ST3131 Regression Analysis

Life Expectancy in Countries Model

Members:

Gong Zhi Qiang [redacted]

Nguyen Thao Ngan [redacted]

Ng Ai Hiang (A0157049Y)

Khoo Hong Liang [redacted]

1. Introduction

In this report, we will be investigating the differences in life expectancy around the world, covering 130 countries. We have developed a linear regression model to predict life expectancy for different countries using various social, economic and health predictor variables. This model can potentially be useful to governments around the world in identifying areas of development in their country that can increase the overall life expectancy.

In this project, we employ two statistical softwares - R and SAS. We first come up with a basic model and test its appropriateness and adequacy. Further improvements are made to the model by considering the statistical results from R regression output and eventually come up with a final model. Our final model may not necessarily be the best model to explain all factors influencing the life expectancy, but it is the best model given the amount of available information.

2. Preliminary Examination of Data

2.1 Source of Data

We obtain our raw data of life expectancy in 2015 and other indicators from a dataset on Kaggle, an online data science community. Data of life expectancy and health indicators for 193 countries is from The Global Health Observatory (GHO) data repository under World Health Organization (WHO) and the corresponding economic data of the same year is from the United Nation website. Validity of the data is justified as it is taken from the database of renowned international organizations.

2.2 Cleansing Data

We first cleanse the data by removing missing observations and NA values, which reduces the dataset from 193 to 130 observations. Most of the variables in the dataset are continuous numerical values. However, the variable “Status” is a categorical variable, hence we converted it to dummy variables, 1 to indicate developed countries and 0 to indicate for developing countries.

2.3 Understanding Dataset

The original dataset consists of 18 variables. *Country* was excluded in our model as each country is already represented by an observation and *Year* was omitted as it is 2015 for all observations. *Life Expectancy* is identified to be the dependent variable and the other 15 variables are potential influential factors. More details on the variables are shown in the appendix.

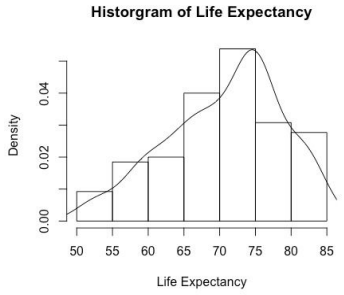


Figure 1- Histogram of Life Expectancy Plot

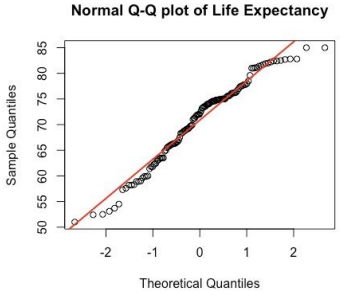


Figure 2 - Normal Q-Q plot

We perform a normality test to check that if our response variable - life expectancy - follows the Gaussian Distribution, which is the empirical assumption in linear regression theory. Refer to Figures 1 and Figure 2, the histogram and Q-Q plot shows that life expectancy is fairly normally distributed. Hence, we do not need to consider any transformation of the variable and can continue with our analysis.

2.4 Choosing an Appropriate Model

2.4.1 Basic Model

$$\textit{Life Expectancy} = \beta_0 + \beta_1 \textit{Status} + \beta_2 \textit{Adult Mortality} + \beta_3 \textit{Infant Deaths} + \beta_4 \textit{Hepatitis.B} + \beta_5 \textit{Measles} + \beta_6 \textit{BMI} + \beta_7 \textit{Under Five Deaths} + \beta_8 \textit{Polio} + \beta_9 \textit{Diphtheria} + \beta_{10} \textit{HIV/Aids} + \beta_{11} \textit{GDP} + \beta_{12} \textit{Population} + \beta_{13} \textit{Thinness from 1 to 19 years} + \beta_{14} \textit{Thinness from 5 to 19 years} + \beta_{15} \textit{Income composition of resources} + \beta_{15} \textit{Schooling} + e$$

The output from R can be referred to Figure 7.3 in the Appendix.

Conducting linear regression on this model, we obtain a R^2 value of 0.9005 with F-Statistic of 63.95 with 16 and 113 df, and a p-value smaller than 2.2e-16. Overall, the model is significant. However, it may not be the most ideal model and we can attempt to get a better model.

2.4.2 Best Fit Model

We employ stepwise regression on SAS to choose the best model. The outputs are based on metric AIC using both forward selection and backward elimination. The selection of the model is important, as under-fitting a model may not capture the true nature of the variability in the outcome variable, while an over-fitted model loses generality. *AIC* is one of the best methods to select the model that best balances these drawbacks. The resultant best model is shown below.

Summary of Stepwise Selection								
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Income_composition_of_resources		1	0.8065	0.8065	95.4487	533.53	<.0001
2	Adult_Mortality		2	0.0632	0.8697	25.1367	61.58	<.0001
3	Hepatitis_B		3	0.0184	0.8881	6.0418	20.76	<.0001
4	HIV_AIDS		4	0.0052	0.8933	2.1303	6.05	0.0153

Figure3: SAS output of stepwise regression

Life Expectancy = β_0 + β_1Adult Mortality + β_4 Hepatitis B + β_{10}HIV Aids + β_{15}Income composition of resources + ε ***-----Model(1)***

We then test for the significance of the model. The model shows strong linearity with a R^2 value of 0.8933, which is only smaller than the R^2 of the original basic model by 0. 0072. Despite a drop in R^2 value, we see a large improvement in AIC value from 541.57 to 258.67. Under H_0: Model (1) is insignificant vs H_1: Model (1) is significant, We obtain a F-Statistic of 261.6 and a p-value 2.2e-16 that is smaller than 0.05. Therefore, we reject H_0 and conclude that the reduced model is significant.

3. Preliminary Model Evaluation and Possible Improvements

Our regression analysis is based on the following assumptions:

- The error term in the regression model follows a normal distribution with a mean 0
- The error term in the regression model has a constant variance.
- The observations are independent.

We will then perform tests to see if these assumptions are met.

3.1 Residual Plot

We plot the residuals versus fitted Life Expectancy, and independent variables to examine if there are any observable patterns. These plots are shown in section 7.5 in Appendix. As there are no patterns in these plots, it seems the residual plots do not show any violation of the assumptions of independence and constant variance.

3.2 Residual Normality Test and Kolmogorov-Smirnov Test

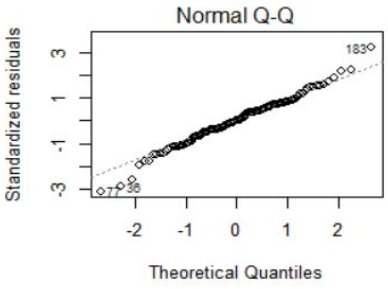


Figure 4 - Q-Q Plot

```
        One-sample Kolmogorov-Smirnov test

data:  res
D = 0.06106, p-value = 0.7176
alternative hypothesis: two-sided
```

Figure 5 - Kolmogorov-Smirnov Test Results

We conducted 2 tests to check for normality and the results are shown above. Since the data points on the Q-Q plot are generally closely aligned to the line on the chart, and the P value from Kolmogorov-Smirnov Test is large (0.7176>0.05). Therefore, we can hold the assumption for normality and conclude that the data points follow a normal distribution.

