variance Influential Factor) is used to check multicollinearity.

```
vif(mdl)
```

Status	Adult.Mortality
1.921824	1.336183
Alcohol	BMI
2.002365	2.225125
Polio	Total.expenditure
2.631057	1.206535
Diphtheria	thinness..1.19.years
2.440882	7.934120
thinness.5.9.years	Income.composition.of.resources
8.074314	2.387211
Schooling	
3.044717	

Inference:

The VIF values for all regressor are less hence there is no autocorrelation between the regressors.

6. Variable Selection

Sub setting of variables is done using **Stepwise Regression** method. Stepwise Regression is a combination of forward and backward technique used for removing variables which are not significant for the model. Here, significant regressors and found out and model is fitted only on them.

```
step(mdl,direction = 'both')
mdl_1=lm(data=train,Life.expectancy~Status+Adult.Mortality+BMI+Polio+
         Total.expenditure+Diphtheria+Schooling)
summary(mdl_1)
```

```
Initial Step: AIC = 1325.64
Step:  AIC=1321.3
Life.expectancy ~ Status + Adult.Mortality + BMI + l + Total.expenditure +
   Diphtheria + thinness..1.19.years + Schooling
                          Df  Sum of Sq  RSS   AIC
<none>                                  11366 1321.3
- thinness..1.19.years     1    72.62   11439 1321.8
+ thinness.5.9.years       1    23.47   11343 1322.5
+ Alcohol                  1    18.19   11348 1322.7
+ Income.composition
.of.resources              1    13.56   11353 1322.8
- Status                   1   120.79   11487 1323.4
- Diphtheria               1   126.46   11493 1323.6
- Total.expenditure        1   126.81   11493 1323.6
- Polio                    1   521.68   11888 1336.6
- BMI                      1   677.54   12044 1341.6
- Schooling                1  1973.86   13340 1380.9
- Adult.Mortality          1  2684.11    14051 1400.9
```

Inference:

From the stepwise regression, we can understand which are the significant regressors.

The model is again fitted on the remaining variables and the R square is 0.7034.

7. Variable transformation

The transformation of variables is used in order to increase the accuracy of the model. Here **Box Tidwell** transformation is used. This method represents an iterative approach and gives the value of lambda. The variable is then transformed by taking its power as the value of lambda. The model is further fitted using these transformed variables.

```
 boxTidwell(train$Life.expectancy~train$Adult.Mortality)
train$Adult.Mortality1=train$Adult.Mortality^(2)
boxTidwell(train$Life.expectancy~train$BMI)
train$BMI1=train$BMI^(2)
boxTidwell(train$Life.expectancy~train$I)
train$polio1=train$I^(4)
boxTidwell(train$Life.expectancy~train$Schooling)
train$Schooling1=train$Schooling^(2)
```

Inference:

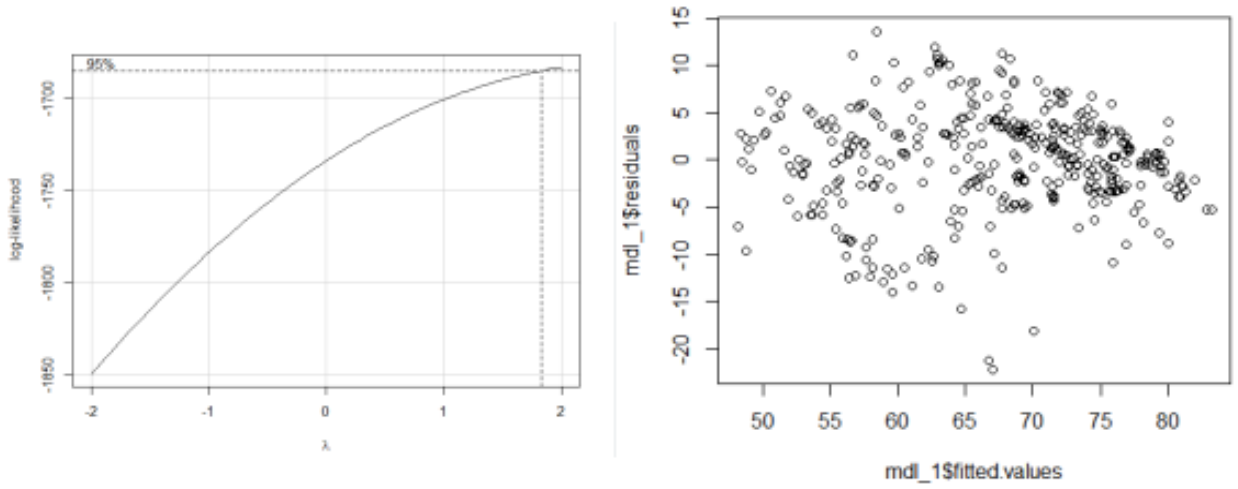
After performing the Box Tidwell transformations on the variables, the values of lambda were found out. The regressor variables were then transformed.

