Life Expectancy (WHO)

Statistical Analysis on impact of factors influencing Life Expectancy

ISyE 6414 - Section A
Team 2

Shiven Amol Barbare	903926493
Karan Nahar	903933893
Prasanthi Toram	903924555
Manikant Thatipalli	903582374
Priyanka Singh	903933981

Table of Contents

1. **Introduction**

2. **Exploratory Data Analysis**

3. **Model Fitting and Diagnostics**

4. **Model Performance Comparison**

5. **Conclusions and Further Scope**

Introduction

Problem Description

Life Expectancy Overview:

- Life expectancy estimates the average years a person is expected to live based on current mortality rates.
- Key indicator for public health and policy-making, reflecting population health and well-being.

Global Research Challenges:

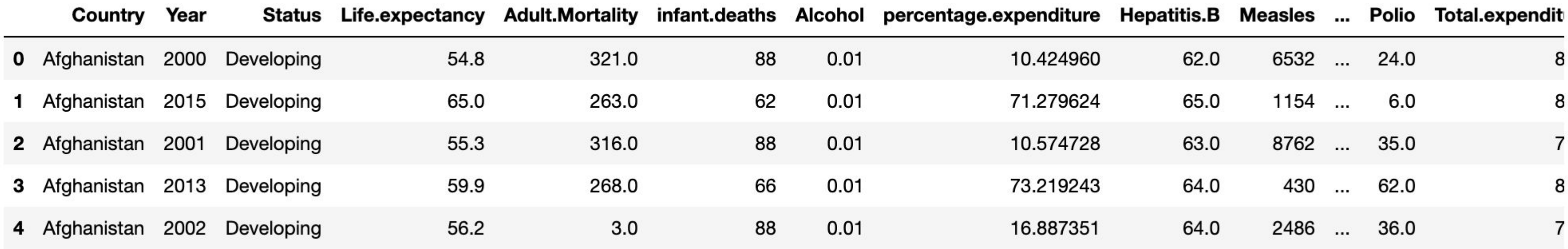
- Active research by governments and health agencies worldwide.
- Developing countries face substantial knowledge gaps, relying on international estimates, leading to potential misrepresentations.

Our Focus:

- Thorough statistical analysis on global factors: immunization, mortality, economics, social aspects, and health-related factors affecting life expectancy.
- Build a holistic model that encompasses these aspects to predict life expectancy across approximately 193 countries, spanning from underdeveloped and developing nations to developed countries.

About the Data

The data set comprises several health-related factors that affect the life expectancy for populations across the globe. The data has been collected for 193 countries over 16 years (2000-2015). The data is a combination of health, social and economic factors that affect the overall health status of the population.



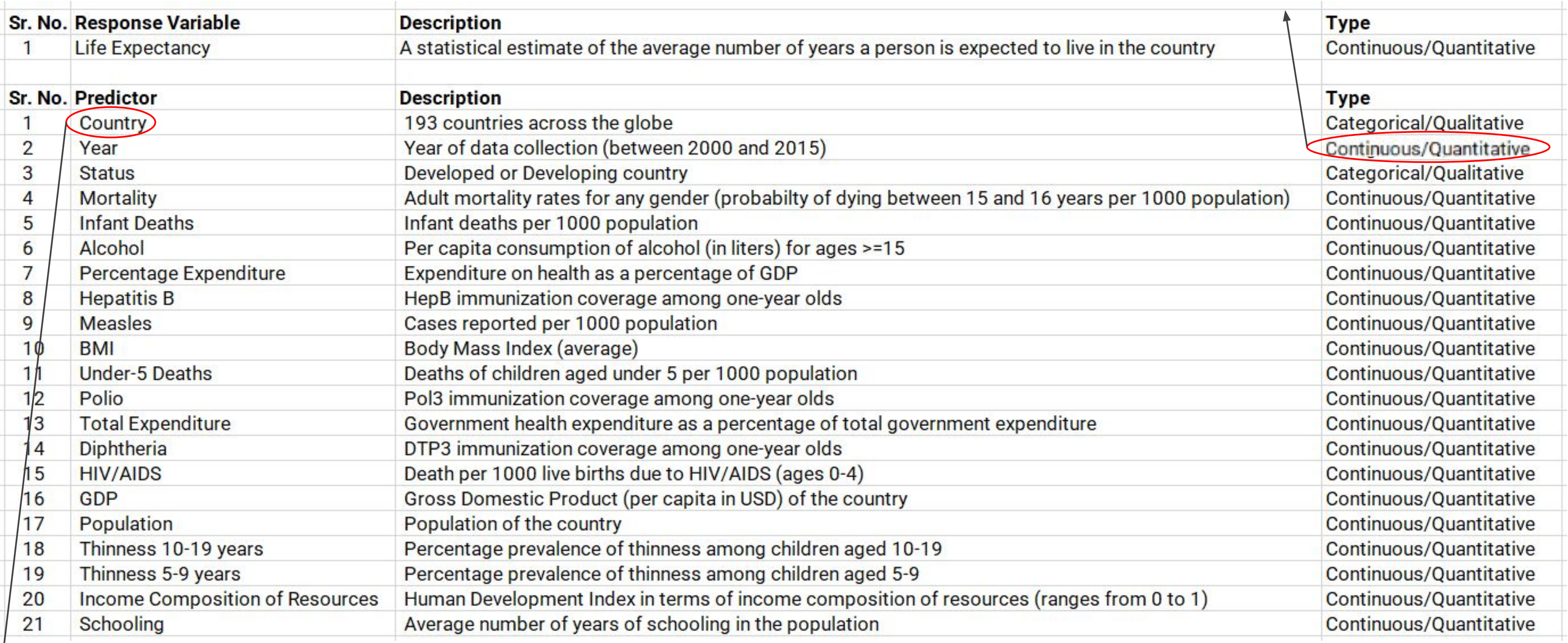
	Country	Year	Status	Life.expectancy	Adult.Mortality	infant.deaths	Alcohol	percentage.expenditure	Hepatitis.B	Measles	...	Polio	Total.expendit
0	Afghanistan	2000	Developing	54.8	321.0	88	0.01	10.424960	62.0	6532	...	24.0	8
1	Afghanistan	2015	Developing	65.0	263.0	62	0.01	71.279624	65.0	1154	...	6.0	8
2	Afghanistan	2001	Developing	55.3	316.0	88	0.01	10.574728	63.0	8762	...	35.0	7
3	Afghanistan	2013	Developing	59.9	268.0	66	0.01	73.219243	64.0	430	...	62.0	8
4	Afghanistan	2002	Developing	56.2	3.0	88	0.01	16.887351	64.0	2486	...	36.0	7

Source of Data

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

About the Data - Variable Description

Too many categories (16); restricts model scope



Sr. No.	Response Variable	Description	Type
1	Life Expectancy	A statistical estimate of the average number of years a person is expected to live in the country	Continuous/Quantitative
Sr. No.	**Predictor**	**Description**	**Type**
1	Country	193 countries across the globe	Categorical/Qualitative
2	Year	Year of data collection (between 2000 and 2015)	Continuous/Quantitative
3	Status	Developed or Developing country	Categorical/Qualitative
4	Mortality	Adult mortality rates for any gender (probabilty of dying between 15 and 16 years per 1000 population)	Continuous/Quantitative
5	Infant Deaths	Infant deaths per 1000 population	Continuous/Quantitative
6	Alcohol	Per capita consumption of alcohol (in liters) for ages >=15	Continuous/Quantitative
7	Percentage Expenditure	Expenditure on health as a percentage of GDP	Continuous/Quantitative
8	Hepatitis B	HepB immunization coverage among one-year olds	Continuous/Quantitative
9	Measles	Cases reported per 1000 population	Continuous/Quantitative
10	BMI	Body Mass Index (average)	Continuous/Quantitative
11	Under-5 Deaths	Deaths of children aged under 5 per 1000 population	Continuous/Quantitative
12	Polio	Pol3 immunization coverage among one-year olds	Continuous/Quantitative
13	Total Expenditure	Government health expenditure as a percentage of total government expenditure	Continuous/Quantitative
14	Diphtheria	DTP3 immunization coverage among one-year olds	Continuous/Quantitative
15	HIV/AIDS	Death per 1000 live births due to HIV/AIDS (ages 0-4)	Continuous/Quantitative
16	GDP	Gross Domestic Product (per capita in USD) of the country	Continuous/Quantitative
17	Population	Population of the country	Continuous/Quantitative
18	Thinness 10-19 years	Percentage prevalence of thinness among children aged 10-19	Continuous/Quantitative
19	Thinness 5-9 years	Percentage prevalence of thinness among children aged 5-9	Continuous/Quantitative
20	Income Composition of Resources	Human Development Index in terms of income composition of resources (ranges from 0 to 1)	Continuous/Quantitative
21	Schooling	Average number of years of schooling in the population	Continuous/Quantitative

Mapped to **Continent**: Instead of using 193 factors, we used 6 factors (for the 6 continents)

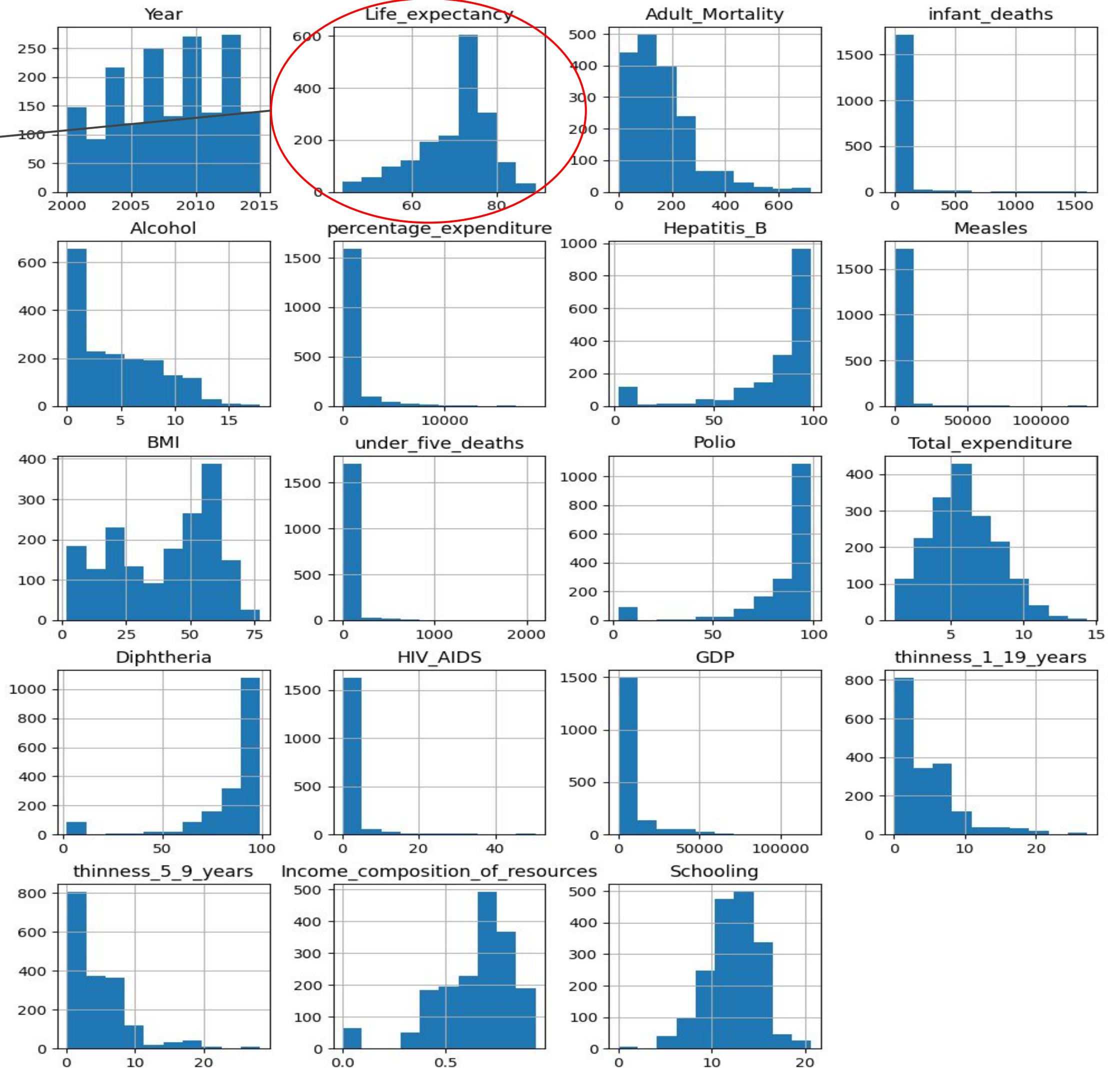
Data Cleaning and Preprocessing

Issue	Resolution
Missing population data	Drop Population column (other variables account for it)
Countries with only one data point	Drop respective rows
Missing/NA values	Drop respective rows
Wide range of predictor values	Min-Max Scaling (0-1) before model fitting