3.3 Collinearity Analysis

We plot a scatter plot between the five variables in the model.. There are some linear relationship exists between Life expectancy (our y variable) and other predictor variables, especially with income composition of resources and Adult mortality rates. There are also no apparent linear correlation between the predictor variables, hence we do not need to consider the problem of collinearity.

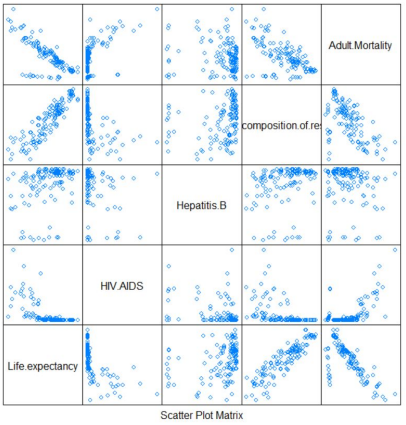
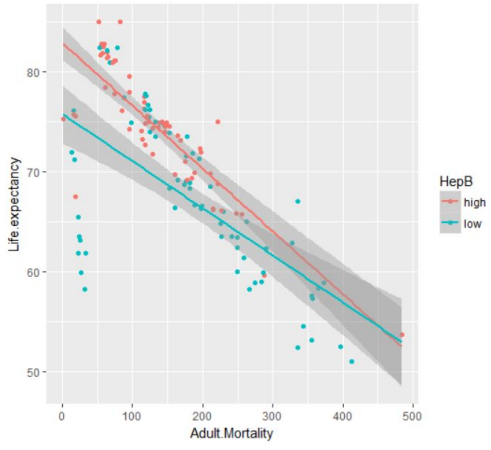


Figure 6: Scatter Plot of the Five Variables

3.4 Consider Interaction Effect and Second Order Variables



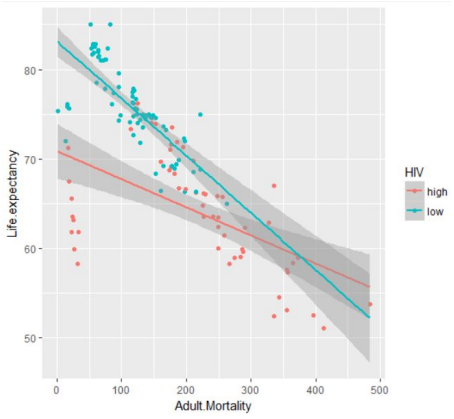


Figure 7 & 8 - Life Expectancy vs Adult mortality based on Hepatitis B Immunization Level and HIV level

We should also consider if there are interaction effects between predictor variables. HIV infection is the leading cause of death among people with AIDS, and Hepatitis B immunization could effectively prevent Hepatitis B infection. Therefore, it is natural to

consider if these two factors would interact with adult mortality rate to affect life expectancy of a country. Figure 7 and 8 show that with high or low rate of HIV death or Hepatitis B immunization, the linear correlation between life expectancy and adult mortality changes. This implies that the predictor variables HIV and Hepatitis B might interact with the adult mortality to affect its linear relationship with life expectancy. This prompts us to consider interaction terms between the predictor variables.

In addition, based on the scatter plots from Figure 5, there are strong linear relationships between each predictor variable against life expectancy. Hence, we do not think there is a need to consider second order variable and propose to improve the preliminary model by adding interaction terms between the predictor variables.

4. Improved Model

4.1 Selection of an improved model

As mentioned, we decided to include the interaction term to model (1) obtained previously. After which, we conduct stepwise regression on the model with independent variables containing first ordered Income composition of resources, Hepatitis B, Adult mortality, HIV/AIDS and all interaction terms between them.Th e In other words, we obtain the following improved model:

Hence the formula of the best-fit model including interaction terms will be:

$y = 45.9713942 + 34.9265658(X2) + 0.0724083(X3) - 2.6055370(X2X5) - 0.0003397(X3X4) +0.01379(X3X5) + \varepsilon$ **-----Model(2)** ,

where X2 = Income composition of resources, X3 = Hepatitis B, X4 = Adult Mortality,
X5 = HIV/AIDS

The output from R can be referred to Figure 7.9.1 in Appendix

4.2 Evaluation of Model (2)

The model has a R^2 value of 0.90987 and the F-test result also proves the overall significance of the model (refer to Figure 7.9.2 in Appendix).

Closer look into this model allows us to have a deeper understanding of how the predictor variables affect life expectancy. Among them, change in income composition of resources contribute most significantly to life expectancy, following by the interaction term between Income composition and HIV. This makes sense because when people in the country generally have lower income, they will have less opportunities and resources to combat AIDS, which would negatively affect the life expectancy of the country (hence the negative sign of the coefficient).

4.3 Test for Coincidence

We now want to introduce one indicator variable, Status of the country. This is to check if the model is the same for both developed and developing countries. Let x1 be the status of the country such that:

$$x1 = \begin{cases} 0, if\ the\ country\ is\ developing \\ 1, \ \ if\ the\ country\ is\ developed \end{cases}$$

After running the test for coincidence, with the p-value is greater than 0.05, we conclude that we cannot reject the null hypothesis that x1 does not improve the model. Hence, we decide not to include this indicator variable in our model (refer to Figure 7.9.3 in Appendix).

4.4 Residual Plot

Similarly, for model(2) we plotted the residuals versus fitted Life Expectancy, and independent variables to examine if there are any observable patterns. As there are no patterns in these plots, it seems the residual plots do not show any violation of the assumptions of independence and constant variance. These plots can be found from Section 7.10 in Appendix.

4.5 Normal Probability Plot and Kolmogorov-Smirnov Test

We also perform the normal probability plot (Figure 7.11.1 in Appendix) and Kolmogorov-Smirnov Test (Figure 7.11.2 in Appendix) to test for the normality of the residual. The graph and the result is shown in the appendix. The residuals lie close to a straight line according to the normal probability plot. The p-value from Kolmogorov-Smirnov Test is 0.9819 which is larger than 0.05. Therefore we do not reject null hypothesis that the residuals follow a normal distribution.

4.6 Influential points

To test that if there is any influential point in the model, we conduct RSTUDENT, DFFITS and DFBETAS tests to check for possible outliers. In the following graph, the numbers corresponds to the *i*th data point in the cleansed data set.

In RSTUDENT test, we take any points with a studentized residuals that have an absolute value greater than 2. We get 23, 35, 43, 84, 96, 114, 130 data point as the potential outliers, which corresponds to Portugal, Algeria, Bangladesh, Gabon, India, Malawi and Niger.

The result of DFFITS and DFBETAS are shown in Figure 7.12.1 in Appendix. For DFFITIS, the threshold for an influential observation is $2\sqrt{\frac{5+1}{130-5-1}} \approx 0.44$, any data points with a DFFITS absolute value greater than 0.44 is a potential influential point. We notice that there are some influential points, which are 23, 52, 77, 84, 100, 105 and 109, 114,130. These points correspond to Portugal, Brazil, Equatorial Guinea, Gabon, Israel, Kiribati and Lebanon, Malawi and Niger.

For DFBETAS, the threshold is $\frac{2}{\sqrt{130}} \approx 0.18$. Any data points with DFBETAS absolute value

greater than 0.18 is a potential influential point, which are 35,54,79,84,100,105,110, which are corresponding to Algeria, Burkina Faso, Estonia, Gabon, Israel, Kiribati and Lesotho.

