|  |
| --- |
| It's a Wonderfully Long Life: Examining the Future of Life Expectancy |
| Life Expectancy Analysis |
| |  |  |  | | --- | --- | --- | | By: Mhatre, Rupalee |  |  | |

**Introduction:**

Life expectancy is defined as an estimate of the average age that members of a particular population group will be when they die. Life expectancy is a key metric used to evaluate population health. It captures mortality along the entire life span of a population. Life expectancy started to improve in early 19th century for the rich countries but it remained low for the poor countries. This inequality of health across the globe has improved significantly since 1900 but still remains a global issue.

The goal of this project is to predict life expectancy based on factors such as demographic, socio economic, immunization and other health factors. The analysis will also aim at answering questions like what are the actual number of variables affecting life expectancy. What changes should a country with low life expectancy implement to improve the life span of its citizens?

**Dataset:**

The dataset for the analysis was a combination of health related data from World Health Organization and related economic data from United Nations. Data from 193 countries from year 2000-2015 totaling 2938 records was used for the analysis. The analysis started with 21 variables including demographic related variables like Country, Status, Population health related variables like BMI, thinness etc. , Immunization related like Polio, Hepatitis ,Measles and Economic like Income and total expenditure.

* **Data Summary:**

The table below is the list of variables from the dataset. Life Expectancy is the dependent variable.

|  |  |
| --- | --- |
| Country | Year |
| Status | Adult.Mortality |
| infant.deaths | Alcohol |
| percentage.expenditure | Hepatitis.B |
| Measles | BMI |
| under.five.deaths | Polio |
| Total.expenditure | Diphtheria |
| HIV.AIDS | GDP |
| Population | thinness..1.19.years |
| thinness.5.9.years | Income.composition.of.resources |
| Schooling | **Life Expectancy(Dependant variable)** |

The analysis of means and medians of variable in the dataset show that that average life expectancy between the years of 2000 and 2015 is about 70 years. There are 512 developed countries in the dataset and 2426 developing countries. (Appendix 1)

* **Missing values:**

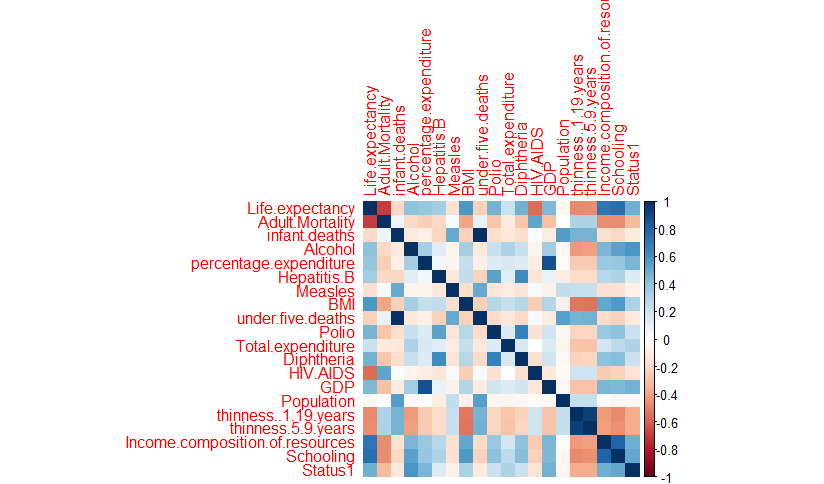
There were 10 missing values for Life expectancy. Data was deleted for these records. Other significant missing values were 652 for population, 553 for Hepatitis, 226 for total expenditure and some more variables. Deleting these rows would have resulted in loss of significant information and probably an incorrect model. So MICE was used to impute missing values. The final dataset with the imputed values was then used for further analysis



* **Binary variables:**

The Status variable had two values, developed and developing. The variable was converted to a binary variable where a new variable, Status 1 was created with the value for developed =1 and the value for developing = 0

* **Analyzing correlation:**



There is a high correlation between Adult mortality and HIV the explanation of that could be if more adult population gets HIV their death rate increases hence adult mortality increases.

Infant deaths and under 5 deaths are redundant data as their correlation is 1. Also Infant death is highly correlated to Measles the reason could be as more infants get measles, the death rate increases.

Percentage Expenditure and GDP have very high correlations. As the GDP goes up so does the percentage expenditure of the country on it population.

There is also high correlation between Alcohol, schooling and status. The correlation between Status and Alcohol means that developed countries have a higher rate of alcohol consumption. The correlation between alcohol and schooling is a bit unexpected.

These correlation were utilized later during the data analysis.

**Data Modeling and Analysis:**

The approach taken to derive data models was to first find multicollinearity of variables using VIF and eliminate the variables that had high influence.

The first model included all the variables. The model was significant as the p value was almost 0 and the R squared and Adjusted R squared were around 83%. But not all variables were significant. The highest VIF (176) was for infant deaths.

In the next model infant deaths variable was removed. This model too was significant with adjusted R squared of 82% but not all variables were significant.

In the next couple of models, variables with high VIFs like thinness 5-9 years were removed. To finally have a model with VIFs less than 5 for all variables. But the model still had variables with p values > than 0.05.

The next model (Model 4) removed all variables that were not significant. So Alcohol, Hepatitis B, Measles, Total expenditure, Population, GDP were removed from the model. All Models from model 4 onwards were considered as candidates and were measured against various factors. (Appendix 2)

List of variables for Model 4:

|  |  |
| --- | --- |
| 1 | Adult.Mortality |
| 2 | percentage.expenditure |
| 3 | BMI |
| 4 | Polio |
| 5 | Diphtheria |
| 6 | under.five.deaths |
| 7 | HIV.AIDS |
| 8 | thinness..1.19.years |
| 9 | Income.composition.of.resources |
| 10 | Schooling |
| 11 | Status |

The normal distribution plot were heavy at the tails. So the residuals were not normally distributed. (Appendix 2)

The fitted vs residuals plot did not have constant variance for lower values. There seemed to be a constant variance thereafter but there were some possible outliers too. (Appendix 2)

**Outliers:**

The models discussed above indicated some outliers. So studentized residuals were used to find the outliers.

The outliers were all for some developing nations. One common factor amongst the outliers was that the Population and Percentage expenditure data varied significantly for consecutive years. So the data most likely was incorrect. Also multiple other variables had a value of 0. So those values were incorrect too. Finding a regression model for incorrect data may not have yielded the desired results. So it was decided to remove the outliers from the analysis. (Appendix 3)

**Model 5:**

The next model 5 was designed with same set of variables as model 4 but without the outliers in the dataset. In this model thinness 1-19 became an insignificant variable. So some of the outlier data had a relationship with thinness being significant in the earlier model. The R-squared and adjusted R-squared were around 83% , the SSE was 42455 and the PRESS value was 42948. Except for the one insignificant variable, thinness 1-19 years, all other confidence intervals were acceptable (Appendix 4).