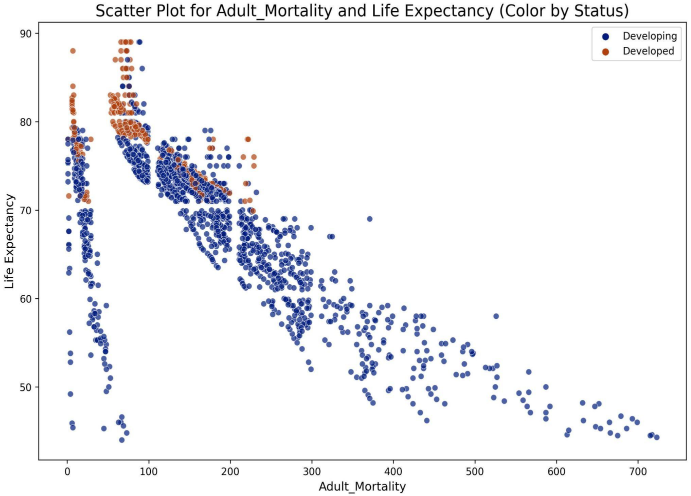
Exploratory Data Analysis

Distribution of Variables

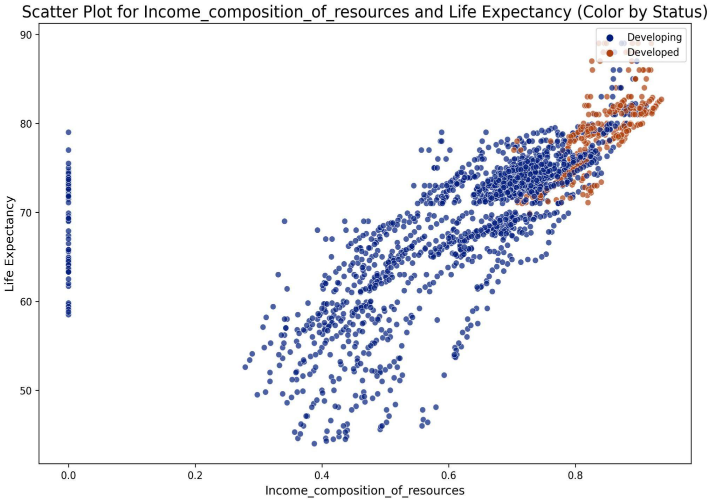
Response variable: Approximately Normal Distribution



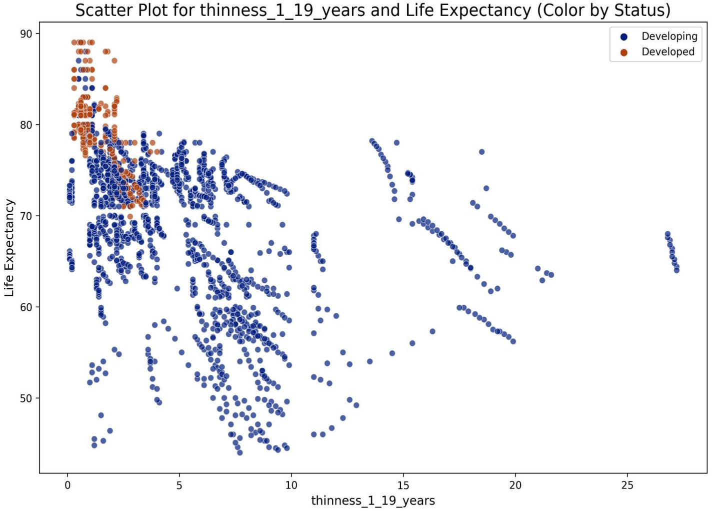
Variation of Life Expectancy with different predictors



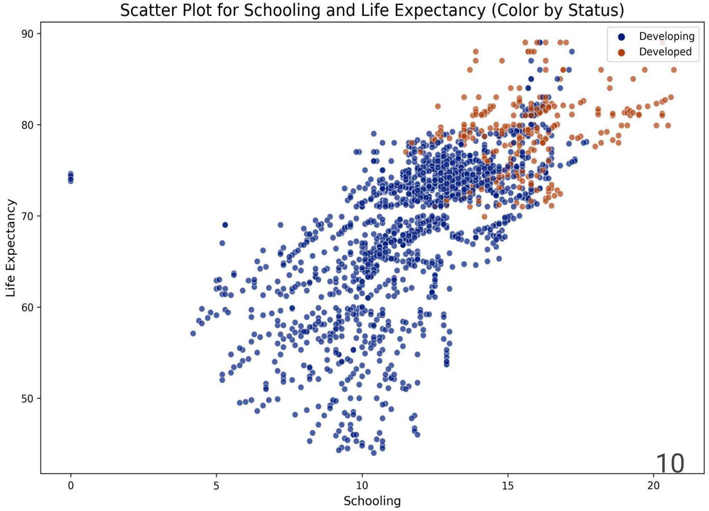
Scatter Plot for Adult_Mortality and Life Expectancy (Color by Status)



Scatter Plot for Income_composition_of_resources and Life Expectancy (Color by Status)

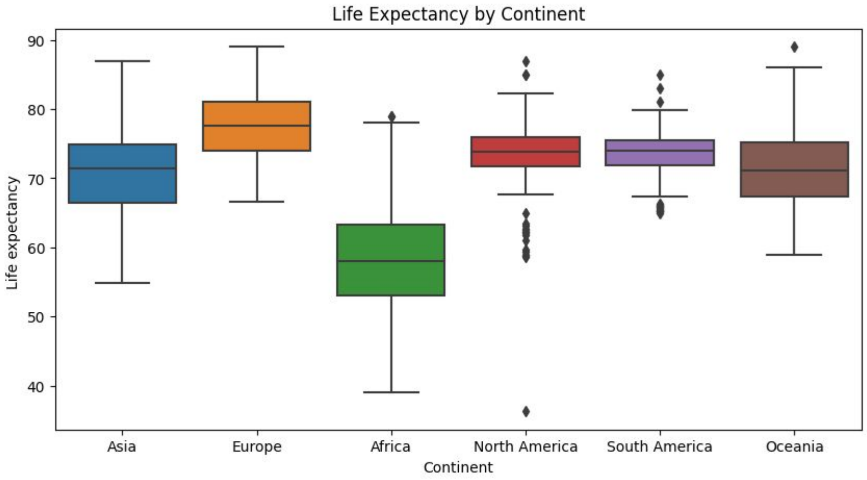


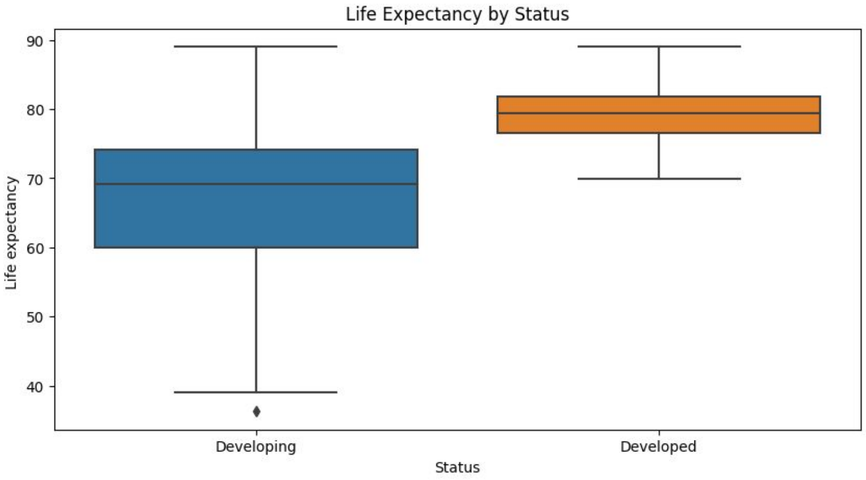
Scatter Plot for thinness_1_19_years and Life Expectancy (Color by Status)



Scatter Plot for Schooling and Life Expectancy (Color by Status)

Variation of Life Expectancy with Continent and Status



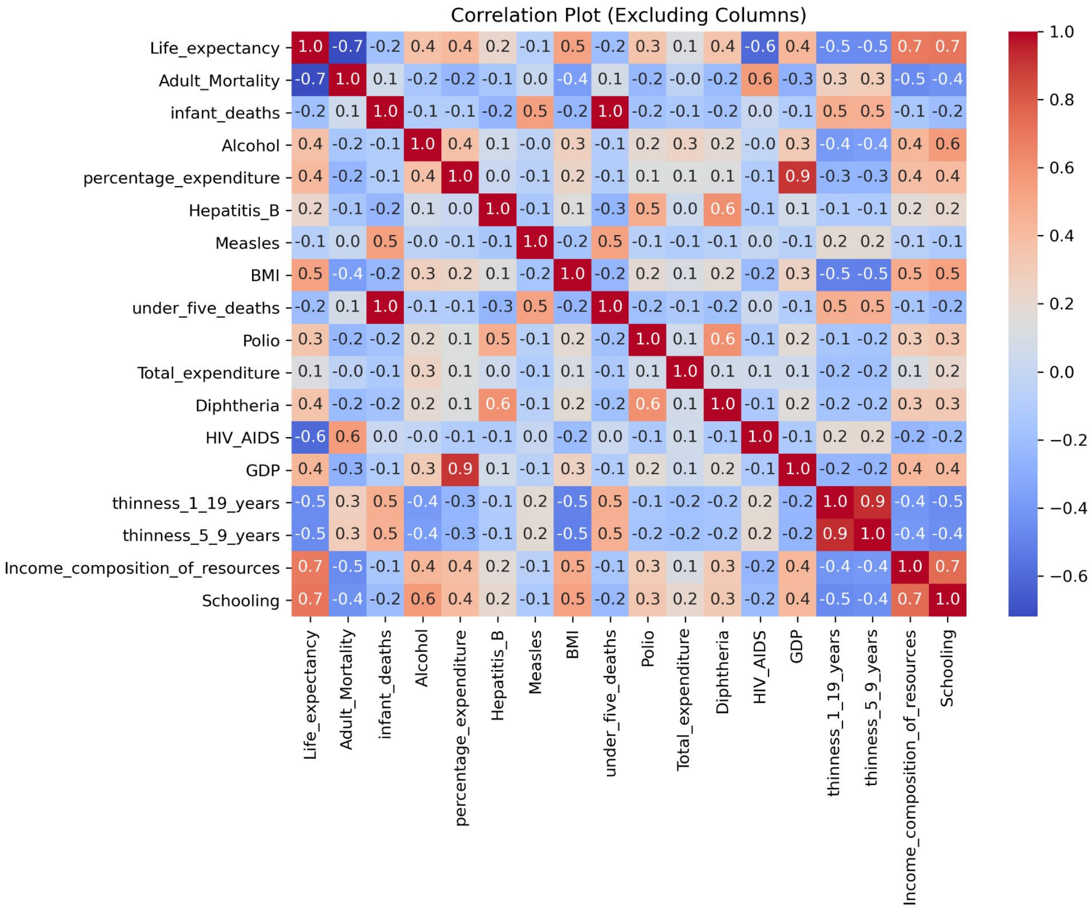


We can observe a clear difference in the median values of life expectancy for different continents. The life expectancy for Europe is much higher than that of Africa, for instance.