To summarise, 84th(Gabon) seems to be most potential influential point as it fits in all three tests. This can be explained by the fact that Gabon is among countries with highest rates of Malaria, Tuberculosis and Endocrine disorder[1]. Such low health profile will definitely affects the life expectancy of the country.

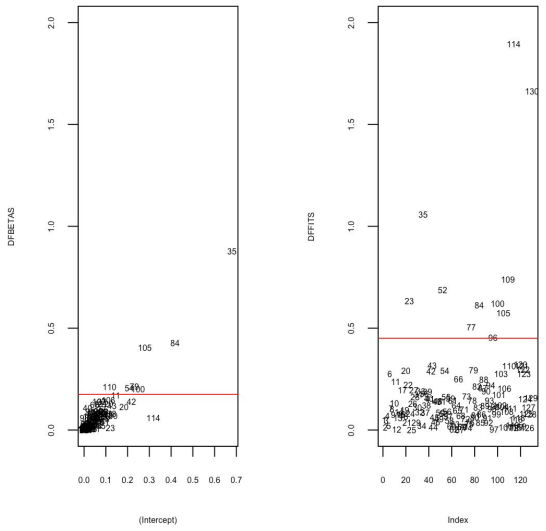


Figure 8: Plot showing the influential point by DFBETAS method and DFFITS method

5. Limitation Discussion

In conclusion, our model (2) provides a relatively good linear fit for the relation between life expectancy and its predictor variables, with a R-square of value greater than 0.9. Although the model is closely fitted to the regression line, there are still limitations to the model that we have yet to consider:

Firstly, we might not have considered all the possible predictor variables that are related to life expectancy. One possible variable that we could consider is availability of primary health services to the population, as without proper and affordable primary health care, even a minor disease could be potentially fatal. We would also be more likely to gain a greater understanding on the relationship between adult mortality and life expectancy if we take into account the effects of the availability of primary health care.

Secondly, there are some influential outliers in the model that could affect the linearity as shown in session 4.6. These outliers should not be excluded to entry or measurement error because the data is collected from UN and should be credible and not miscalculated. They should be further studied carefully by looking into the specific conditions in each outlier countries. For example, one of the influential point - Malawi - is an African country which suffers from frequent outbreak of violence and political instability. All these might affect the predictor variables and cause this data point to behave differently from the rest.

Thirdly, although the residual plots for both model(1) and (2) had no distinctive patterns, some of the plots (refer to Appendix) were x-axis unbalanced, where most of the plots are skewed towards one side of the x-axis. However upon plotting the actual life expectancy vs fitted life expectancy plot, we can clearly see from figure 7.13 in Appendix, there is a strong linear correlation between the actual values and fitted values. Therefore, we can conclude that even though the plots might be x-axis unbalanced, the model is still significant and accurate.

Lastly, Lack of fit test was not conducted as there were no repeated set of values. Similarly, Run and Durbin-Watson tests were also not conducted to test for the presence of serial correlation.Since we only limit our data to the year 2015, the data set has no time series and does not warrant these tests.

6. Conclusion

We have proposed a model which identifies key factors that affect life expectancy. We first consider a preliminary model (model1) reduced from the original model between life expectancy and all 15 predictor variables using stepwise selection based on AIC. We then run several tests to diagnose the preliminary model and suggest an improved model (model2) from the preliminary model with the involvement of interaction terms. Finally, we test for the significance, assumption and influential points for model 2 and prove that it is a significant linear model that could be used to predict life expectancy. Our final proposed best model is:

$$y = 45.9713942 + 34.9265658(X2) + 0.0724083(X3) - 2.6055370(X2X5) - 0.0003397(X3X4) + 0.01379(X3X5) + \varepsilon$$

where y = Life expectancy, X2 = Income composition of resources, X3 = Hepatitis B, X4 = Adult Mortality, X5 = HIV/AIDS.

The proposed model can serve as a guide to governments around the world who seek to continuously develop their nation. One method of measuring a country's development is the Human Development Index (HDI)[2], used by the United Nations. One of the three main indicators in HDI calculation is the Life Expectancy of the country and this is why our model comes in place. Based on our model, governments can consider allocating more funds to increase Hepatitis B immunization coverage or channel more resources to combating HIV and AIDS.

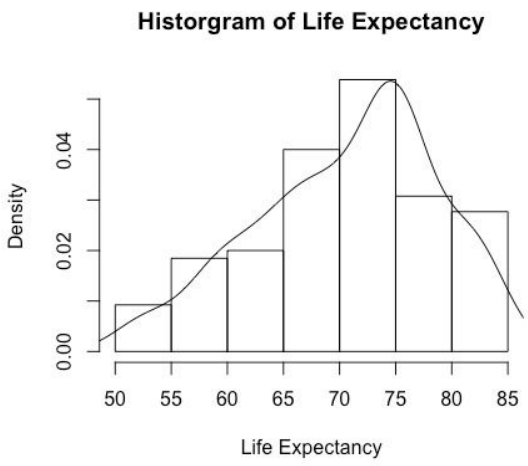
\

7. Appendix

7.1 Dataset

Variable	Definition
Country	Country
Year	Year, all observations were taken in year 2015
Status	Developed or Developing Status
Life Expectancy	Life Expectancy in age
Adult Mortality	Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)
Infant Deaths	Number of Infant Deaths per 1000 population
Hepatitis B	Hepatitis B (HepB) immunization coverage among 1-year-olds (%)
Measles	Measles - number of reported cases per 1000 population
BMI	Average Body Mass Index of entire population
Under-five deaths	Number of under-five deaths per 1000 population
Polio	Polio (Pol3) immunization coverage among 1-year-olds (%)
Total Expenditure	General government expenditure on health as a percentage of total government expenditure (%)
Diphtheria	Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)
HIV/AIDS	Deaths per 1 000 live births HIV/AIDS (0-4 years)
GDP	Gross Domestic Product per capita (in USD)
Population	Population of the country
Thinness 1-19 years	Prevalence of thinness among children and adolescents for Age 1 to 19 (%)
Thinness 5-9 years	Prevalence of thinness among children and adolescents for Age 5 to 9 (%)
Income Composition of Resources	Income index specific to each country (range from 0 to 1)
Schooling	Number of years of Schooling(years)