The normal distribution plot was still heavily tailed but looked better than the one for previous model. The fitted vs residual plot was the same as previous model. (Appendix 4)

The list of variables for Model 5:

|  |  |
| --- | --- |
| 1 | Adult.Mortality |
| 2 | percentage.expenditure |
| 3 | BMI |
| 4 | Polio |
| 5 | Diphtheria |
| 6 | under.five.deaths |
| 7 | HIV.AIDS |
| 8 | thinness..1.19.years |
| 9 | Income.composition.of.resources |
| 10 | Schooling |
| 11 | Status |

**Model 6:**

In the previous model, Thinness 1-19 was an insignificant variable so removed that variable to derive Model 6. The model was significant and also all variables were now significant. All variables had VIFs below 5. The R-squared and adjusted R-squared were around 83% , the SSE was around 42475 and PRESS was 42936.93.The confidence intervals for all variables had acceptable values. (Appendix 5)

Like the previous models, the normal plot still was heavily tailed and there was very little change in the fitted vs residual plot from the previous ones. (Appendix 5)

The list of variables for Model 6:

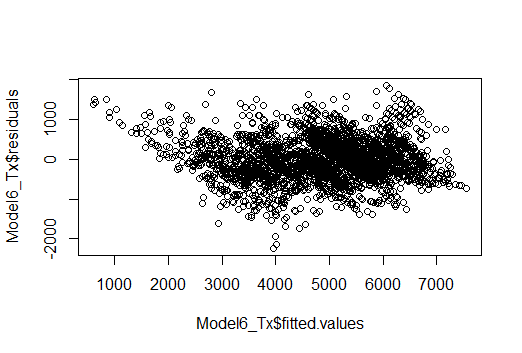
|  |  |
| --- | --- |
| 1 | Adult.Mortality |
| 2 | percentage.expenditure |
| 3 | BMI |
| 4 | Polio |
| 5 | Diphtheria |
| 6 | under.five.deaths |
| 7 | HIV.AIDS |
| 8 | Income.composition.of.resources |
| 9 | Schooling |
| 10 | Status |

**Transformations:**

The fitted vs residual plots for all models so far showed non-constant error variance amongst the lower end of prediction values. These values were for developing nations like Botswana, Kenya, Lesotho etc., These countries had higher adult mortality and under five death values which could be a factor. (Appendix 6)

Boxcox transformation to evaluate if there was any change in the error variance was performed. Λ was calculated to be around 1.8 so it was decided to square the values. A model was built with the transformed values and fitted vs residual values were plotted. The transformations did not impact the plot. So it was decided to keep the non-transformed values for further analysis.

Plot with Transformed values:



Note: Model 7 was derived incorrectly so the details were skipped here.

**Model 8:**

To derive Model 8, regression subsets were used and Status was not indicated as a significant variable in any of the models from the subsets. So status was removed for Model 8. The r-squared and Adjusted r-squared remained at 83%, SSE was 43730 and PRESS was 43169. The normal probability and fitted vs residual plots remained the same. (Appendix 7)

Variables for Model 8:

|  |  |
| --- | --- |
| 1 | Adult.Mortality |
| 2 | percentage.expenditure |
| 3 | BMI |
| 4 | Polio |
| 5 | Diphtheria |
| 6 | under.five.deaths |
| 7 | HIV.AIDS |
| 8 | Income.composition.of.resources |
| 9 | Schooling |

**Model 9:**

For Model 9, StepAIC was used on the full model with all variables. The model did not do well compared to all the previous models. It had 13 variables with high VIF values. The r-squared and Adjusted r-squared remained at 83%, SSE was 44643 and PRESS was 45230. So this model was rejected. (Appendix 8)

**Model 10:**

The goal of next few models was to reduce the number of variables and compare the models. In Model 10, the correlation between Income composition of resources and schooling was considered. Their VIFs were a little over 3. Model 10 was derived by removing schooling from model 6. The R-squared and Adjusted R-squared dropped to 80%, SSE was 50113 and PRESS was 50663. The Normal probability plot also was move heavily tailed now. This model did not look like an improvement over other models. (Appendix 9)

Variables for Model 10:

|  |  |
| --- | --- |
| 1 | Adult.Mortality |
| 2 | percentage.expenditure |
| 3 | BMI |
| 4 | Polio |
| 5 | Diphtheria |
| 6 | under.five.deaths |
| 7 | HIV.AIDS |
| 8 | Income.composition.of.resources |
| 9 | Status |

**Model 11:**

For Model 11, Schooling was used and Income composition of resources was removed. The r-squared and Adjusted R-squared went up from the previous model to 82% and SSE was 45296 and PRESS was 45728. The normal probability plot looked better for higher values but was heavily tailed for lower values. The fitted vs residual plot remained the same. (Appendix 10)

List of variables for Model 11:

|  |  |
| --- | --- |
| 1 | Adult.Mortality |
| 2 | percentage.expenditure |
| 3 | BMI |
| 4 | Polio |
| 5 | Diphtheria |
| 6 | under.five.deaths |
| 7 | HIV.AIDS |
| 8 | Schooling |
| 9 | Status |

It was observed that removing ICR gave better values of R squared compared to the model without Schooling. So removing ICR from some future models was considered (Model 13)

**Model 12:**

There was high correlation observed between Diphtheria and polio. In Model 12 Diphtheria was removed while keeping ICR and schooling. The R-squared and Adjusted R-squared went up from the previous model to 83% and SSE was 43136 and PRESS was 43574. The normal probability plot and fitted vs residual plots remained unchanged. (Appendix 11).This model without Diphtheria had better R squared and other values than the model without ICR.

List of variables for Model 12:

|  |  |
| --- | --- |
| 1 | Adult.Mortality |
| 2 | percentage.expenditure |
| 3 | BMI |
| 4 | Polio |
| 5 | under.five.deaths |
| 6 | HIV.AIDS |
| 7 | Income.composition.of.resources |
| 8 | Schooling |
| 9 | Status |

**Model 13:**

Model 13 analyzed the impact of removing both diphtheria and ICR. The R-squared and Adjusted R-squared went down from the previous model to 82% and SSE was 46232 and PRESS was 46646. The normal probability plot and fitted vs residual plots remained unchanged. (Appendix 12)

List of variables for Model 13:

|  |  |
| --- | --- |
| 1 | Adult.Mortality |
| 2 | percentage.expenditure |
| 3 | BMI |
| 4 | Polio |
| 5 | under.five.deaths |
| 6 | HIV.AIDS |
| 7 | Schooling |
| 8 | Status |

**Model 14:**

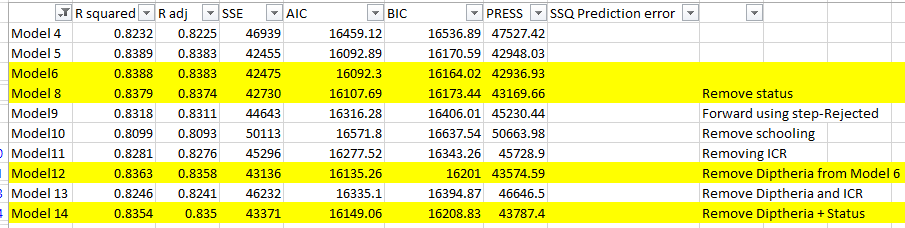
For Model 14, Diphtheria and status were removed and ICR was kept. The R-squared and Adjusted R-squared went up from the previous model to 83% and SSE was 43371 and PRESS was 43787. The normal probability plot and fitted vs residual plots remained unchanged. (Appendix 13)

This Model performed better than the one without Diphtheria and ICR.