The variability is also significantly different for some continent pairs like North America and Africa.

We can observe a clear difference in the median values of life expectancy for developed and developing countries. The life expectancy for developed countries is much higher than that of developing countries, which is makes sense, intuitively.

Correlation Matrix

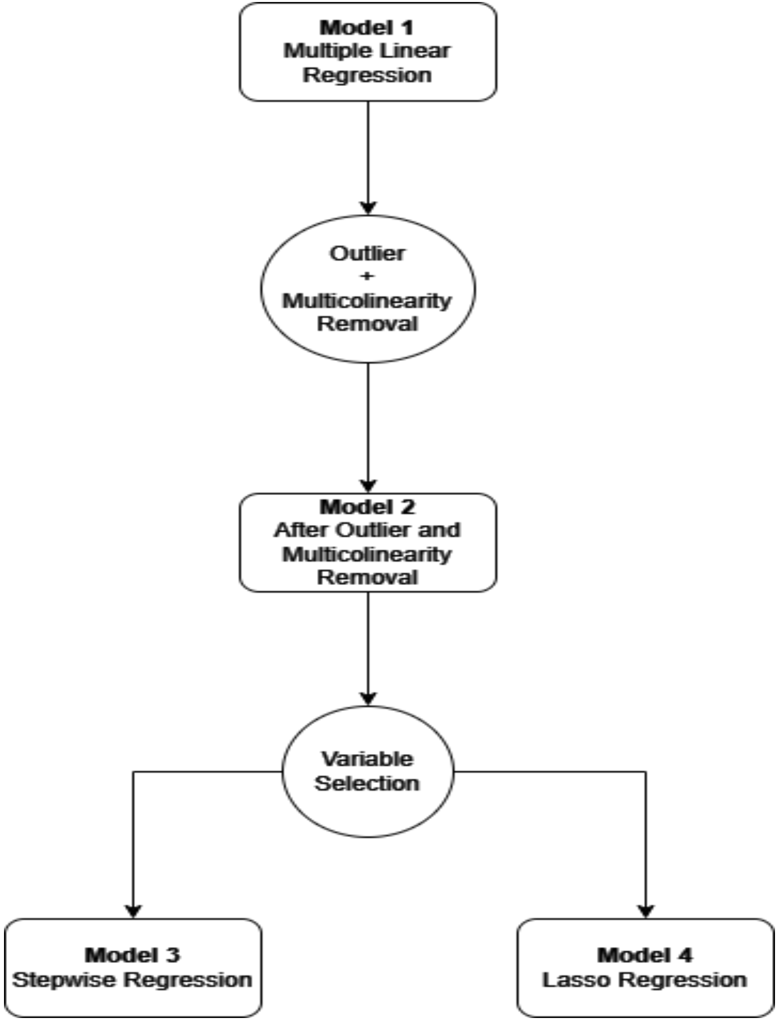


Correlation Plot (Excluding Columns)

	Life_expectancy	Adult_Mortality	infant_deaths	Alcohol	percentage_expenditure	Hepatitis_B	Measles	BMI	under_five_deaths	Polio	Total_expenditure	Diphtheria	HIV_AIDS	GDP	thinness_1_19_years	thinness_5_9_years	Income_composition_of_resources	Schooling
Life_expectancy	1.0	-0.7	-0.2	0.4	0.4	0.2	-0.1	0.5	-0.2	0.3	0.1	0.4	-0.6	0.4	-0.5	-0.5	0.7	0.7
Adult_Mortality	-0.7	1.0	0.1	-0.2	-0.2	-0.1	0.0	-0.4	0.1	-0.2	-0.0	-0.2	0.6	-0.3	0.3	0.3	-0.5	-0.4
infant_deaths	-0.2	0.1	1.0	-0.1	-0.1	-0.2	0.5	-0.2	1.0	-0.2	-0.1	-0.2	0.0	-0.1	0.5	0.5	-0.1	-0.2
Alcohol	0.4	-0.2	-0.1	1.0	0.4	0.1	-0.0	0.3	-0.1	0.2	0.3	0.2	-0.0	0.3	-0.4	-0.4	0.4	0.6
percentage_expenditure	0.4	-0.2	-0.1	0.4	1.0	0.0	-0.1	0.2	-0.1	0.1	0.1	0.1	-0.1	0.9	-0.3	-0.3	0.4	0.4
Hepatitis_B	0.2	-0.1	-0.2	0.1	0.0	1.0	-0.1	0.1	-0.3	0.5	0.0	0.6	-0.1	0.1	-0.1	-0.1	0.2	0.2
Measles	-0.1	0.0	0.5	-0.0	-0.1	-0.1	1.0	-0.2	0.5	-0.1	-0.1	-0.1	0.0	-0.1	0.2	0.2	-0.1	-0.1
BMI	0.5	-0.4	-0.2	0.3	0.2	0.1	-0.2	1.0	-0.2	0.2	0.1	0.2	-0.2	0.3	-0.5	-0.5	0.5	0.5
under_five_deaths	-0.2	0.1	1.0	-0.1	-0.1	-0.3	0.5	-0.2	1.0	-0.2	-0.1	-0.2	0.0	-0.1	0.5	0.5	-0.1	-0.2
Polio	0.3	-0.2	-0.2	0.2	0.1	0.5	-0.1	0.2	-0.2	1.0	0.1	0.6	-0.1	0.2	-0.1	-0.2	0.3	0.3
Total_expenditure	0.1	-0.0	-0.1	0.3	0.1	0.0	-0.1	0.1	-0.1	0.1	1.0	0.1	0.1	0.1	-0.2	-0.2	0.1	0.2
Diphtheria	0.4	-0.2	-0.2	0.2	0.1	0.6	-0.1	0.2	-0.2	0.6	0.1	1.0	-0.1	0.2	-0.2	-0.2	0.3	0.3
HIV_AIDS	-0.6	0.6	0.0	-0.0	-0.1	-0.1	0.0	-0.2	0.0	-0.1	0.1	-0.1	1.0	-0.1	0.2	0.2	-0.2	-0.2
GDP	0.4	-0.3	-0.1	0.3	0.9	0.1	-0.1	0.3	-0.1	0.2	0.1	0.2	-0.1	1.0	-0.2	-0.2	0.4	0.4
thinness_1_19_years	-0.5	0.3	0.5	-0.4	-0.3	-0.1	0.2	-0.5	0.5	-0.1	-0.2	-0.2	0.2	-0.2	1.0	0.9	-0.4	-0.5
thinness_5_9_years	-0.5	0.3	0.5	-0.4	-0.3	-0.1	0.2	-0.5	0.5	-0.2	-0.2	-0.2	0.2	-0.2	0.9	1.0	-0.4	-0.4
Income_composition_of_resources	0.7	-0.5	-0.1	0.4	0.4	0.2	-0.1	0.5	-0.1	0.3	0.1	0.3	-0.2	0.4	-0.4	-0.4	1.0	0.7
Schooling	0.7	-0.4	-0.2	0.6	0.4	0.2	-0.1	0.5	-0.2	0.3	0.2	0.3	-0.2	0.4	-0.5	-0.4	0.7	1.0

- Predictor variables such as Schooling, Income composition of resources have high positive correlation with the response variable.
- Predictor variables such as Adult Mortality, HIV AIDS have high negative correlation with the response variable.

Model Fitting and Diagnostics



Model 1
Multiple Linear Regression

Outlier
+
Multicolinearity Removal

Model 2
After Outlier and Multicolinearity Removal

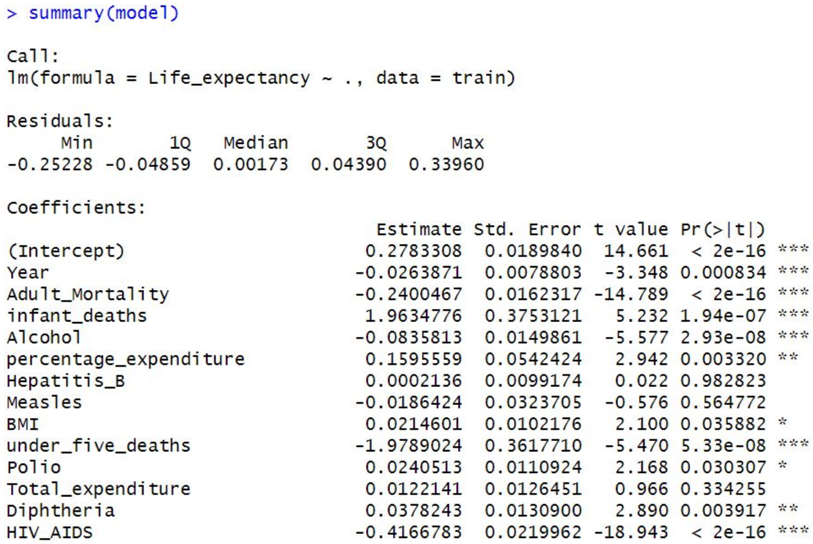
Variable Selection

Model 3
Stepwise Regression

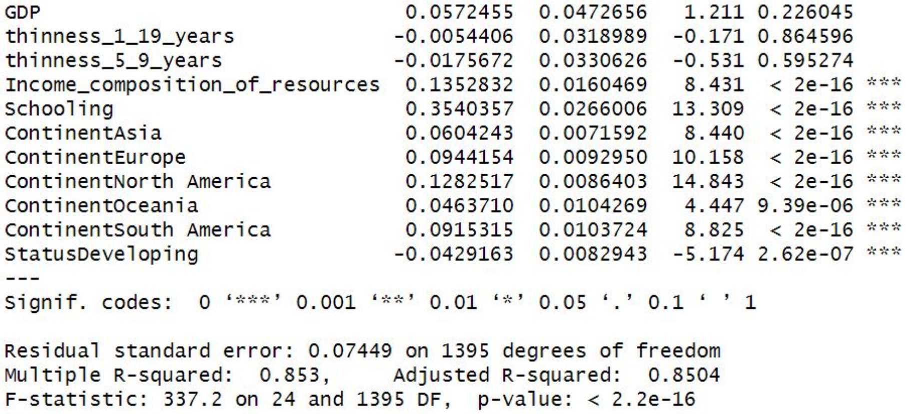
Model 4
Lasso Regression

Modeling Overview

Model I: Full Model with all Predictors



```
> summary(model)

Call:
lm(formula = Life_expectancy ~ ., data = train)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.25228 -0.04859  0.00173  0.04390  0.33960 

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)                     0.2783308  0.0189840  14.661  < 2e-16 ***
Year                           -0.0263871  0.0078803  -3.348 0.000834 ***
Adult_Mortality                -0.2400467  0.0162317 -14.789  < 2e-16 ***
infant_deaths                   1.9634776  0.3753121   5.232 1.94e-07 ***
Alcohol                        -0.0835813  0.0149861  -5.577 2.93e-08 ***
percentage_expenditure          0.1595559  0.0542424   2.942 0.003320 ** 
Hepatitis_B                     0.0002136  0.0099174   0.022 0.982823    
Measles                        -0.0186424  0.0323705  -0.576 0.564772    
BMI                             0.0214601  0.0102176   2.100 0.035882 *  
under_five_deaths              -1.9789024  0.3617710  -5.470 5.33e-08 ***
Polio                           0.0240513  0.0110924   2.168 0.030307 *  
Total_expenditure               0.0122141  0.0126451   0.966 0.334255    
Diphtheria                      0.0378243  0.0130900   2.890 0.003917 ** 
HIV_AIDS                       -0.4166783  0.0219962 -18.943  < 2e-16 ***
```