7.2 Normality Test on Response Variable



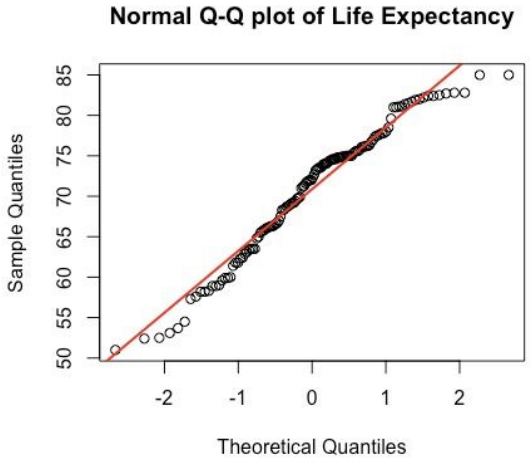


Figure 7.2.1- Histogram of Life Expectancy Plot

Figure 7.2.2 - Normal Q-Q plot

7.3 Linear Regression on Basic Model

```
Call:
lm(formula = data1$Life.expectancy ~ ., data = data1)

Residuals:
    Min      1Q  Median      3Q     Max
-7.3416 -1.4379  0.0359  1.5459  7.9402

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     5.075e+01  3.017e+00  16.822  < 2e-16 ***
Status                         -3.424e-01  8.299e-01  -0.413   0.6807
Adult.Mortality                -2.092e-02  3.607e-03  -5.799 6.15e-08 ***
infant.deaths                   6.601e-02  3.285e-02   2.010   0.0469 *
Hepatitis.B                     4.333e-02  2.260e-02   1.917   0.0577 .
Measles                        -5.119e-05  5.729e-05  -0.893   0.3735
BMI                            -8.580e-03  1.550e-02  -0.554   0.5809
under.five.deaths              -4.783e-02  2.354e-02  -2.032   0.0445 *
Polio                           1.147e-02  1.267e-02   0.905   0.3676
Diphtheria                     -1.106e-02  2.630e-02  -0.420   0.6750
HIV.AIDS                       -4.847e-01  2.239e-01  -2.165   0.0325 *
GDP                             5.064e-06  3.003e-05   0.169   0.8664
Population                     -1.010e-09  9.564e-09  -0.106   0.9161
thinness..1.19.years           -1.228e-01  2.338e-01  -0.525   0.6004
thinness.5.9.years             -1.735e-02  2.287e-01  -0.076   0.9396
Income.composition.of.resources  3.325e+01  4.981e+00   6.676 9.59e-10 ***
Schooling                      -4.796e-02  2.402e-01  -0.200   0.8421
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.695 on 113 degrees of freedom
Multiple R-squared:  0.9005,    Adjusted R-squared:  0.8865
F-statistic: 63.95 on 16 and 113 DF,  p-value: < 2.2e-16
```

Figure 7.3 - Linear Regression Results

7.4 Stepwise Regression

```
Start:  AIC=541.57
data1$Life.expectancy ~ 1

                                 Df Sum of Sq    RSS    AIC F Value     Pr(F)
+ Income.composition.of.resources  1    6654.2 1596.4 330.04  533.53 < 2.2e-16 ***
+ Schooling                        1    5360.9 2889.7 407.18  237.46 < 2.2e-16 ***
+ Adult.Mortality                  1    4411.4 3839.2 444.11  147.08 < 2.2e-16 ***
+ HIV.AIDS                         1    3176.8 5073.8 480.36   80.14 3.436e-15 ***
+ BMI                              1    2450.5 5800.1 497.75   54.08 2.033e-11 ***
+ Polio                            1    2008.9 6241.7 507.29   41.20 2.449e-09 ***
+ GDP                              1    1956.9 6293.7 508.37   39.80 4.213e-09 ***
+ Status                           1    1920.5 6330.1 509.12   38.83 6.151e-09 ***
+ Diphtheria                       1    1793.4 6457.2 511.70   35.55 2.267e-08 ***
+ thinness..1.19.years             1    1739.4 6511.2 512.79   34.19 3.918e-08 ***
+ thinness.5.9.years               1    1707.3 6543.3 513.43   33.40 5.413e-08 ***
+ Hepatitis.B                      1    1142.4 7108.2 524.19   20.57 1.305e-05 ***
+ under.five.deaths                1     479.3 7771.3 535.79    7.89   0.00574 **
+ infant.deaths                    1     361.4 7889.2 537.74    5.86   0.01685 *
<none>                                       8250.6 541.57
+ Measles                         1      20.1 8230.5 543.25    0.31   0.57748
+ Population                      1       6.3 8244.3 543.47    0.10   0.75532
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 7.4.1 - Stepwise Regression Results - Start (R)

```
Step:  AIC=258.67
data1$Life.expectancy ~ Income.composition.of.resources + Adult.Mortality +
    Hepatitis.B + HIV.AIDS

                                 Df Sum of Sq     RSS    AIC F Value     Pr(F)
<none>                                       880.43 258.67
+ thinness..1.19.years            1     12.25  868.18 258.85   1.750 0.1882940
+ thinness.5.9.years              1      9.48  870.95 259.27   1.350 0.2475363
+ Population                      1      9.23  871.20 259.30   1.313 0.2539949
+ Polio                           1      7.61  872.82 259.55   1.081 0.3004765
+ under.five.deaths               1      6.40  874.03 259.73   0.907 0.3426737
+ infant.deaths                   1      4.11  876.32 260.07   0.581 0.4473146
+ Measles                         1      1.14  879.29 260.51   0.160 0.6897780
+ Status                          1      1.03  879.40 260.52   0.146 0.7033638
+ Diphtheria                      1      0.40  880.03 260.61   0.057 0.8121827
+ Schooling                       1      0.28  880.15 260.63   0.040 0.8420810
+ BMI                             1      0.24  880.18 260.64   0.034 0.8530040
+ GDP                             1      0.00  880.43 260.67   0.000 0.9901663
- HIV.AIDS                        1     42.62  923.04 262.82   6.050 0.0152699 *
- Hepatitis.B                     1     94.47  974.90 269.92  13.413 0.0003675 ***
- Adult.Mortality                 1    270.66 1151.09 291.52  38.427 7.597e-09 ***
- Income.composition.of.resources 1   2213.65 3094.08 420.06 314.285 < 2.2e-16 ***
---
```

Figure 7.4.2 - Stepwise Regression Results - Final (R)

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Income_composition_of_resources		1	0.8065	0.8065	95.4487	533.53	<.0001
2	Adult_Mortality		2	0.0632	0.8697	25.1367	61.58	<.0001
3	Hepatitis_B		3	0.0184	0.8881	6.0418	20.76	<.0001
4	HIV_AIDS		4	0.0052	0.8933	2.1303	6.05	0.0153

Figure 7.4.3- Stepwise Regression Results(SAS)

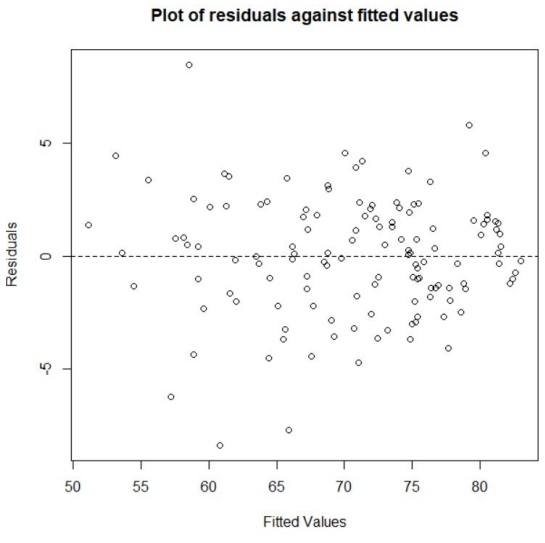
```
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
data1$Life.expectancy ~ 1

Final Model:
data1$Life.expectancy ~ Income.composition.of.resources + Adult.Mortality +
    Hepatitis.B + HIV.AIDS