List of variables for Model 14:

|  |  |
| --- | --- |
| 1 | Adult.Mortality |
| 2 | percentage.expenditure |
| 3 | BMI |
| 4 | Polio |
| 5 | under.five.deaths |
| 6 | HIV.AIDS |
| 7 | Income.composition.of.resources |
| 8 | Schooling |

**Model Comparison:**

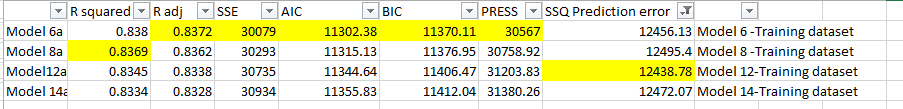


Model 4 and 5 were initial models that had some insignificant variables in them. So they were not considered. Model 9 had around 13 variables so it was rejected. Model 10 was derived by removing schooling from Model 6 but it had the lowest R-squared and adjusted R-squared and highest PRESS value. So this model was not considered. Model 11 and 13 again had lower values of R-squared and adjusted R-squared and higher values of PRESS and SSE as compared to some other models so these two models were not considered for further analysis.

Model 6, Model 8, Model 12 and Model 14 were good candidates for model selection. Based on all the performance measures, Model 6 performed the best. Although Model 8 without the Status variable and Model 12 without Diphtheria were not too far away in these performance values. Model 14 with fewer variables had parameter values close to the remaining three models.

**Out of Sample Model comparison:**

The data was split into training and test sets in the ratio 70:30. The above four models Model 6, Model 8, Model 12, Model 14 were run on the training dataset and compared for values like r-squared, adjusted r-squared, SSE, AIC,BIC and PRESS. Using the above four models, predictions were calculated for life expectancy on the test dataset and the sum of squared prediction errors were also calculated for the four contending models. The values for sum of squared prediction errors and other parameters were as shown in the screenshot below. From these values, Model 6 was still the best performing model although the other models were not too far away.



**Result: Final Model selection**

Analysis of models from both the entire dataset and out of sample data concluded that Model 6 performed better for most performance measures. The other three models though, were not very distant in performance from Model 6. This made the final model selection irresolute. Should the final model selection depend just on the performance measures or should the fact that a model with fewer variables like model 14 performed quite close also be considered? Using fewer variables in a model has benefits like practicality, availability of information, less computational time and complexity, less dependency on observed data, ease of interpretation and generalization. Based on the performance measures and considering the benefits of fewer variables in a model, ***Model 14 was selected as the final Model.***

List of variables for Model 14

|  |  |
| --- | --- |
| 1 | Adult.Mortality |
| 2 | percentage.expenditure |
| 3 | BMI |
| 4 | Polio |
| 5 | under.five.deaths |
| 6 | HIV.AIDS |
| 7 | Income.composition.of.resources |
| 8 | Schooling |

**Conclusion:**

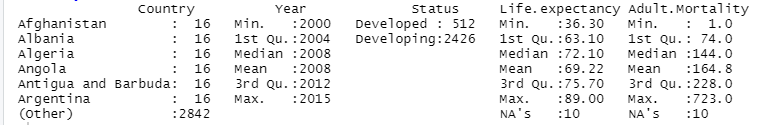
Life expectancy as defined in the beginning, is based on an estimate of the average age that members of a particular population group will be when they die. The analysis above revealed that factors like adult mortality , percentage expenditure, BMI, Polio, under five deaths, HIV, ICR and schooling were important factors in determining the life expectancy of a population. From an array of about twenty one variables which included factors like demographic, socio economic, immunization and other health factors, it was concluded that the above eight factors were important in determining the life expectancy of a population. Quick glance at the summary revealed that the average global life expectancy till the year 2015 was about 70 years. Availability and accuracy of data was one of the challenges faced when predicting life expectancy. Due to this some data had to be deleted while remaining was imputed. The contending four models with very similar performance values made it challenging to select a final model. Due to this, the selection of the final model came with a bit of uncertainty. The decision for the model was based on the fact that it had fewer variables than the other models while performing similar to other models. As a future plan, these four models should be analyzed by a domain expert in the field. This expert opinion will provide a different dimension to this analysis and will offer better confidence. The data used for the above analysis was from 2000-2015. Analysis could be done on recent data beyond year 2015 to solidify this analysis.

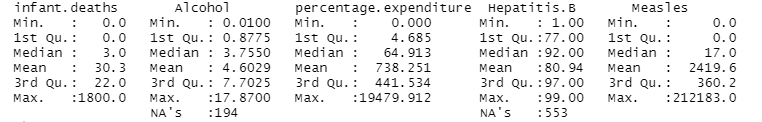
Finally, the hope is to utilize this analysis to improve life expectancy especially in developing countries by monitoring the major factors impacting life expectancy for their population. Factors like HIV rates in a country can impact the adult mortality hence taking measures to lower HIV rates can positively impact life expectancy. Ensuring that children get vaccinated regularly may also see an impact on life expectancy in the long term. The developed countries can monitor their income composition and percentage expenditure on the population to ensure high life expectancy.