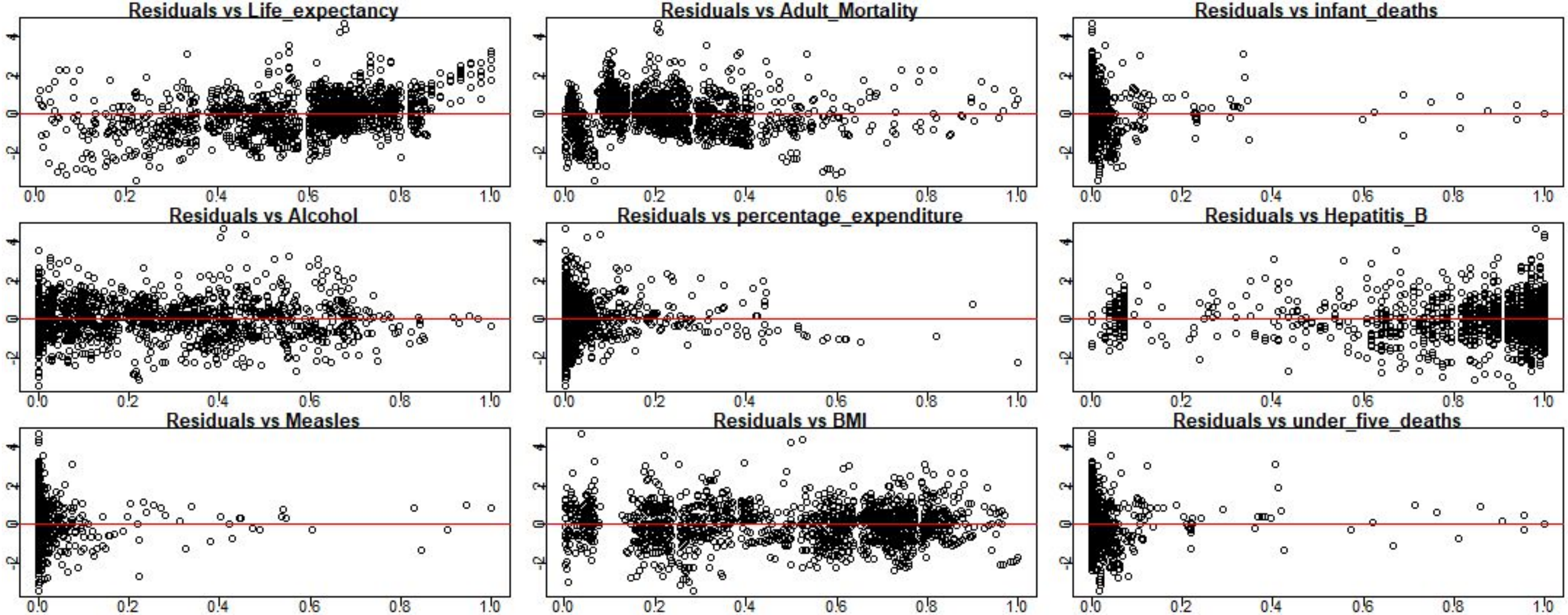


```
GDP                             0.0572455  0.0472656   1.211 0.226045    
thinness_1_19_years            -0.0054406  0.0318989  -0.171 0.864596    
thinness_5_9_years             -0.0175672  0.0330626  -0.531 0.595274    
Income_composition_of_resources  0.1352832  0.0160469   8.431  < 2e-16 ***
Schooling                       0.3540357  0.0266006  13.309  < 2e-16 ***
ContinentAsia                   0.0604243  0.0071592   8.440  < 2e-16 ***
ContinentEurope                 0.0944154  0.0092950  10.158  < 2e-16 ***
ContinentNorth America          0.1282517  0.0086403  14.843  < 2e-16 ***
ContinentOceania                0.0463710  0.0104269   4.447 9.39e-06 ***
ContinentSouth America          0.0915315  0.0103724   8.825  < 2e-16 ***
StatusDeveloping               -0.0429163  0.0082943  -5.174 2.62e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07449 on 1395 degrees of freedom
Multiple R-squared:  0.853,	Adjusted R-squared:  0.8504 
F-statistic: 337.2 on 24 and 1395 DF,  p-value: < 2.2e-16
```

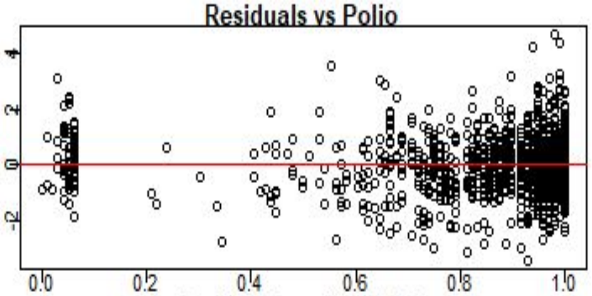
Findings:

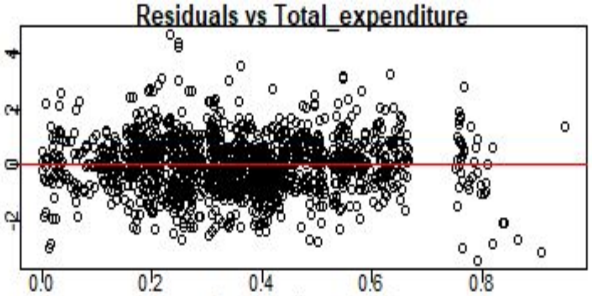
- Adjusted R^2 is 0.8504 i.e. model explains about 85% variability in the data
- The F-statistic is large and corresponding p-value is small, so the model as a whole is significant

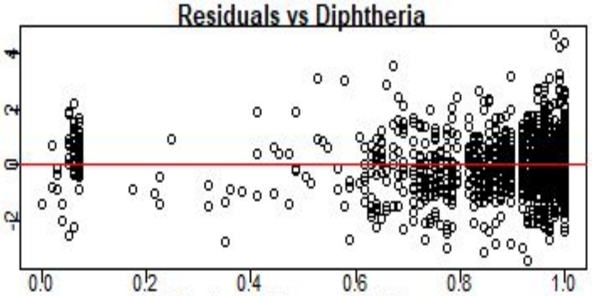
Model I: Residual Analysis (Linearity)

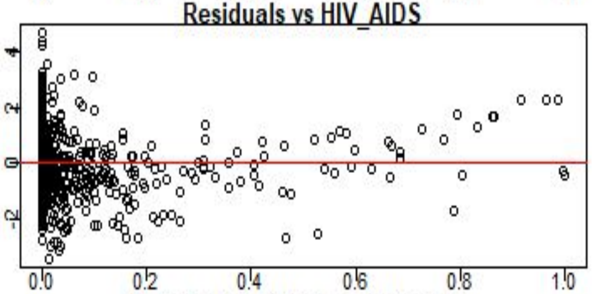


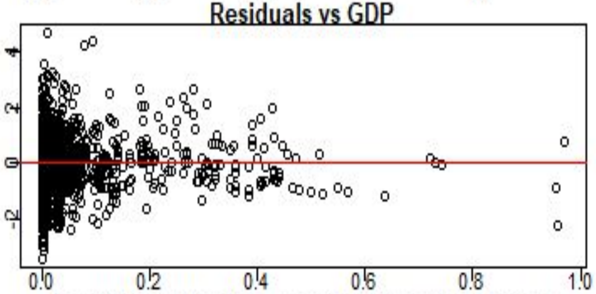
Model I: Residual Analysis (Linearity)

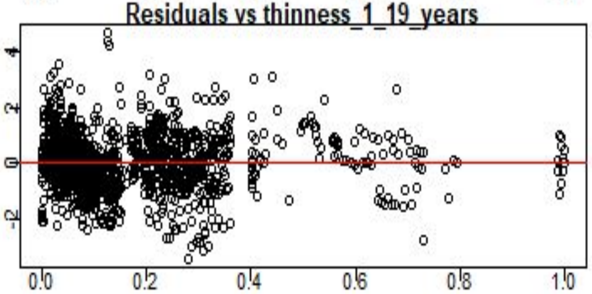


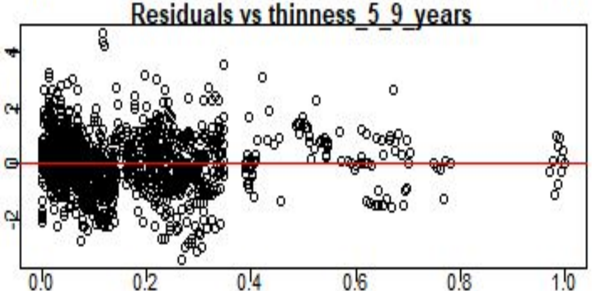


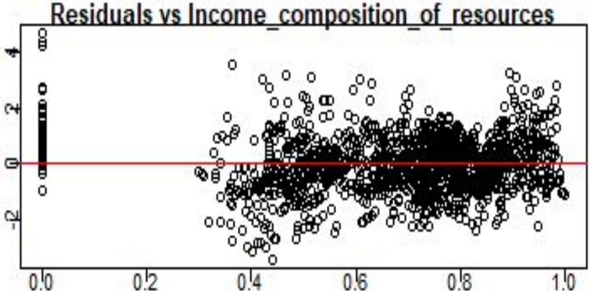








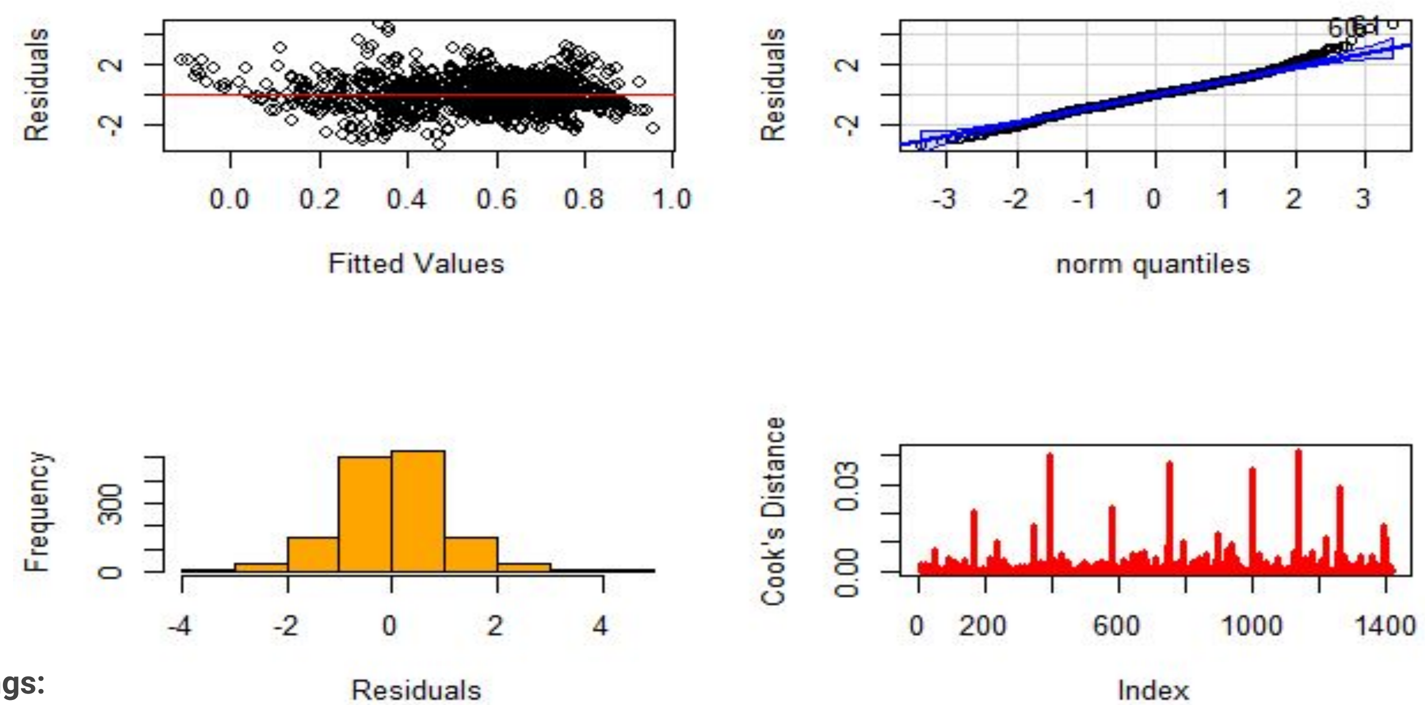




Findings:

- Residuals are randomly distributed across predictor variables implying linear relation with the response variable.