                              Step Df  Deviance Resid. Df Resid. Dev      AIC
1                                                    129  8250.5957 541.5658
2 + Income.composition.of.resources  1 6654.1861       128  1596.4096 330.0371
3                  + Adult.Mortality  1  521.2926       127  1075.1170 280.6445
4                      + Hepatitis.B  1  152.0722       126   923.0448 262.8186
5                         + HIV.AIDS  1   42.6157       125   880.4291 258.6737
```

Figure 7.4.3 - Stepwise Model Path

```
Call:
lm(formula = data1$Life.expectancy ~ Income.composition.of.resources +
    Adult.Mortality + Hepatitis.B + HIV.AIDS, data = data1)

Residuals:
    Min      1Q  Median      3Q     Max
-8.3646 -1.4775  0.0805  1.7418  8.4728

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    47.782061   1.697385  28.150  < 2e-16 ***
Income.composition.of.resources 35.111815   1.980575  17.728  < 2e-16 ***
Adult.Mortality                -0.021118   0.003407  -6.199  7.6e-09 ***
Hepatitis.B                     0.037515   0.010244   3.662 0.000368 ***
HIV.AIDS                       -0.520103   0.211445  -2.460 0.015270 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.654 on 125 degrees of freedom
Multiple R-squared:  0.8933,    Adjusted R-squared:  0.8899
F-statistic: 261.6 on 4 and 125 DF,  p-value: < 2.2e-16
```

Figure 7.4.4 - Best Fit Model from Stepwise Regression

7.5 Residual Plots for Model (1)



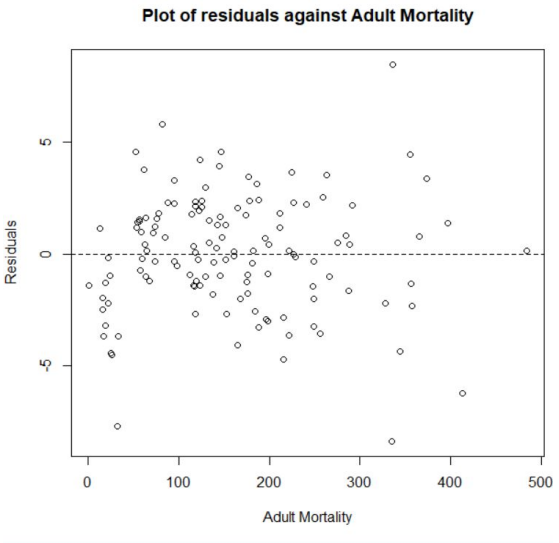


Figure 7.5.1 - Residual vs FItted Values Plot *Figure 7.5.2 - Residual vs Adult Mortality Plot*

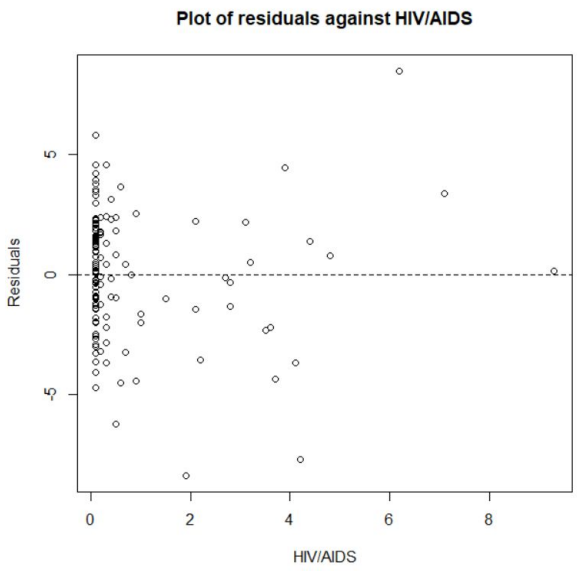


Figure 7.5.3 - Residual vs HIV/AIDS Plot

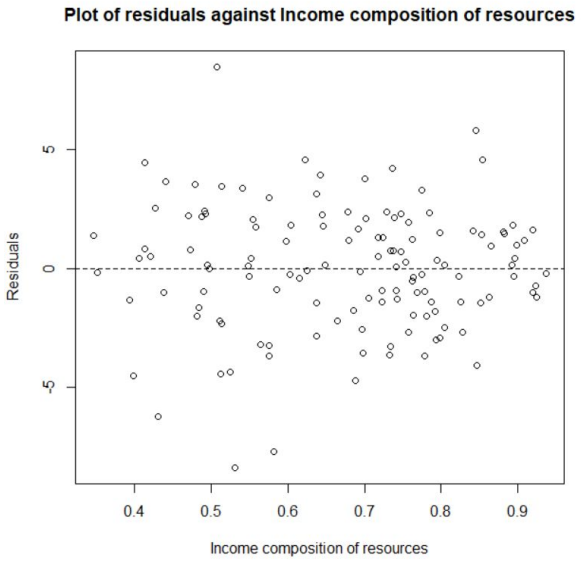


Figure 7.5.4 - Residual vs Income com. Plot

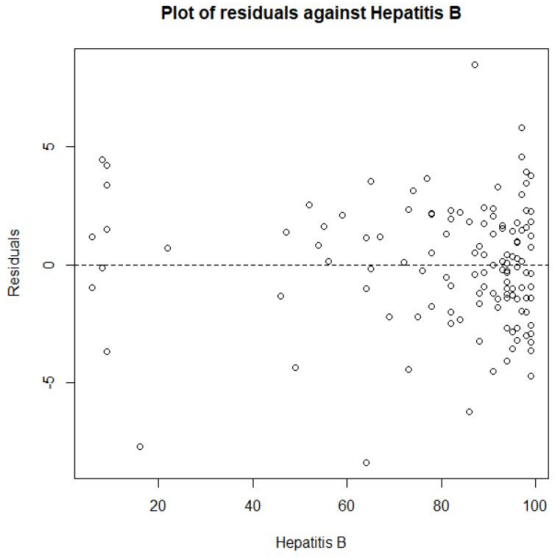


Figure 7.5.5 - Residual vs Hepatitis B Plot

7.6 Residual Normality Test and Kolmogorov-Smirnov Test for Model (1)

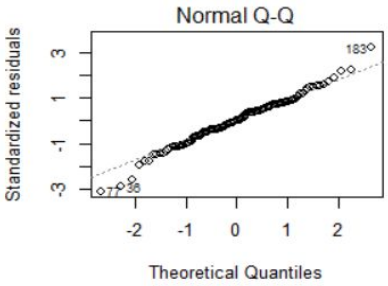


Figure 7.6.1 - Q-Q Plot