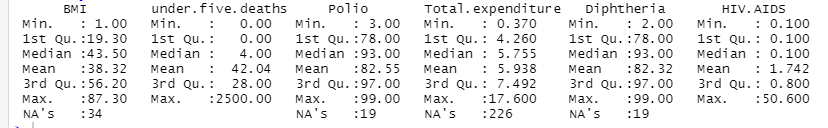
Appendix:

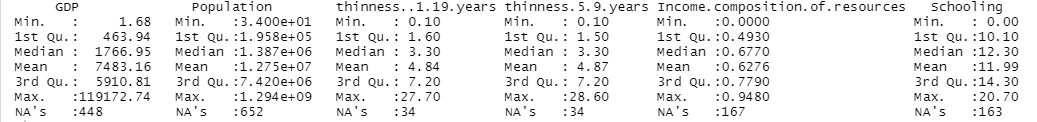
**Appendix 1:**

Means and Medians



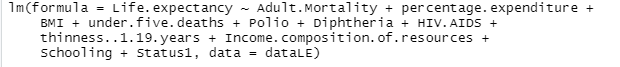






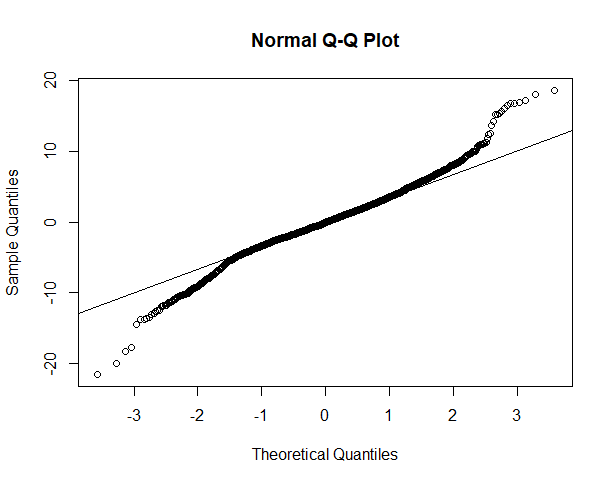
**Appendix 2:**

**Model 4:**

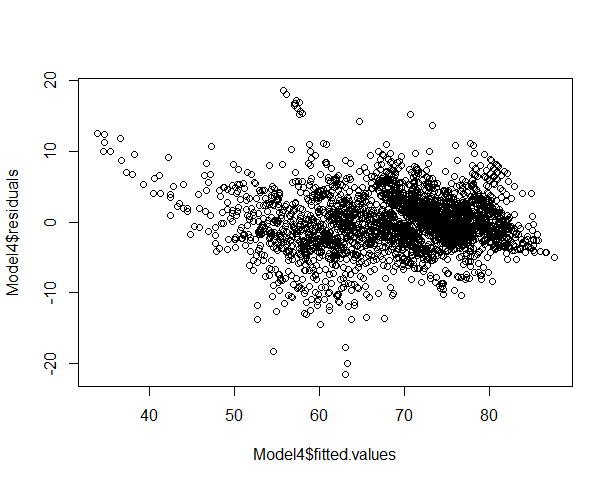




**Model 4-Normal probability plot**

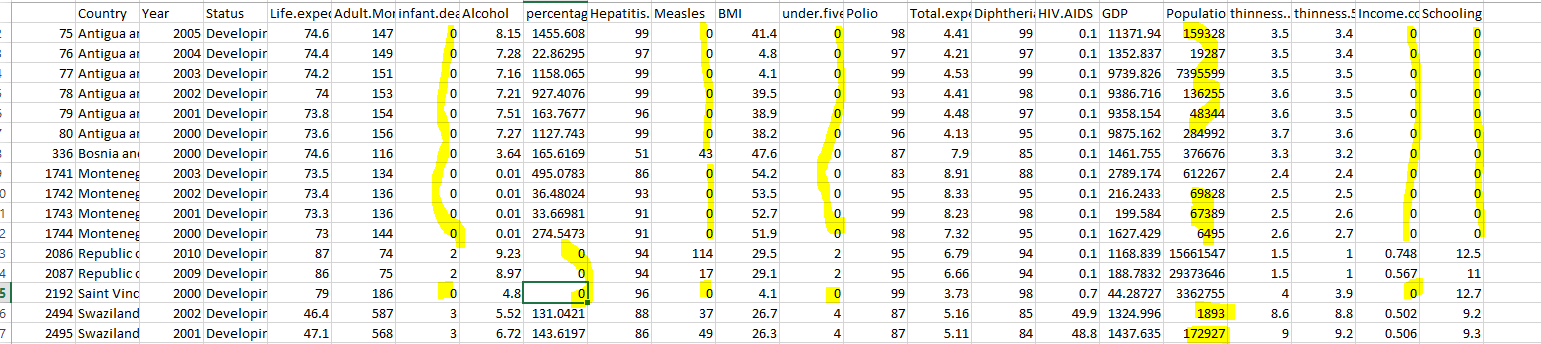


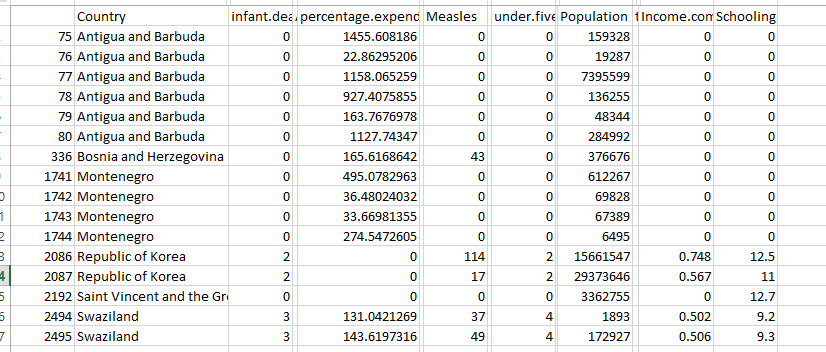
**Model 4: Fitted vs residual plot**



**Appendix 3:**

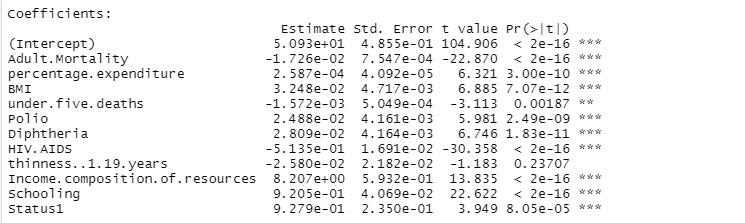
Outliers:



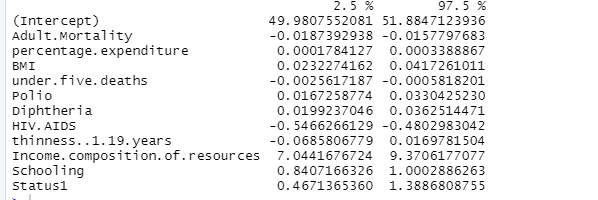


**Appendix 4:**

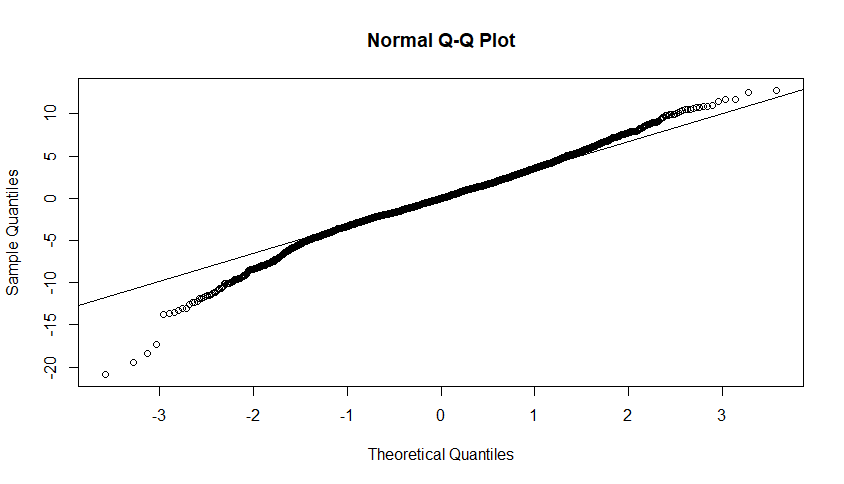
Model 5:



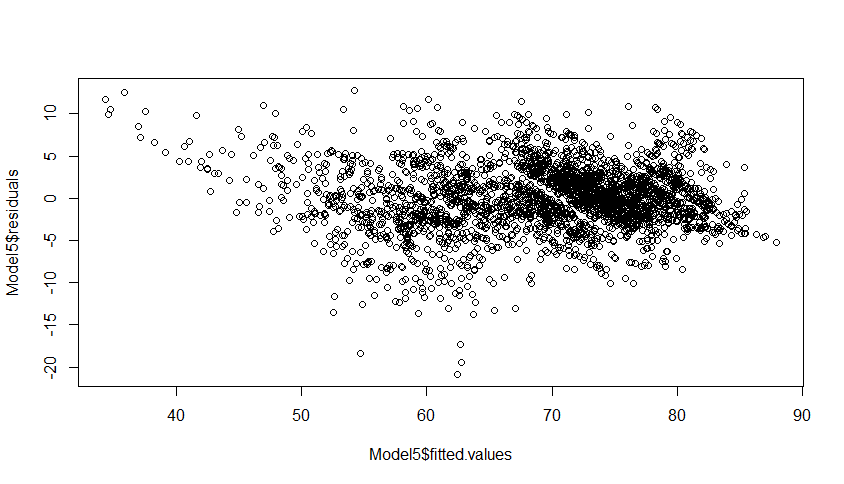




**Model 5-Normal probability plot**

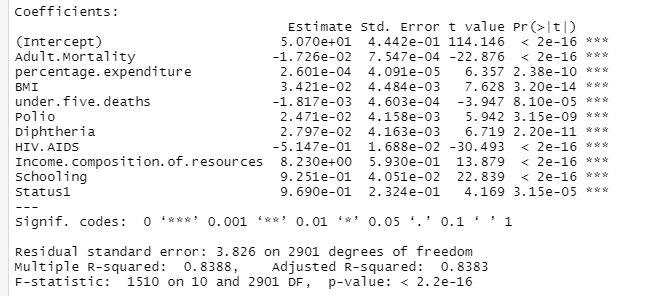


**Model 5: Fitted vs residual plot**



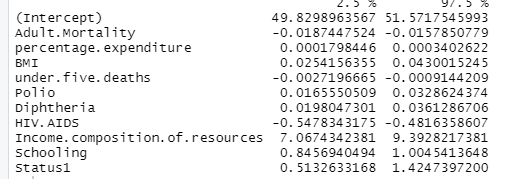
**Appendix 5:**

**Model 6:**

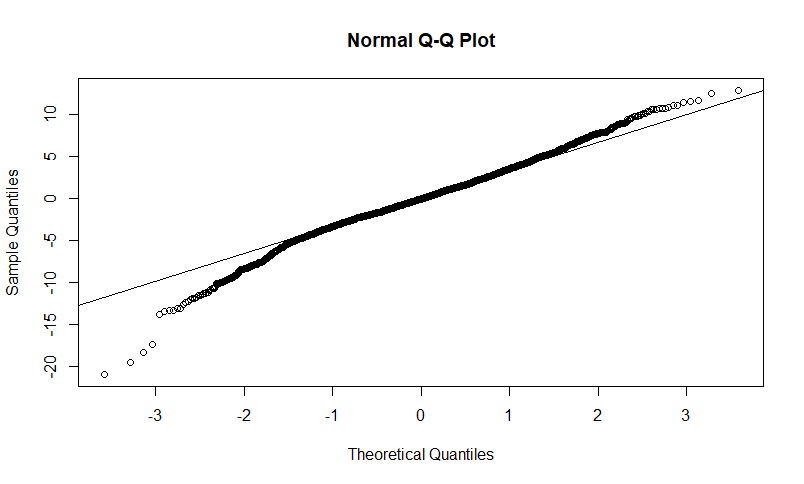




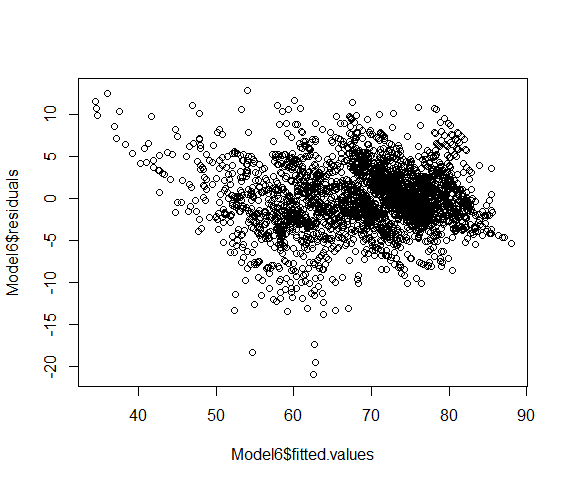
**Model 6: Confidence interval**



**Model 6-Normal probability plot**

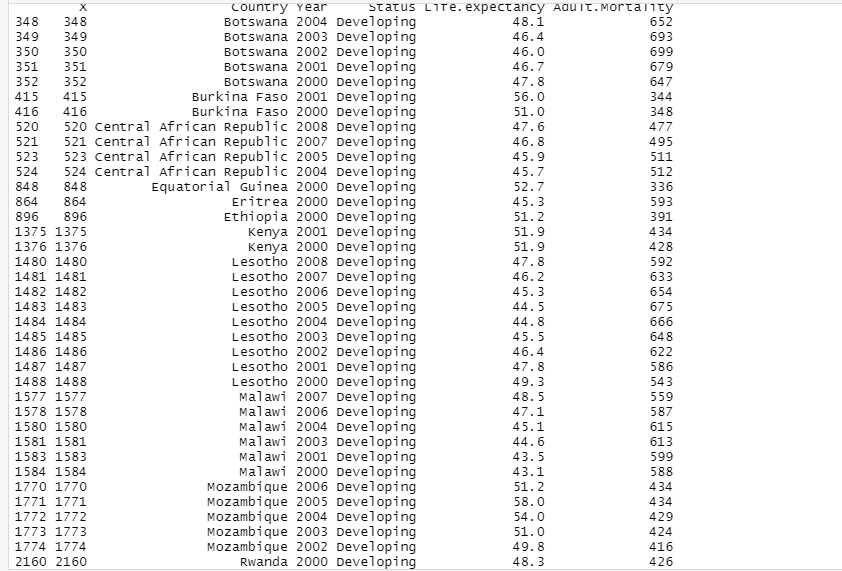


**Model 6: Residual vs fitted values**

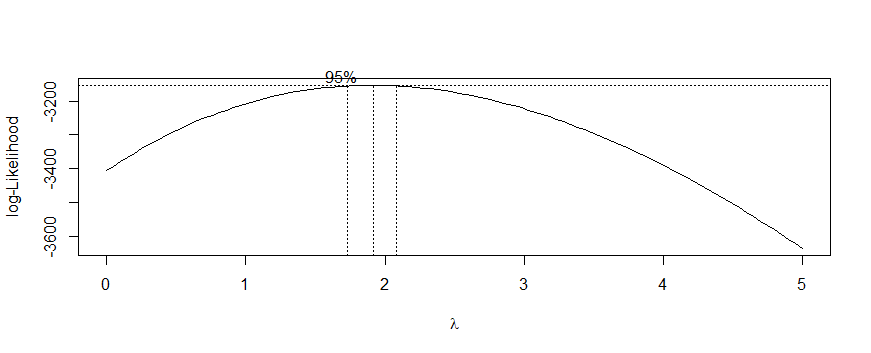


**Appendix 6:**

**Countries with non-constant error variance:**

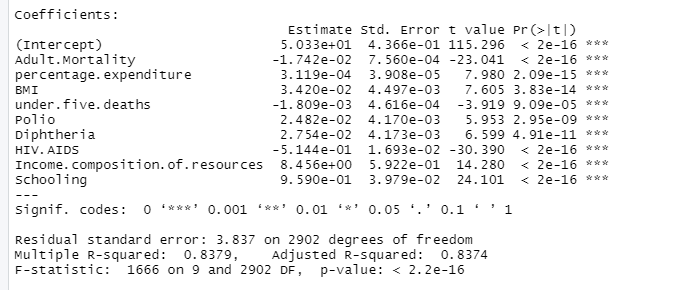


**Boxcox Transformation:**

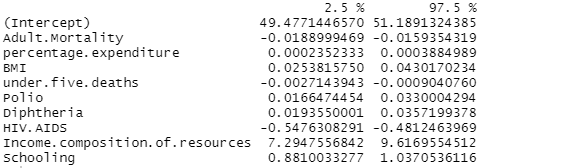


**Appendix 7:**

Model 8:



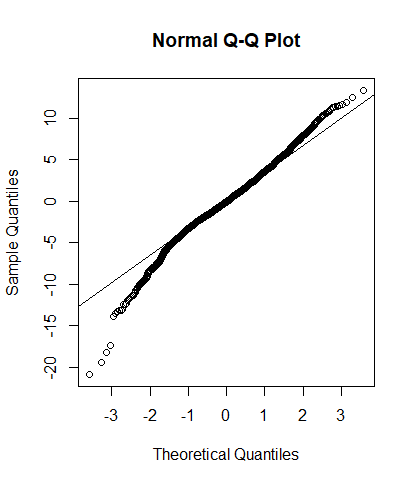
**Model 8 confidence intervals::**



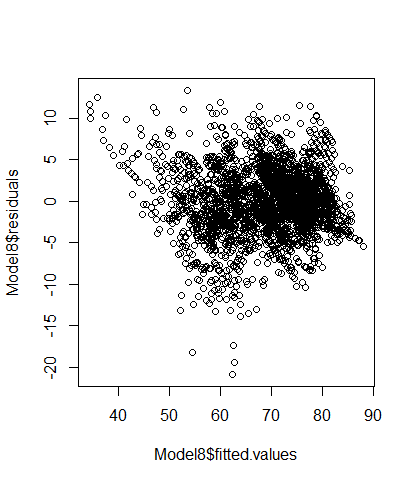
**Model 8:**



**Model 8: Normal probability plot**

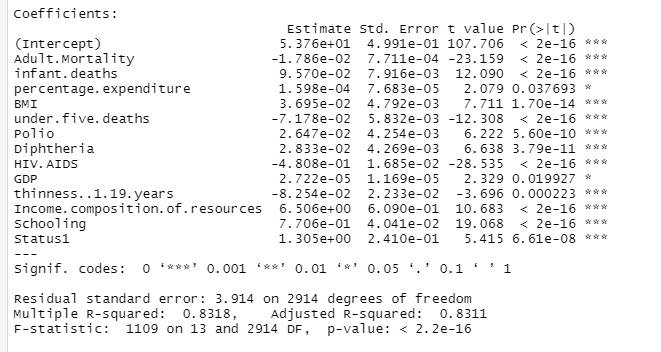


**Model 8: Residual vs fitted values**



**Appendix 8:**

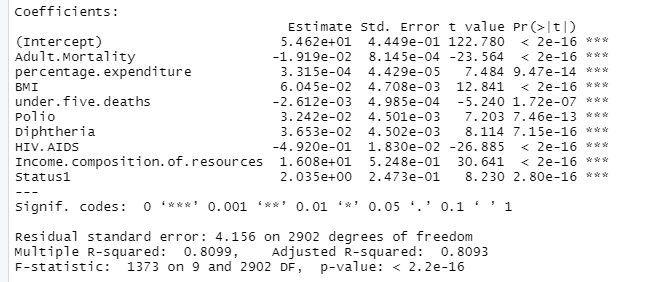
**Model 9:**





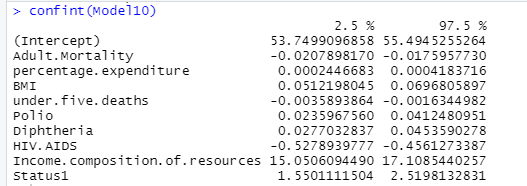
**Appendix 9:**

**Model 10**

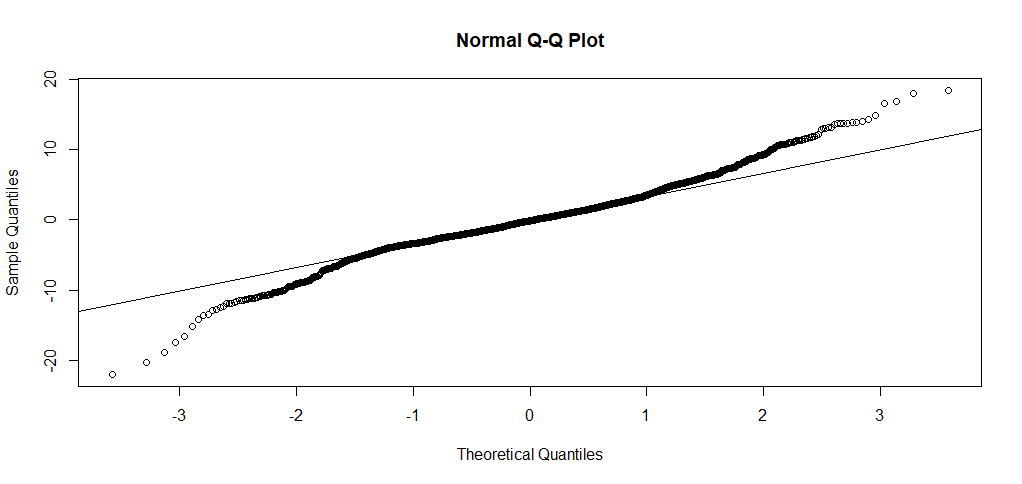




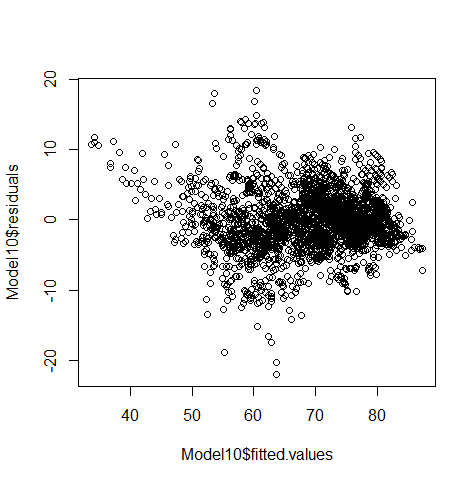
**Model 10 confidence intervals:**



**Model 10: Normal probability plot**

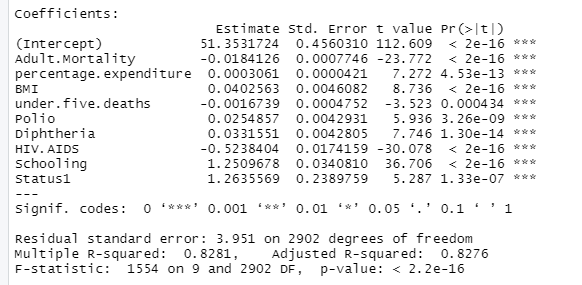