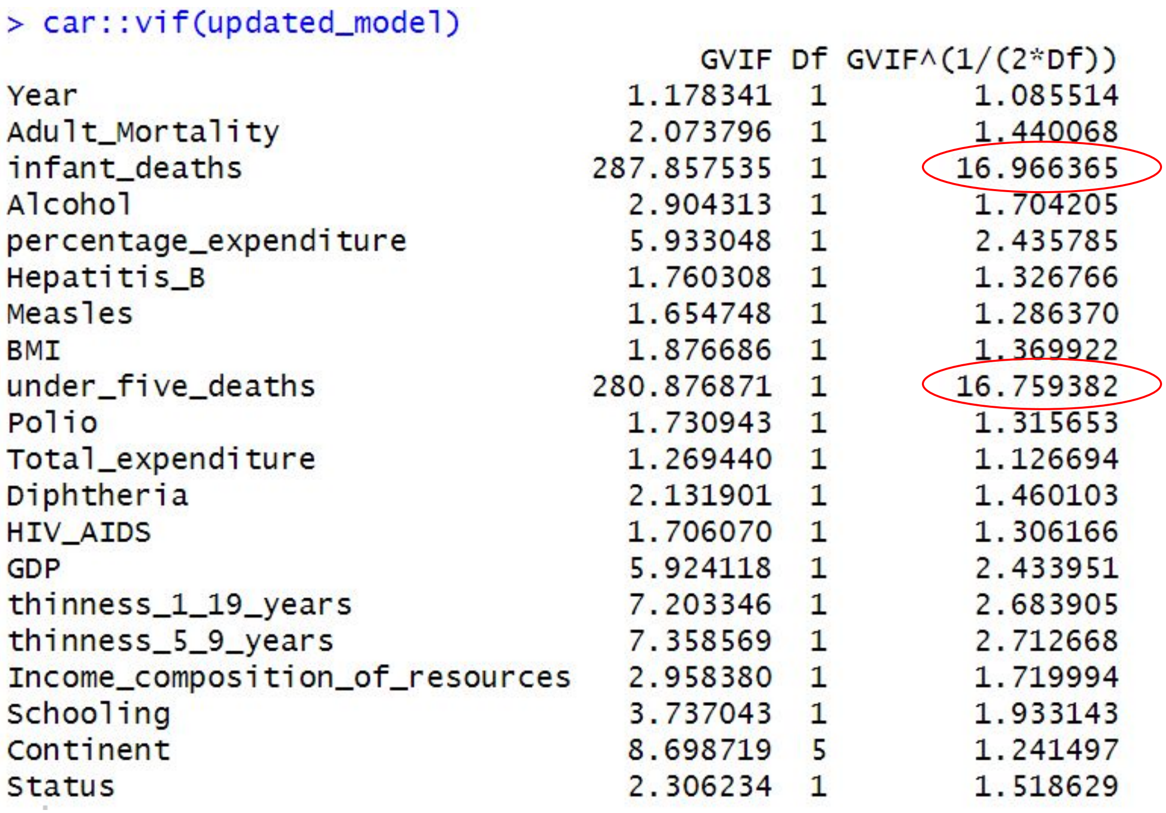
Model I: Residual Analysis (Variance & Normality)



Findings:

- The residual vs fitted values curve shows randomly distributed residuals with no pattern, implying uncorrelated errors and constant variance
- The QQ plot shows that the residuals have an approximately normal distribution but the distribution has **tail** which needs to be examined.
- The Cook's distance plot shows the presence of a few **outliers**

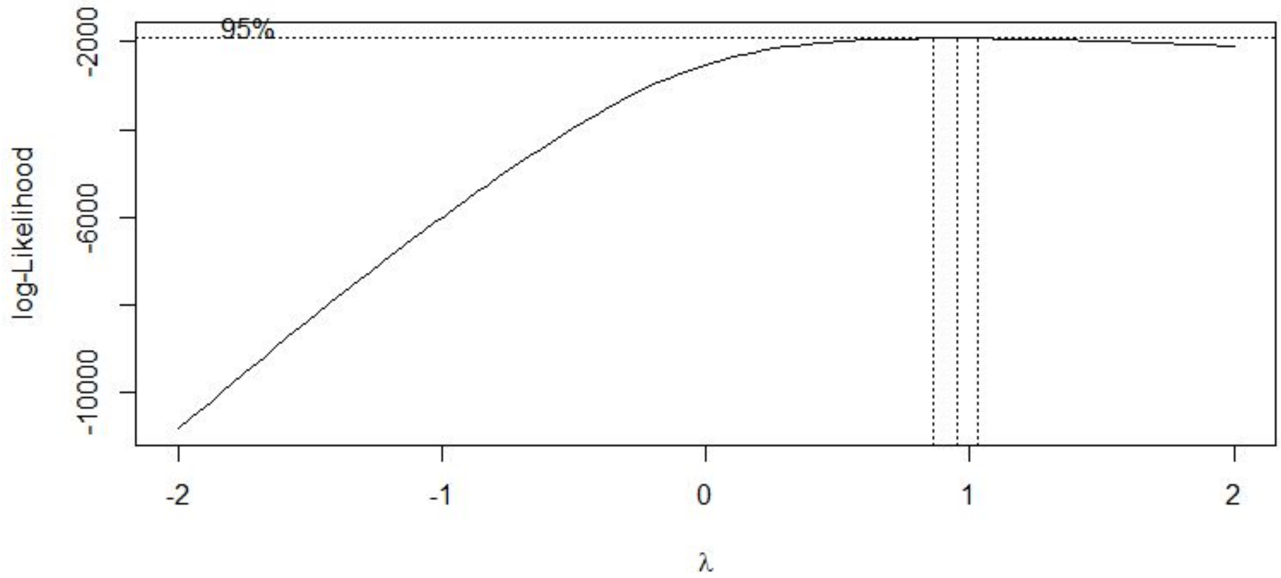
Multicollinearity



```
> car::vif(updated_model)
                                      GVIF Df GVIF^(1/(2*Df))
Year                              1.178341  1        1.085514
Adult_Mortality                   2.073796  1        1.440068
infant_deaths                   287.857535  1       16.966365
Alcohol                           2.904313  1        1.704205
percentage_expenditure            5.933048  1        2.435785
Hepatitis_B                       1.760308  1        1.326766
Measles                           1.654748  1        1.286370
BMI                               1.876686  1        1.369922
under_five_deaths               280.876871  1       16.759382
Polio                             1.730943  1        1.315653
Total_expenditure                 1.269440  1        1.126694
Diphtheria                        2.131901  1        1.460103
HIV_AIDS                          1.706070  1        1.306166
GDP                               5.924118  1        2.433951
thinness_1_19_years               7.203346  1        2.683905
thinness_5_9_years                7.358569  1        2.712668
Income_composition_of_resources   2.958380  1        1.719994
Schooling                         3.737043  1        1.933143
Continent                         8.698719  5        1.241497
Status                            2.306234  1        1.518629
```

- Variables - infant_deaths and under_five_deaths have very high VIF values
- This suggests that **multicollinearity** may be an issue in the model

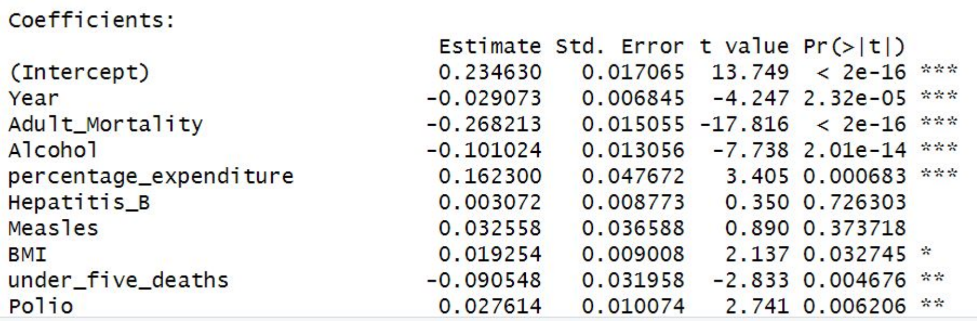
Box Cox Transformation for Response



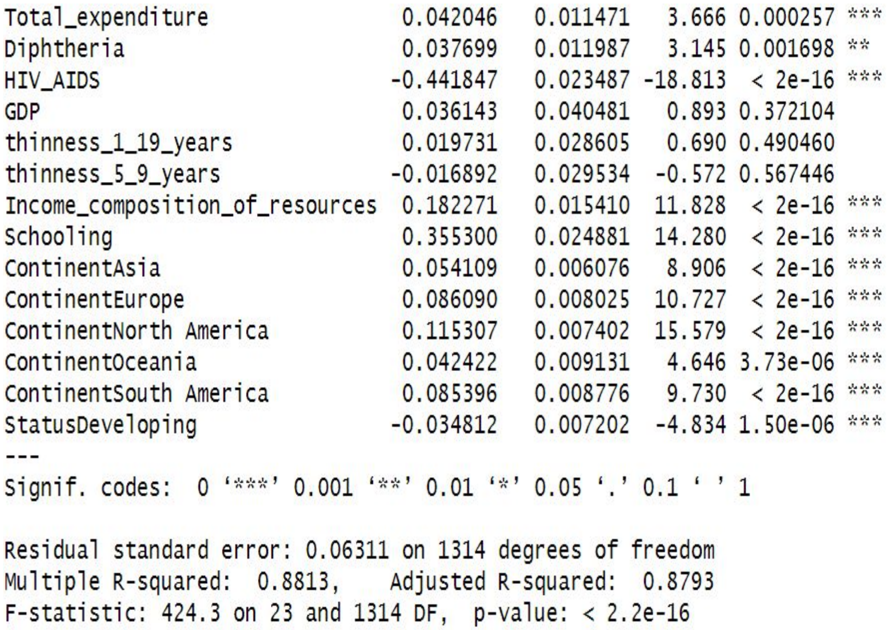
Findings:

- The optimal value of λ (power) provided by the Box Cox Transformation comes out to be around 0.9 (close to 1)
- This suggests that there is **no need for transforming** the response variable

Model II: Outlier Removal & Accounting for Multicollinearity



```
Coefficients:
                           Estimate Std. Error t value Pr(>|t|)    
(Intercept)                0.234630   0.017065  13.749  < 2e-16 ***
Year                      -0.029073   0.006845  -4.247 2.32e-05 ***
Adult_Mortality           -0.268213   0.015055 -17.816  < 2e-16 ***
Alcohol                   -0.101024   0.013056  -7.738 2.01e-14 ***
percentage_expenditure     0.162300   0.047672   3.405 0.000683 ***
Hepatitis_B                0.003072   0.008773   0.350 0.726303    
Measles                    0.032558   0.036588   0.890 0.373718    
BMI                        0.019254   0.009008   2.137 0.032745 *  
under_five_deaths         -0.090548   0.031958  -2.833 0.004676 ** 
Polio                      0.027614   0.010074   2.741 0.006206 ** 
```