```
        One-sample Kolmogorov-Smirnov test

data:  res
D = 0.06106, p-value = 0.7176
alternative hypothesis: two-sided
```

Figure 7.6.2 - Kolmogorov-Smirnov Test Results

7.7 Scatter Plot for Model (1)

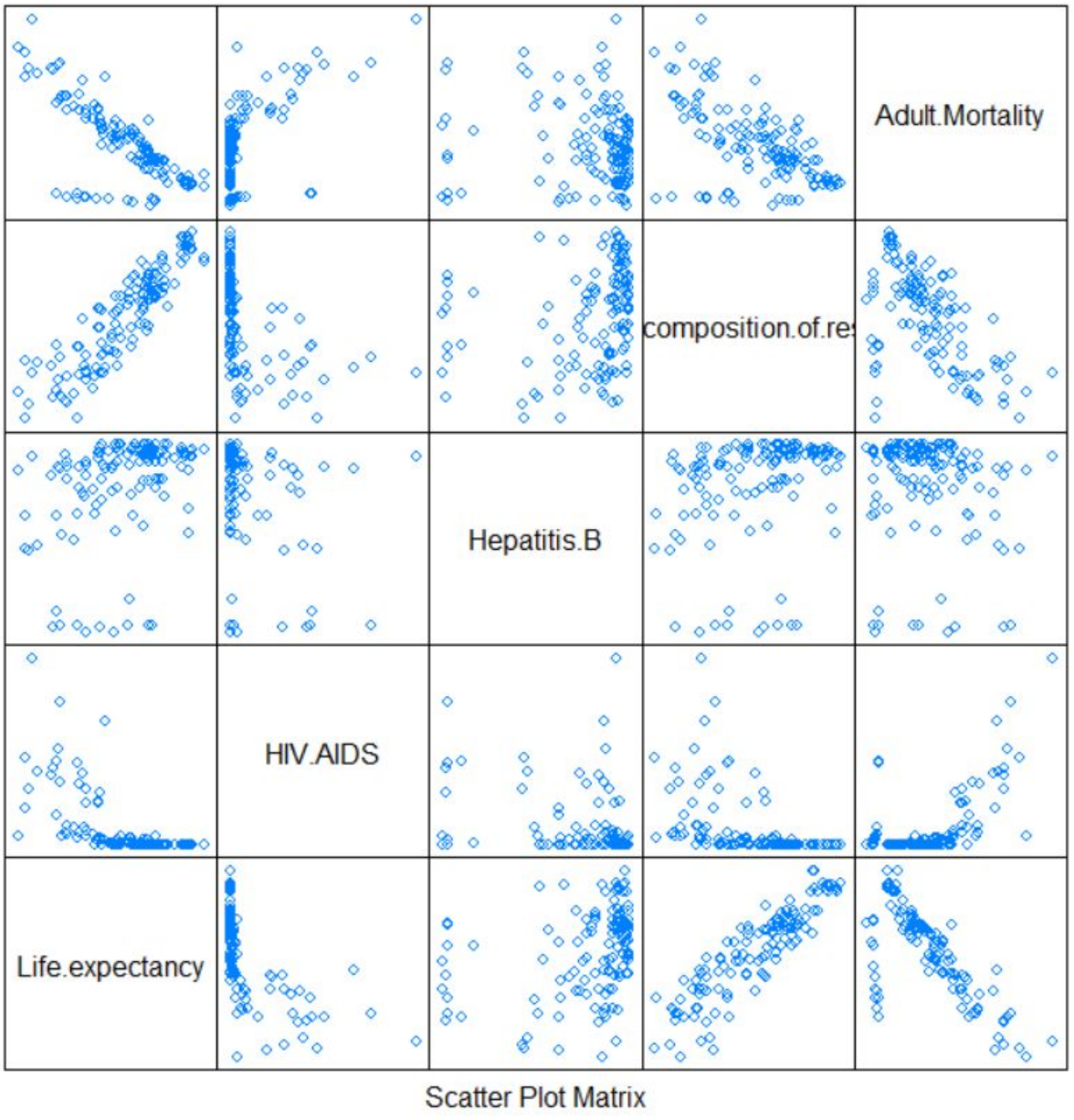


Figure 7.7 - Scatter Plot Matrix

7.8 Plots to Check for Interaction Effect

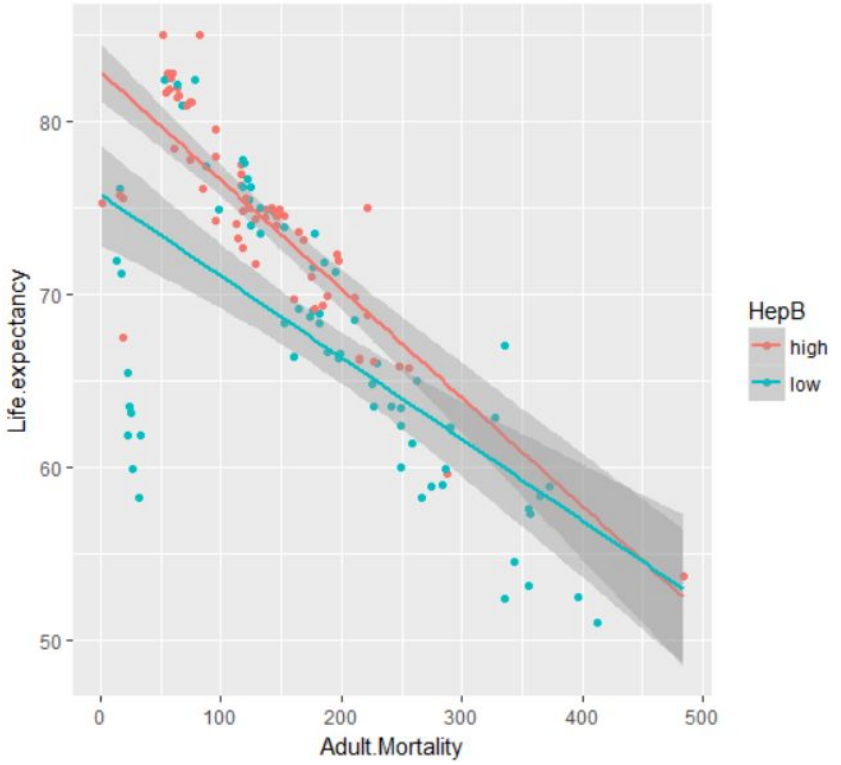


Figure 7.8.1 - Life expectancy vs Adult mortality based on Hepatitis B Immunization Level

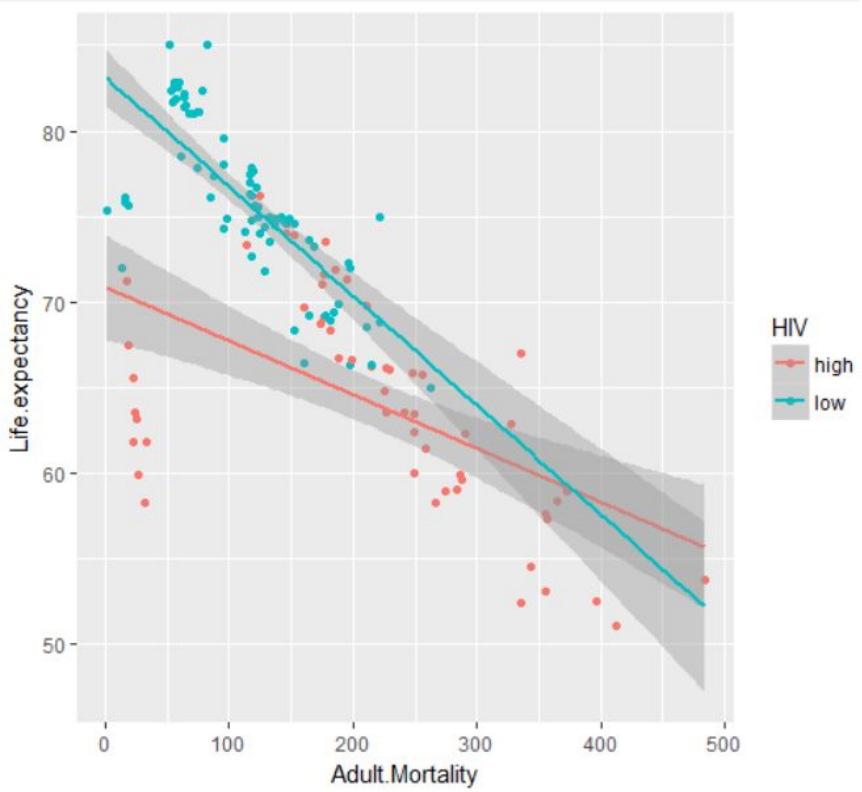


Figure 7.8.2 - Life expectancy vs Adult mortality based on HIV

7.9 Improve model with interaction terms - Model (2)