**Model 10: Residual vs fitted values**



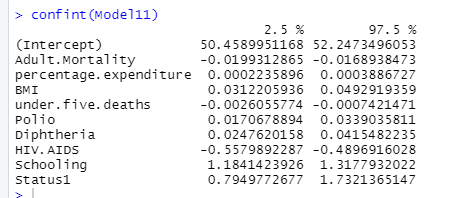
**Appendix 10:**

**Model 11:**

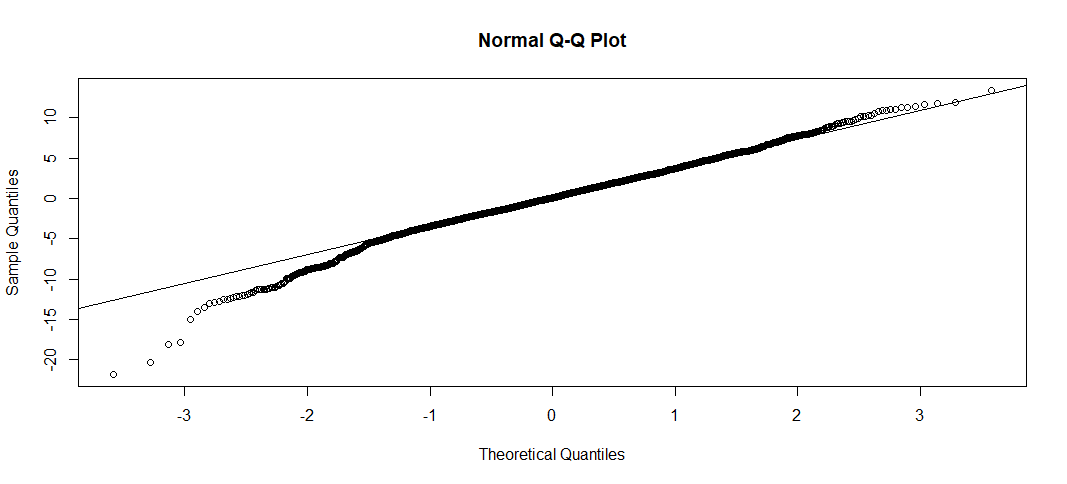




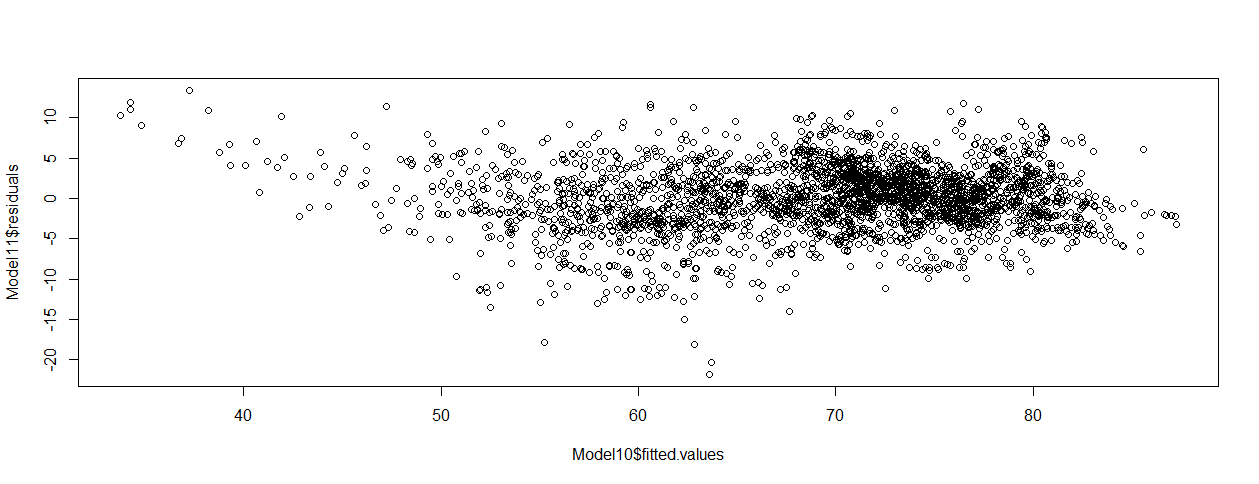
**Model 11 Confidence intervals:**



**Model 11 Normal Probability plot**

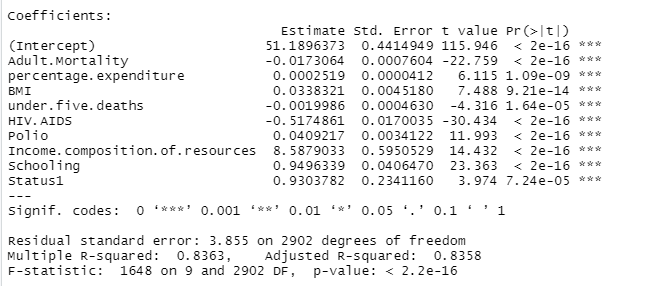


**Model 11 Fitted vs residual plot:**



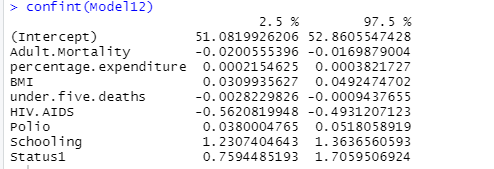
**Appendix 11:**

**Model 12:**

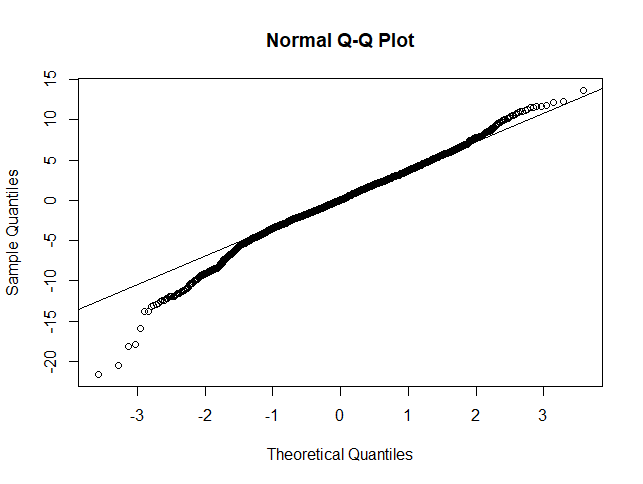




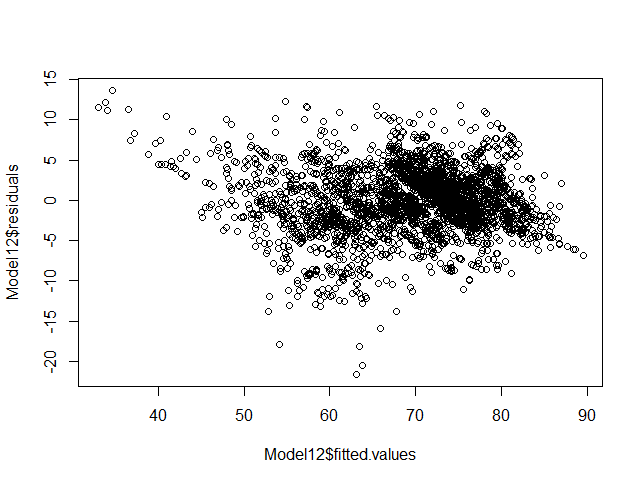