Box Cox Transformation

Box Cox transformation is a way to transform dependent variables. The model fit gets better if the response variable is transformed correctly.

Here, Box Cox transformation was applied on the response variable the Life expectancy to make the model better.

```
boxcox(mdl_1)
plot(mdl_1$fitted.values,mdl_1$residual
b<-boxcox(mdl_1)
 plot(mdl_1$fitted.values,mdl_1$residuals)
b$x[which.max(b$y)] train$Life.expectancy1=train$Life.expectancy^2
```



Inference:

After performing the Boxcox transformations on the model, the values of lambda were found out. The response variables were then transformed.

Model after the transformation

After all the transformations done on the regressor and the response variable, the model was fitted to these variables to check for the coefficients and adjusted R square.

```
mdl_2=lm(formula = Life.expectancy1 ~ Status + Adult.Mortality1 + BMI1 + polio1 + Diphtheria + Schooling1, data = train)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-2481.57 -361.50   62.26  379.99 1500.04

Coefficients:
                 Estimate Std. Error  t value  Pr(>|t|)
(Intercept)     3295.474     130.319  25.288  < 2e-16 ***
Status           189.582     100.671   1.883   0.0604 .
Adult.Mortality1 -156.028     11.900  -13.112  < 2e-16 ***
BMI1              53.805      11.592   4.641  4.78e-06 ***
l1                 3.728       0.832   4.481  9.86e-06 ***
Diphtheria       109.128      56.099   1.945   0.0525 .
Schooling1        77.263       8.804   8.776  < 2e-16 ***
--
Residual standard error: 627 on 378 degrees of freedom
Multiple R-squared:  0.767,   Adjusted R-squared:  0.7633
F-statistic: 207.4 on 6 and 378 DF,  p-value: < 2.2e-16
```

Inference:

The R square in this transformed model is 0.7633 .

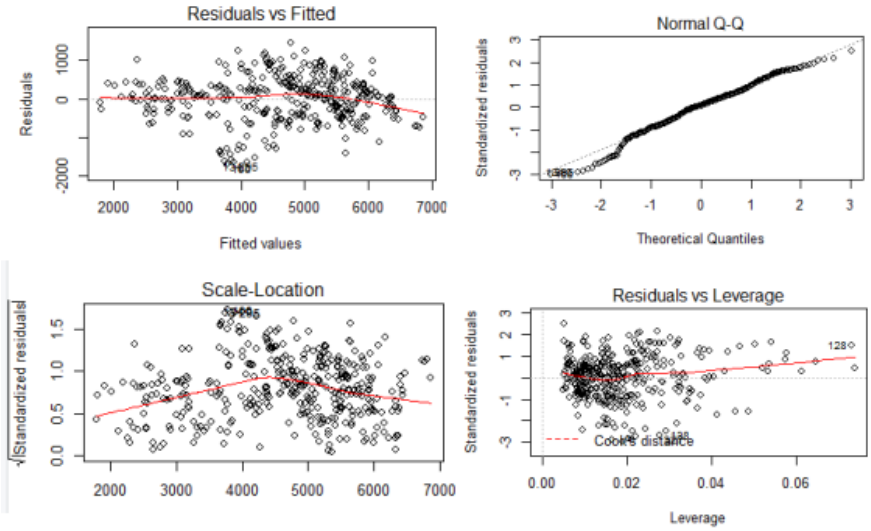
8. Model Adequacy

The fitting of linear regression model, estimation of parameters are based on following major assumptions:

1. The relationship between the study variable and explanatory variables is linear, atleast approximately.
2. The error term has zero mean.
3. The error term has constant variance.
4. The errors are uncorrelated.
5. The errors are normally distributed.

The validity of these assumption is needed for the results to be meaningful.

Model Adequacy plots are plotted to check these assumptions.