```
Residuals:
    Min      1Q  Median      3Q     Max
-8.5162 -1.4785 -0.0238  1.5804  7.7038

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.597e+01  1.417e+00  32.445  < 2e-16 ***
x2           3.493e+01  1.845e+00  18.931  < 2e-16 ***
x3           7.241e-02  1.291e-02   5.608 1.26e-07 ***
x25         -2.606e+00  6.013e-01  -4.333 3.01e-05 ***
x34         -3.397e-04  4.328e-05  -7.850 1.68e-12 ***
x35          1.379e-02  4.469e-03   3.085  0.00251 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.464 on 124 degrees of freedom
Multiple R-squared:  0.9087,    Adjusted R-squared:  0.9051
F-statistic: 246.9 on 5 and 124 DF,  p-value: < 2.2e-16
```

Figure 7.9.1 - Summary of the improved model with interaction terms (model 2)

```
Analysis of Variance Table

Response: y
           Df Sum Sq Mean Sq   F value    Pr(>F)
x2          1 6654.2  6654.2 1095.8241 < 2.2e-16 ***
x3          1  131.0   131.0   21.5767 8.539e-06 ***
x25         1  338.2   338.2   55.6957 1.295e-11 ***
x34         1  316.4   316.4   52.1073 4.623e-11 ***
x35         1   57.8    57.8    9.5202  0.002507 **
Residuals 124  753.0     6.1
```

Figure 7.9.2 - F-test for the overall significance of model 2

```
Analysis of Variance Table

Model 1: y ~ x2 + x3 + x25 + x34 + x35
Model 2: y ~ x1 + x2 + x3 + x25 + x34 + x35 + x12 + x13 + x125 + x134 +
    x135
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    124 752.97
2    120 713.11  4    39.855 1.6767 0.1598
```

Figure 7.9.3 - Test for Independence

7.10 Residual Plots for Model (2)

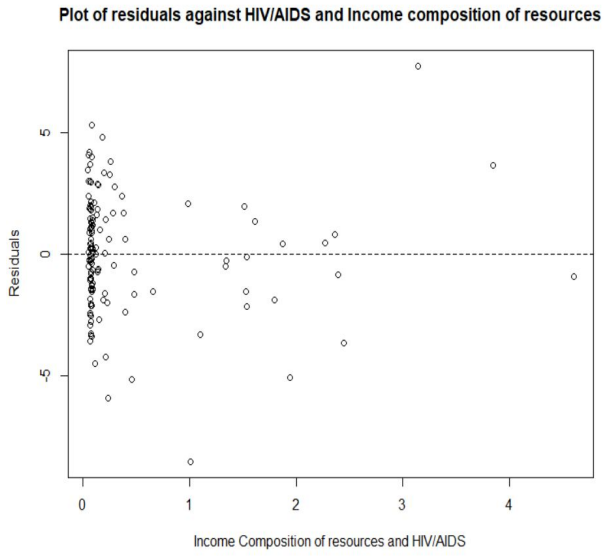


Figure 7.10.1 - Residual vs Income com. and HIV Plot

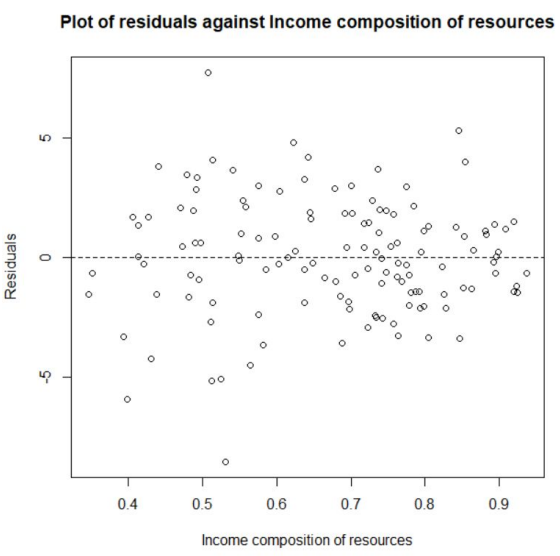


Figure 7.10.2 - Residual vs Income com. Plot

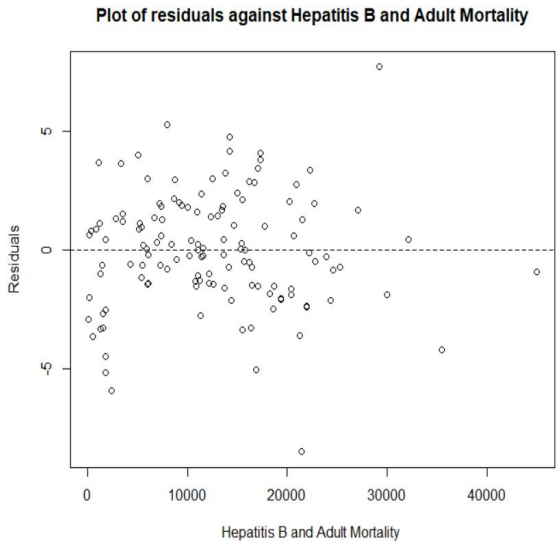


Figure 7.10.3 - Residual vs Hep. B and Adult mort. Plot

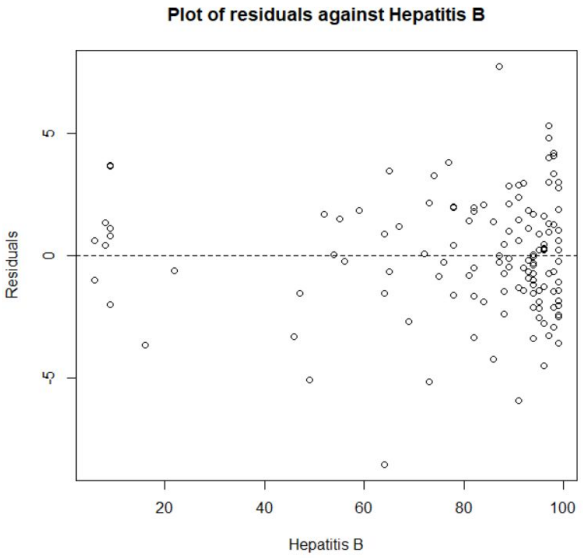


Figure 7.10.2 - Residual vs Hep. B Plot

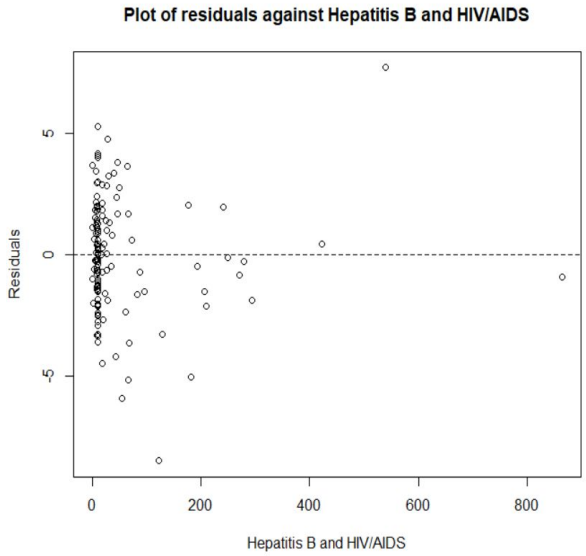


Figure 7.10.5 - Residual vs Hep. B and HIV Plot

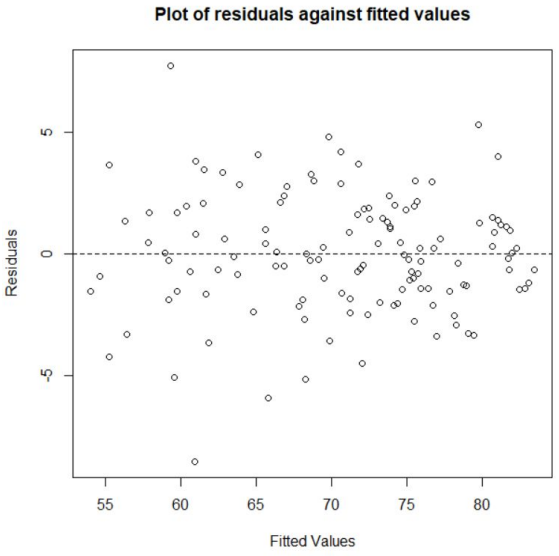


Figure 7.10.2 - Residual vs Fitted values Plot

7.11 Residual Normality Test and Kolmogorov-Smirnov Test for Model (2)

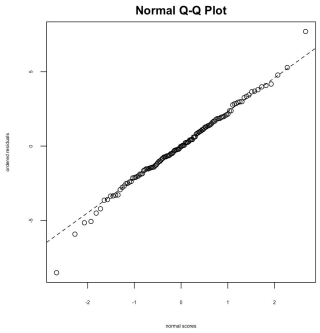


Figure 7.11.1 - Q-Q Plot