**Model 12 confidence interval:**



**Model 12: Normal probability plot**

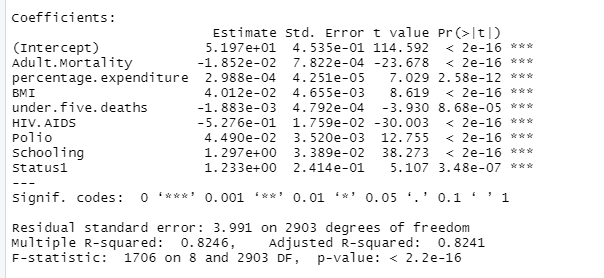


**Model 12: Fitted vs residual plot**



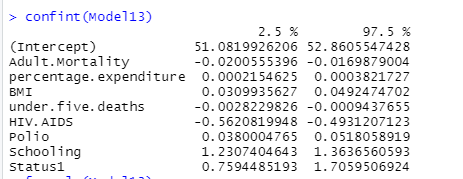
**Appendix 12:**

**Model 13:**



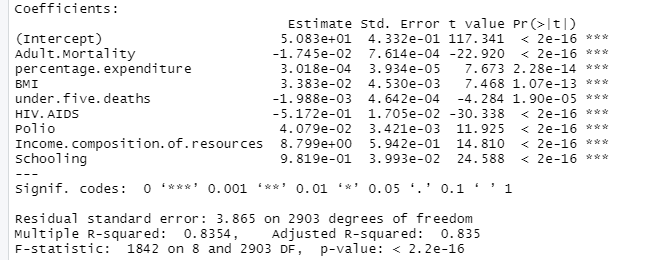


**Model 13 confidence interval:**



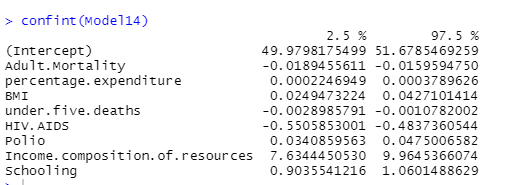
**Appendix 13:**

**Model 14:**

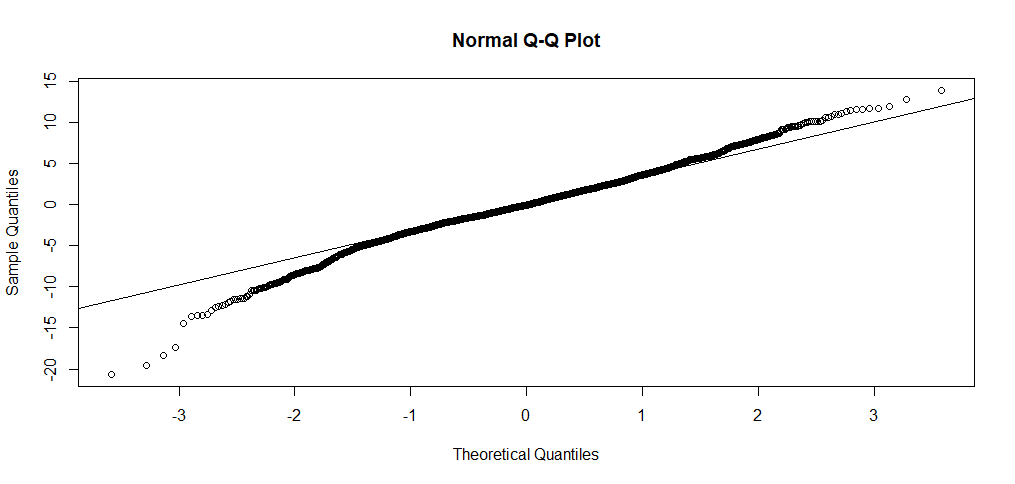




**Model 14: Confidence interval**



**Model 14 Normal probability plot:**



**Model 14 fitted vs residual**