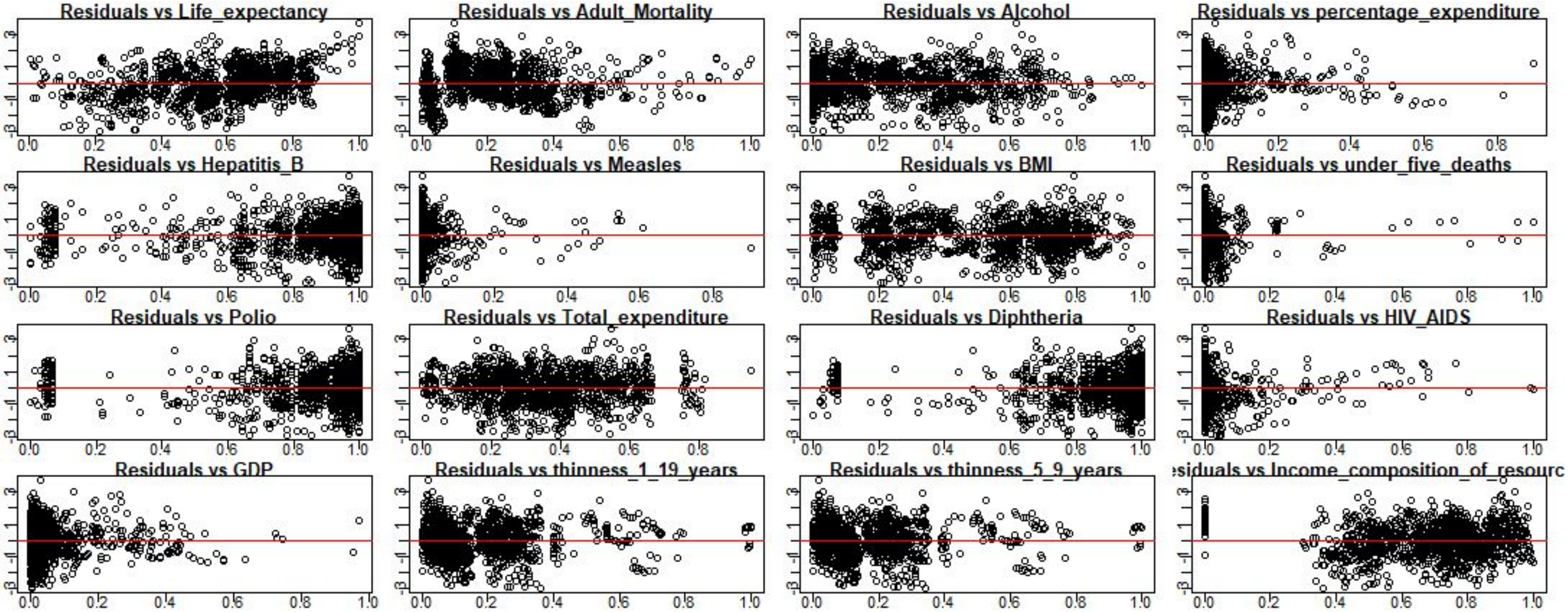


```
Total_expenditure                0.042046   0.011471   3.666 0.000257 ***
Diphtheria                       0.037699   0.011987   3.145 0.001698 ** 
HIV_AIDS                        -0.441847   0.023487 -18.813  < 2e-16 ***
GDP                              0.036143   0.040481   0.893 0.372104    
thinness_1_19_years              0.019731   0.028605   0.690 0.490460    
thinness_5_9_years              -0.016892   0.029534  -0.572 0.567446    
Income_composition_of_resources  0.182271   0.015410  11.828  < 2e-16 ***
Schooling                        0.355300   0.024881  14.280  < 2e-16 ***
ContinentAsia                    0.054109   0.006076   8.906  < 2e-16 ***
ContinentEurope                  0.086090   0.008025  10.727  < 2e-16 ***
ContinentNorth America           0.115307   0.007402  15.579  < 2e-16 ***
ContinentOceania                 0.042422   0.009131   4.646 3.73e-06 ***
ContinentSouth America           0.085396   0.008776   9.730  < 2e-16 ***
StatusDeveloping                -0.034812   0.007202  -4.834 1.50e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06311 on 1314 degrees of freedom
Multiple R-squared:  0.8813,	Adjusted R-squared:  0.8793 
F-statistic: 424.3 on 23 and 1314 DF,  p-value: < 2.2e-16
```

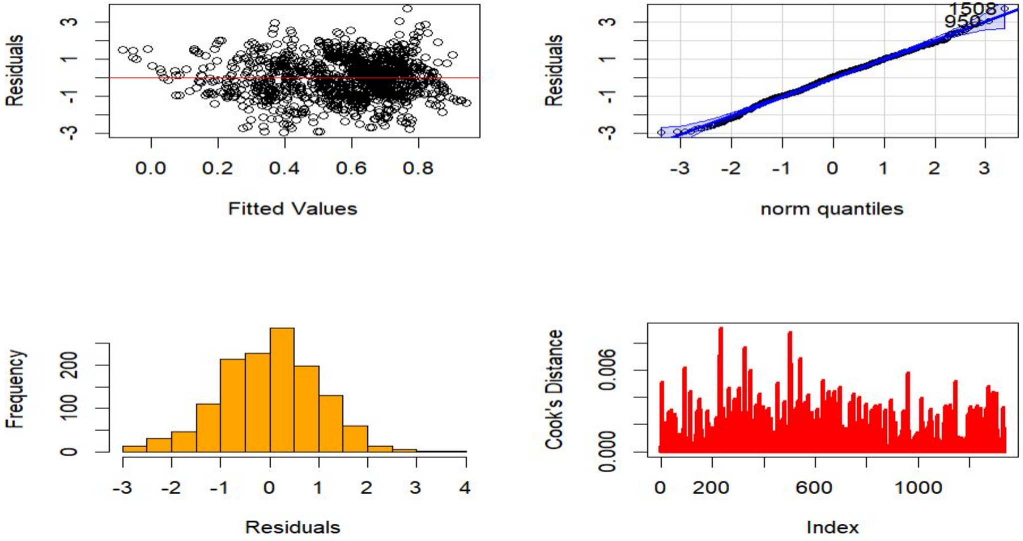
Findings:

- Adjusted R^2 is 0.8793 i.e. model explains about 88% variability in the data (**some improvement**)
- The F-statistic is large and corresponding p-value is small, so the model as a whole is significant

Model II: Residual Analysis (Linearity)



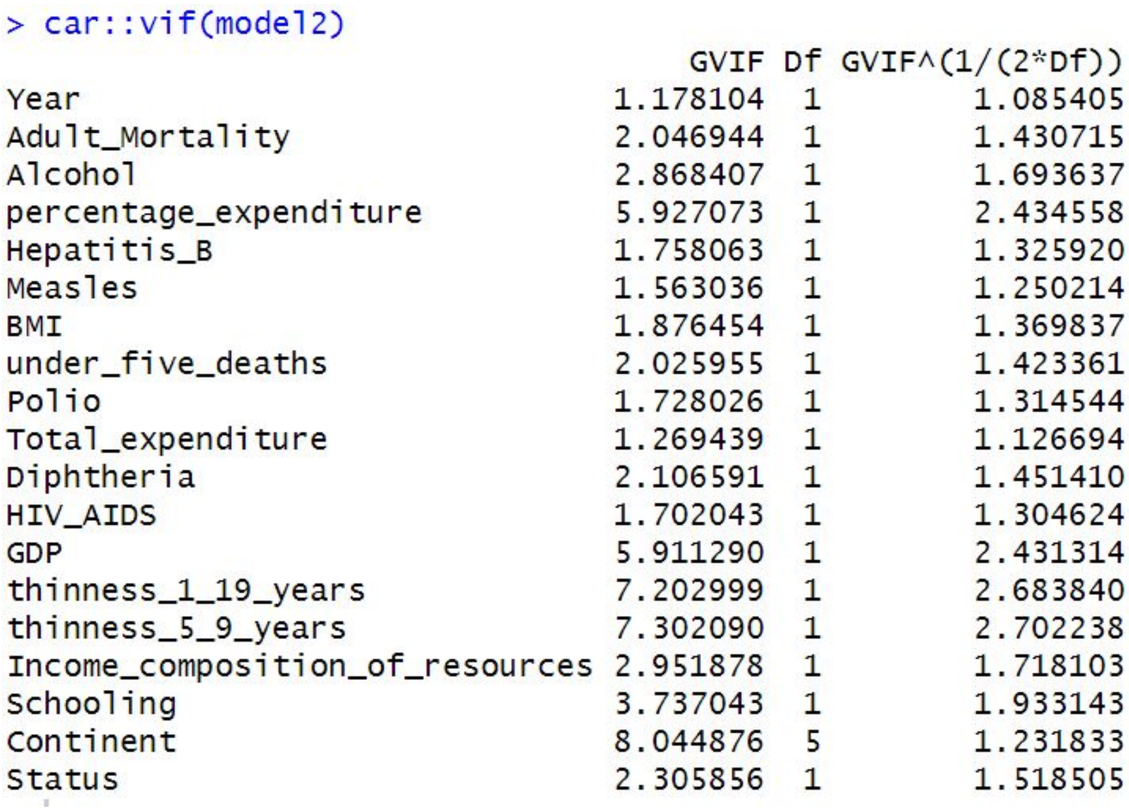
Model II: Residual Analysis (Variance and Normality)



Findings:

- The residual vs fitted values curve shows randomly distributed residuals with no pattern, implying uncorrelated errors and constant variance
- The QQ plot and histogram show that the residuals have an approximately normal distribution
- The cook's distance plot clearly shows that the outliers from Model 1 have been removed

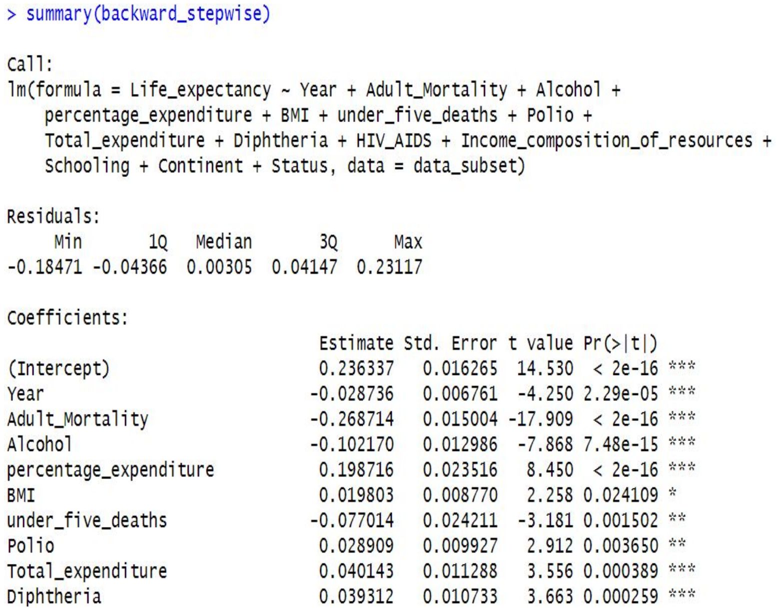
Multicollinearity



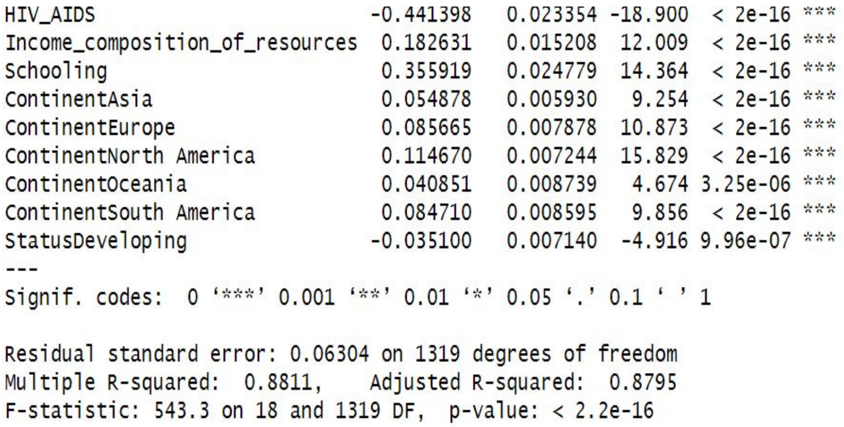
```
> car::vif(model2)
                                    GVIF Df GVIF^(1/(2*Df))
Year                            1.178104  1        1.085405
Adult_Mortality                 2.046944  1        1.430715
Alcohol                         2.868407  1        1.693637
percentage_expenditure          5.927073  1        2.434558
Hepatitis_B                     1.758063  1        1.325920
Measles                         1.563036  1        1.250214
BMI                             1.876454  1        1.369837
under_five_deaths               2.025955  1        1.423361
Polio                           1.728026  1        1.314544
Total_expenditure               1.269439  1        1.126694
Diphtheria                      2.106591  1        1.451410
HIV_AIDS                        1.702043  1        1.304624
GDP                             5.911290  1        2.431314
thinness_1_19_years             7.202999  1        2.683840
thinness_5_9_years              7.302090  1        2.702238
Income_composition_of_resources 2.951878  1        1.718103
Schooling                       3.737043  1        1.933143
Continent                       8.044876  5        1.231833
Status                          2.305856  1        1.518505
```