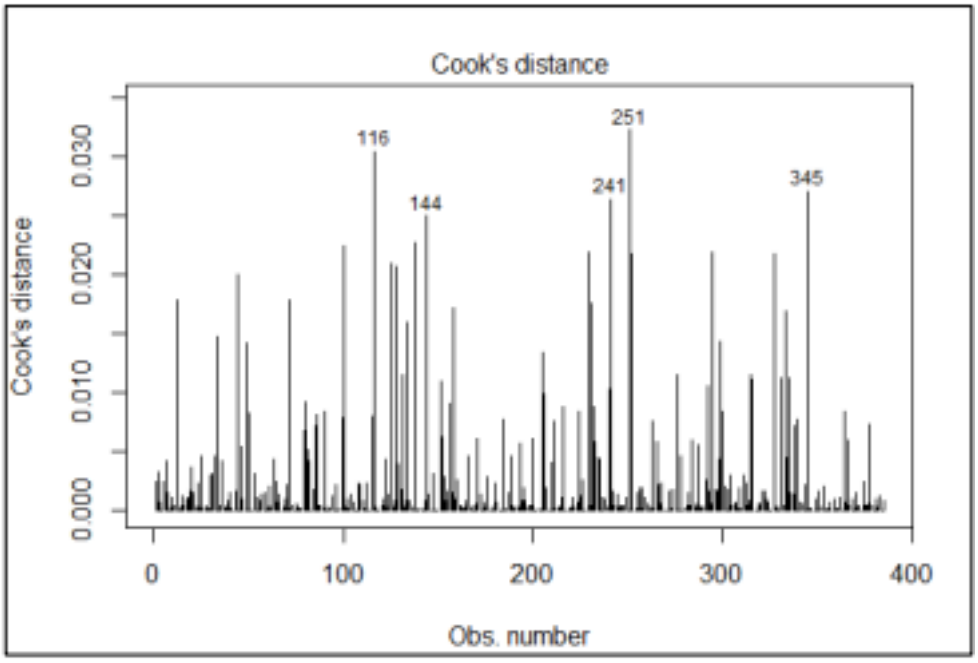


Inference:

• From the 1st plot, we can see that the residuals are independently distributed.
• From the 2nd plot, the assumption about the normality of the errors is followed.
• From the 3rd plot, it is observed that the assumption about the constant variance of the error is not satisfies even after the response transformations.
• In the 4th plot, influential points are observed, which are treated in the next part.

9. Influential and Leverage points

From the model adequacy the presence of leverage and influential points is observed. Further analysis on this variables is made using Cooks distance method.



Inference:

The observation numbers 116,114,241,251,345 indicate the influential points. These observation and corresponding rows are completely deleted from the data. The model is again fitted with the remaining data points. This is a iterative method which is terminated at a point where adjusted R square starts decreasing.

```
plot(mdl_2,4,id.n = 5)
 train1<-train[-c(116,144,241,251,345),]
mdl_2=lm(data=train1,Life.expectancy1~Status+Adult.Mortality1+BMI1+
          polio1+Diphtheria+Schooling1)
```

Inference:

After the model 3 was fitted the R square value was 0.7633.

10. Model Fitting-Test data

To validate the model it is fitted on test data. The test data is prepared in the same fashion as the train data. Further the values of the response variable are predicted for the test data.

```
test$PredVal <- predict(object = mdl_8, newdata = test)
```

Given are the values of the predicted values.

Life.expectancy	PredVal1
46.0	25.56480
75.0	73.40371
74.7	72.88852
74.4	89.39665
59.3	61.07867
57.7	83.59641
51.6	69.06454
51.5	83.37974
79.4	64.17159
77.3	85.61289
72.2	75.60518

PREDICTIVE MODEL

Equation of the model is

Life expectancy1= 3295.47 + 189.58*(Status) – 156.02(Adult Mortality1) + 53.8*(BMI1) + 3.72*(Polio1) + 109.12*(Diptheria) + 77.26*(Schooling)

The Adjusted R square is 0.7633

LIMITATION AND SCOPE

- The effective variables have removed due to humongous outliers.
- Due to lack of exposure to advanced knowledge, the further scope of use of higher techniques apart from linear regression was ruled out
- In real time it is not possible to satisfy all the assumptions.

BIBLIOGRAPHY

http://www.sthda.com/english/articles/39-regression-model-diagnostics/161-linear_regression assumptions-and-diagnostics-in-r-essentials/

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

https://stackoverflow.com/questions/8211318/multiple-regression-analysis

https://www.ibm.com/analytics/learn/linear-regression?S_PKG=-&cm_mmc=Search_Bing-_-Hybrid+Cloud_Business+Analytics-_-WW_IN-_-+linear++regression_Broad_-&cm_mmca1=000000OA&cm_mmca2=10000380&cm_mmca7=116073&cm_mmca8=kwd-81776279707291:loc-90&cm_mmca9=60323ca1-11b4-4d73-a6cc_c55d3527a5b5&cm_mmca10=81776224001569&cm_mmca11=p&mkwid=60323ca1-11b4-4d73-a6cc_c55d3527a5b5|479|452412&cvosrc=ppc.bing.%2Blinear%20%2Bregression&cvo_campaign=000000OA&cvo_crid=81776224001569&Matchtype=p&msclkid=cbadd0b95814159b49398c295d33ad98&utm_source=bing&utm_medium=cpc&utm_campaign=Search%7CGeneric%20-%20SPSS%20-%20Linear%20Regression%7C000000OA%7CWW%7CEN%7CIN%7CBMM%7C10000380%7CNULL&utm_term=%2Blinear%20%2Bregression&utm_content=Linear%20Regression%20-%20Generic%20-%20BM