```
        One-sample Kolmogorov-Smirnov test

data:  res
D = 0.040814, p-value = 0.9819
alternative hypothesis: two-sided
```

Figure 7.11.2 - Kolmogorov-Smirnov Test

7.12 Influential Points

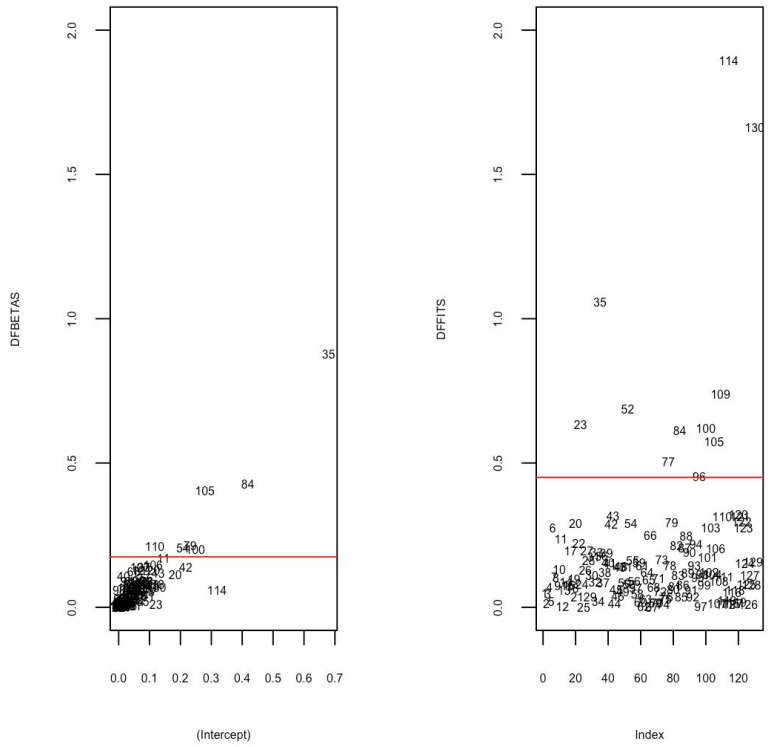
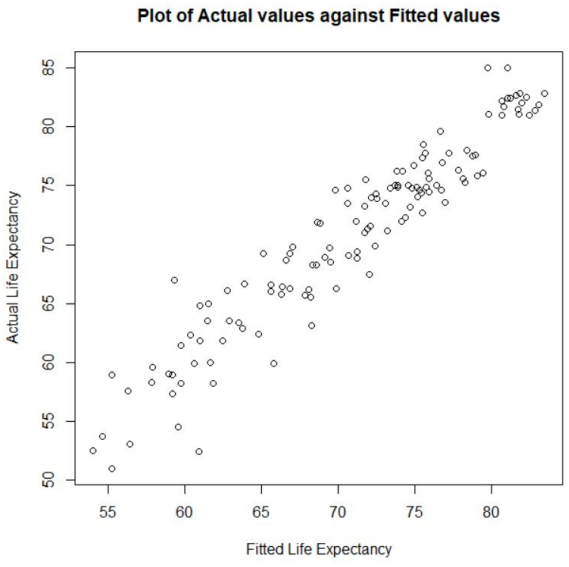


Figure 7.12.1 - Influential point based on DFBETAS DFFITS

7.13 Actual values vs Fitted Values



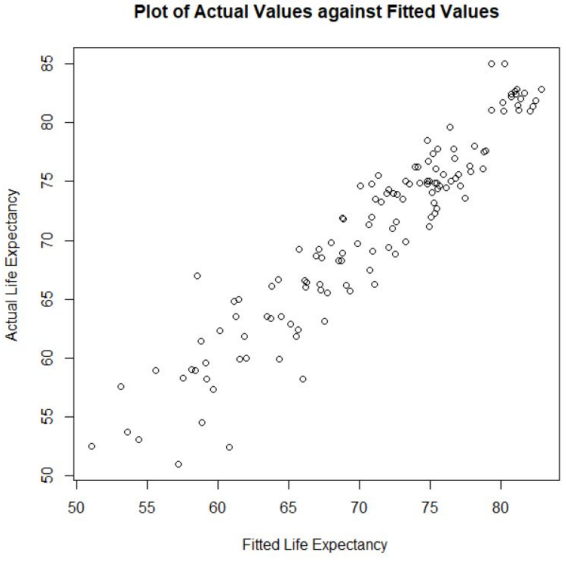


Figure 7.13.1 - Model(1): Actual vs Fitted Values

Figure 7.13.2 - Model(2) : Actual vs Fitted Values

8. References

1. "HEALTH PROFILE GABON." World Life Expectancy. Accessed April 17, 2018. http://www.worldlifeexpectancy.com/country-health-profile/gabon.
2. "Human Development Reports." Human Development Index (HDI) | Human Development Reports. Accessed April 17, 2018. http://hdr.undp.org/en/content/human-development-index-hdi.