- VIF values are small for all predictors, which implies that multicollinearity is no longer a problem

Model III: Stepwise Regression



```
> summary(backward_stepwise)

Call:
lm(formula = Life_expectancy ~ Year + Adult_Mortality + Alcohol + 
    percentage_expenditure + BMI + under_five_deaths + Polio + 
    Total_expenditure + Diphtheria + HIV_AIDS + Income_composition_of_resources + 
    Schooling + Continent + Status, data = data_subset)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.18471 -0.04366  0.00305  0.04147  0.23117 

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)                      0.236337   0.016265  14.530  < 2e-16 ***
Year                            -0.028736   0.006761  -4.250 2.29e-05 ***
Adult_Mortality                 -0.268714   0.015004 -17.909  < 2e-16 ***
Alcohol                         -0.102170   0.012986  -7.868 7.48e-15 ***
percentage_expenditure           0.198716   0.023516   8.450  < 2e-16 ***
BMI                              0.019803   0.008770   2.258 0.024109 *  
under_five_deaths               -0.077014   0.024211  -3.181 0.001502 ** 
Polio                            0.028909   0.009927   2.912 0.003650 ** 
Total_expenditure                0.040143   0.011288   3.556 0.000389 ***
Diphtheria                       0.039312   0.010733   3.663 0.000259 ***
```

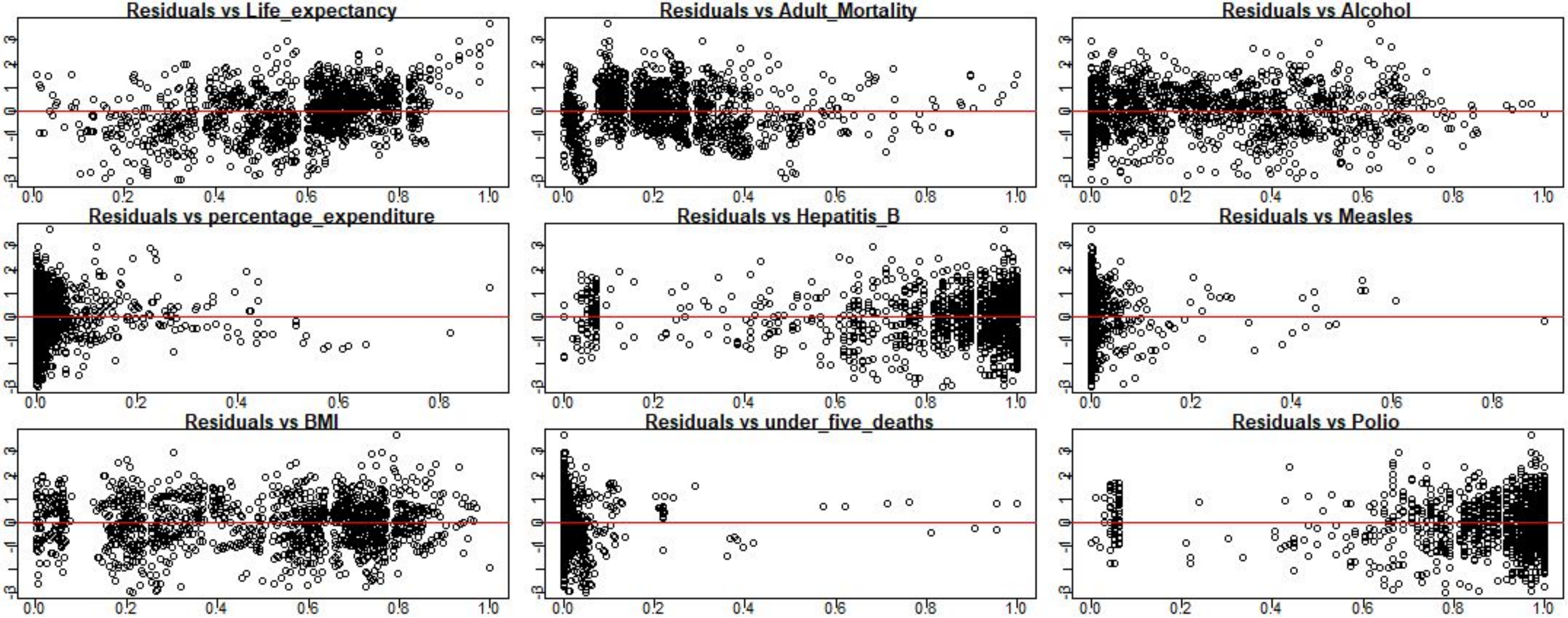


```
HIV_AIDS                        -0.441398   0.023354 -18.900  < 2e-16 ***
Income_composition_of_resources  0.182631   0.015208  12.009  < 2e-16 ***
Schooling                        0.355919   0.024779  14.364  < 2e-16 ***
ContinentAsia                    0.054878   0.005930   9.254  < 2e-16 ***
ContinentEurope                  0.085665   0.007878  10.873  < 2e-16 ***
ContinentNorth America           0.114670   0.007244  15.829  < 2e-16 ***
ContinentOceania                 0.040851   0.008739   4.674 3.25e-06 ***
ContinentSouth America           0.084710   0.008595   9.856  < 2e-16 ***
StatusDeveloping                -0.035100   0.007140  -4.916 9.96e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06304 on 1319 degrees of freedom
Multiple R-squared:  0.8811,	Adjusted R-squared:  0.8795 
F-statistic: 543.3 on 18 and 1319 DF,  p-value: < 2.2e-16
```

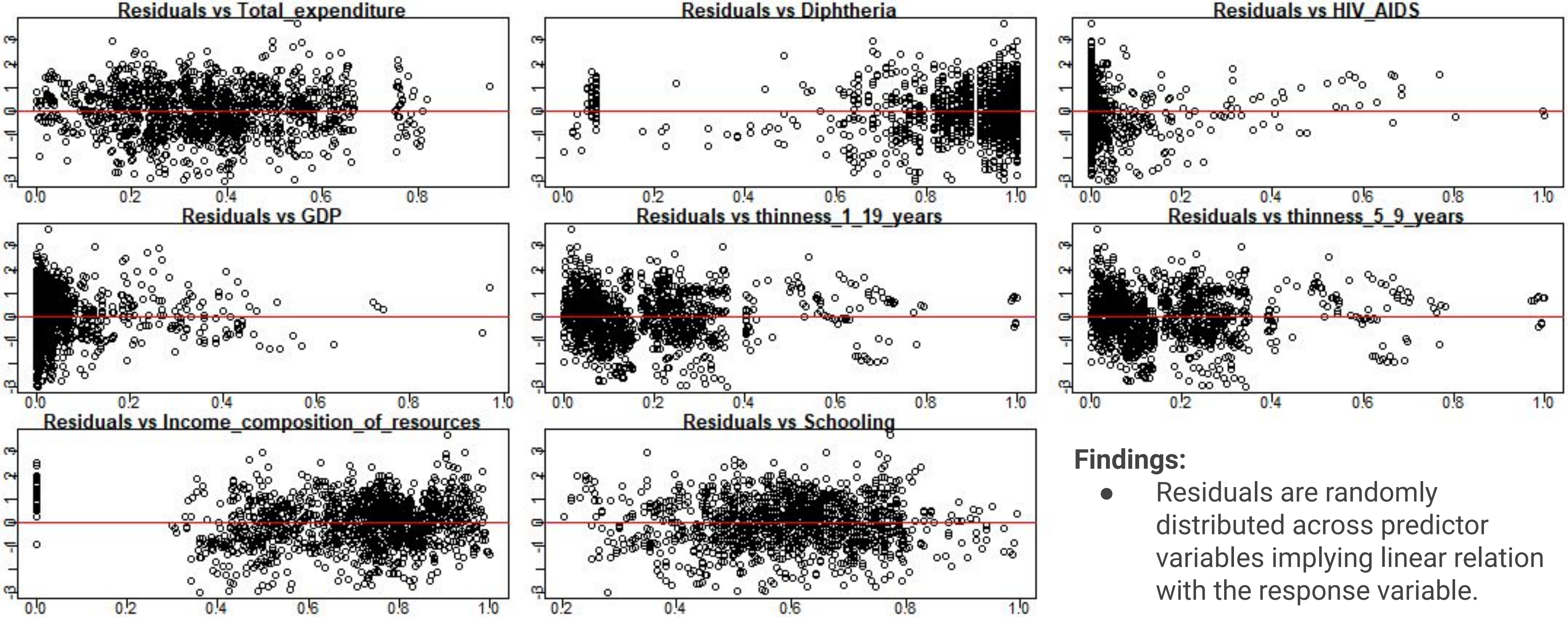
Findings:

- 14 variables are selected based on AIC values
- There is a slight improvement in adjusted R^2 value

Model III: Residual Analysis (Linearity)



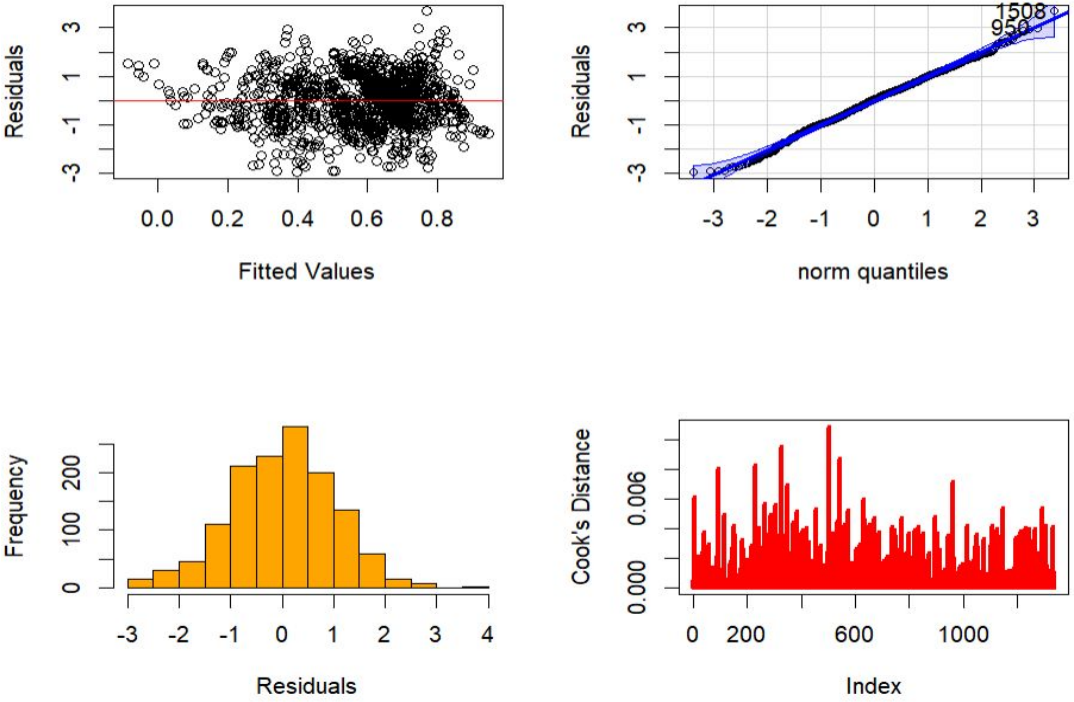
Model III: Residual Analysis (Linearity)



Findings:

- Residuals are randomly distributed across predictor variables implying linear relation with the response variable.

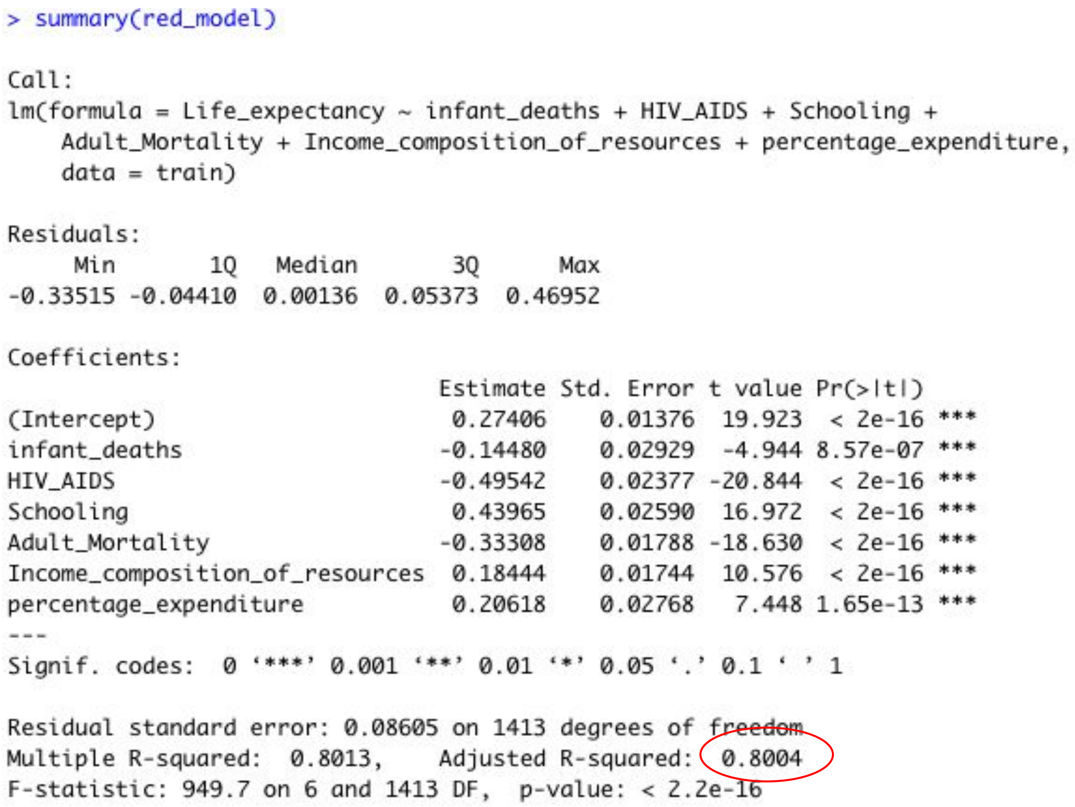
Model III: Residual Analysis (Variance and Normality)



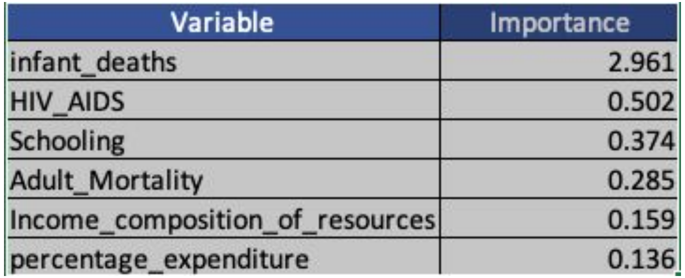
Findings:

- The residual vs fitted values curve shows randomly distributed residuals with no pattern, implying uncorrelated errors and constant variance
- The QQ plot and histogram show that the residuals have an approximately normal distribution
- Cook’s distances are smaller than the 4/n threshold so we do not have outliers

Model IV: Lasso Regression



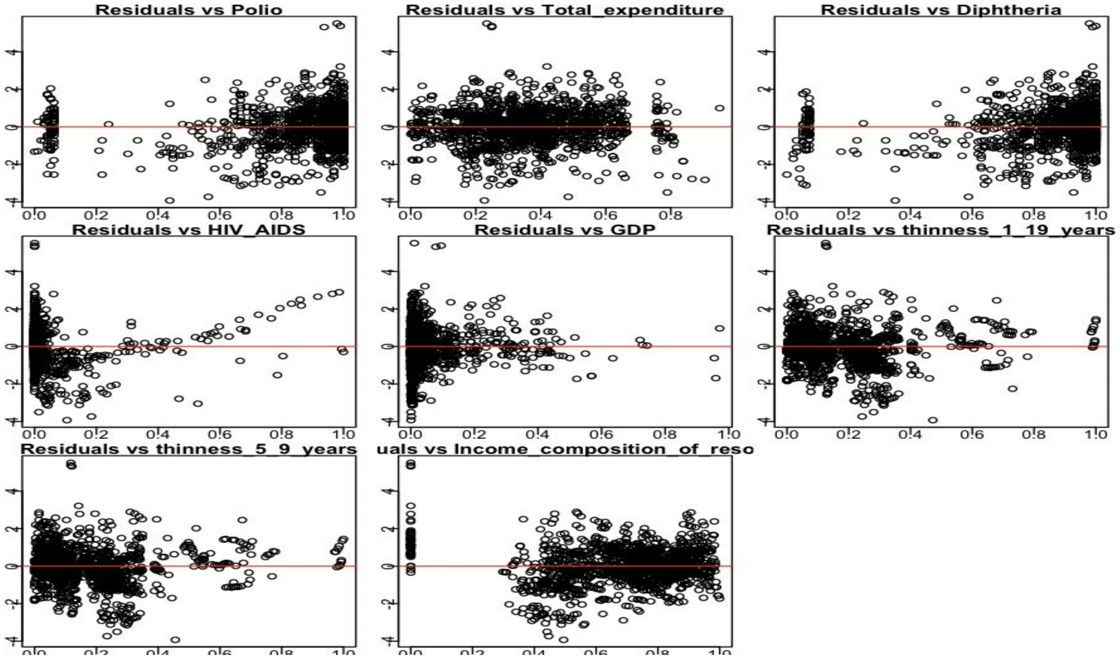
```
> summary(red_model)

Call:
lm(formula = Life_expectancy ~ infant_deaths + HIV_AIDS + Schooling +
    Adult_Mortality + Income_composition_of_resources + percentage_expenditure,
    data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-0.33515 -0.04410  0.00136  0.05373  0.46952

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                      0.27406    0.01376  19.923  < 2e-16 ***
infant_deaths                   -0.14480    0.02929  -4.944 8.57e-07 ***
HIV_AIDS                        -0.49542    0.02377 -20.844  < 2e-16 ***
Schooling                        0.43965    0.02590  16.972  < 2e-16 ***
Adult_Mortality                 -0.33308    0.01788 -18.630  < 2e-16 ***
Income_composition_of_resources  0.18444    0.01744  10.576  < 2e-16 ***
percentage_expenditure           0.20618    0.02768   7.448 1.65e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08605 on 1413 degrees of freedom
Multiple R-squared:  0.8013,    Adjusted R-squared:  0.8004
F-statistic: 949.7 on 6 and 1413 DF,  p-value: < 2.2e-16
```



Variable	Importance
infant_deaths	2.961
HIV_AIDS	0.502
Schooling	0.374
Adult_Mortality	0.285
Income_composition_of_resources	0.159
percentage_expenditure	0.136

Variable importance summary

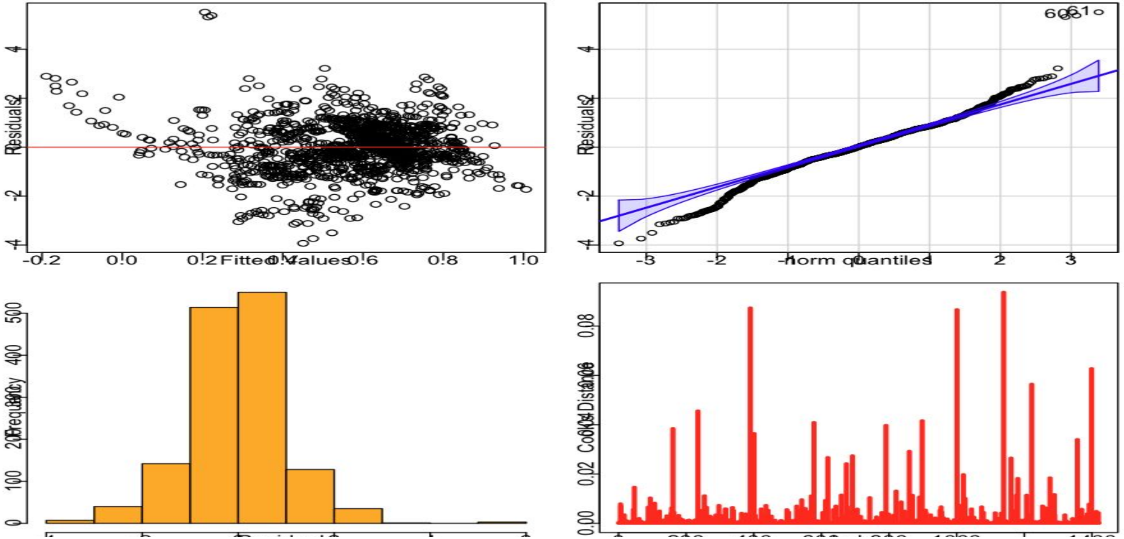
Model IV: Residual Analysis (Linearity)



Findings:

- Residuals are randomly distributed across predictor variables implying linear relation with the response variable.

Model IV: Residual Analysis (Variance and Normality)



Findings:

- The residual vs fitted values curve shows randomly distributed residuals with no pattern, implying uncorrelated errors and constant variance
- The QQ plot and histogram show that the residuals DO NOT have a normal distribution so the normality assumption is violated

Model Performance Comparison

Model Performance Comparison

		Adjusted R^2	MSPE	MAE	PM
	Model I	0.8504	0.00565	0.054641	0.14665
	Model II	0.8793	0.00596	0.055251	0.15468
Best	**Model III**	0.8795	0.006	0.055435	0.15570
	Model IV	0.8023	0.0074	0.060923	0.19433

- Model performance was calculated on test data comprising of **20%** of points from **random subsampling**

Conclusions

- Around 14 out of the 21 predictors seem to affect life expectancy significantly
- Life expectancy is significantly correlated with macroeconomic factors, health parameters as well as geography
- Government expenditure on healthcare as well as education can significantly improve life expectancy
- Viral outbreaks, like HIV-AIDS can negatively impact life expectancy
- Developed countries have higher life expectancy probably due to improved quality of life and availability of resources

Further Scope

- The data is limited in the sense that it does not capture all relevant health/macroeconomic factors; including more parameters and more data points can improve the model's prediction accuracy
- Data is for the years 2000-2016 and hence needs to be collected for more recent years for the study to be relevant
- MLR was used since the dataset was reasonably small, however advanced techniques like Random Forest or Gradient Boosting

Thank You!