Life Expectancy Predictive Model (WHO)

Statistical Analysis on factors influencing Life Expectancy

Patricia Attah

Introduction

Although there have been a lot of studies undertaken in the past on factors affecting life expectancy considering demographic variables, income composition and mortality rates. It was found that the effect of immunization and human development index was not taken into account in the past. Also, some of the past research was done considering multiple linear regression based on a data set of one year for all the countries. Hence, this gives motivation to resolve both the factors stated previously by formulating a regression model based on mixed effects model and multiple linear regression while considering data from a period of 2000 to 2015 for all the countries. Important immunization like Hepatitis B, Polio and Diphtheria will also be considered. In a nutshell, this study will focus on immunization factors, mortality factors, economic factors, social factors and other health related factors as well. Since the observations in this dataset are based on different countries, it will be easier for a country to determine the predicting factor which is contributing to lower value of life expectancy. This will help in suggesting a country which area should be given importance in order to efficiently improve the life expectancy of its population. *The objective of this study is to attempt to build a model that would identify and point out key relationships in the life expectancy of countries in the data as are related to the variables provided.*

Data Description

The project relies on accuracy of data. The Global Health Observatory (GHO) data repository under World Health Organization (WHO) which was made available to the public for the purpose of health data analysis. The data-set related to life expectancy, In this project data covers from the year 2000-2015 for 193 countries for further analysis. Missing data was handled in R software by using Missmap command. The result indicated that most of the missing data was for population, Hepatitis B and GDP. The missing data were from less known countries like Vanuatu, Tonga, Togo, Cabo Verde etc. Finding all data for these countries was difficult and hence, it was decided that we exclude these countries from the final model data-set. The final merged file(final dataset) consists of 22 Columns and 2938 rows which meant 20 predicting variables.

EDA: Exploratory Data Analysis

From 64,636 total observations, missing data 25% from columns adversely affecting countries that were difficult to collect data per WHO/UN (Figure 1). There are 3 categorical columns, country, over 65 years life expectancy, and status in which only 20% of countries are developed. The 20 continuous variables are representative of various immunization related factors, Mortality factors, Economical factors and Social factors (Figure 2). Observations of population, hepatitis b, GDP, total expenditure, alcohol, composition of resources and schooling make up 90% of missing data (Figure 3). 11 of the 20 continuous variables are right-skewed and were log-transformed, while only 3 vaccine related variables were left-skewed and cube^3-

transformed (Figure 4 & 5). For the question in hand, life expectancy was split into categorical age 65 above vs < 65 age. Outliers were left as part of the data and observations were omitted on a case-by-case basis depending on the model variables to increase degrees of freedom.

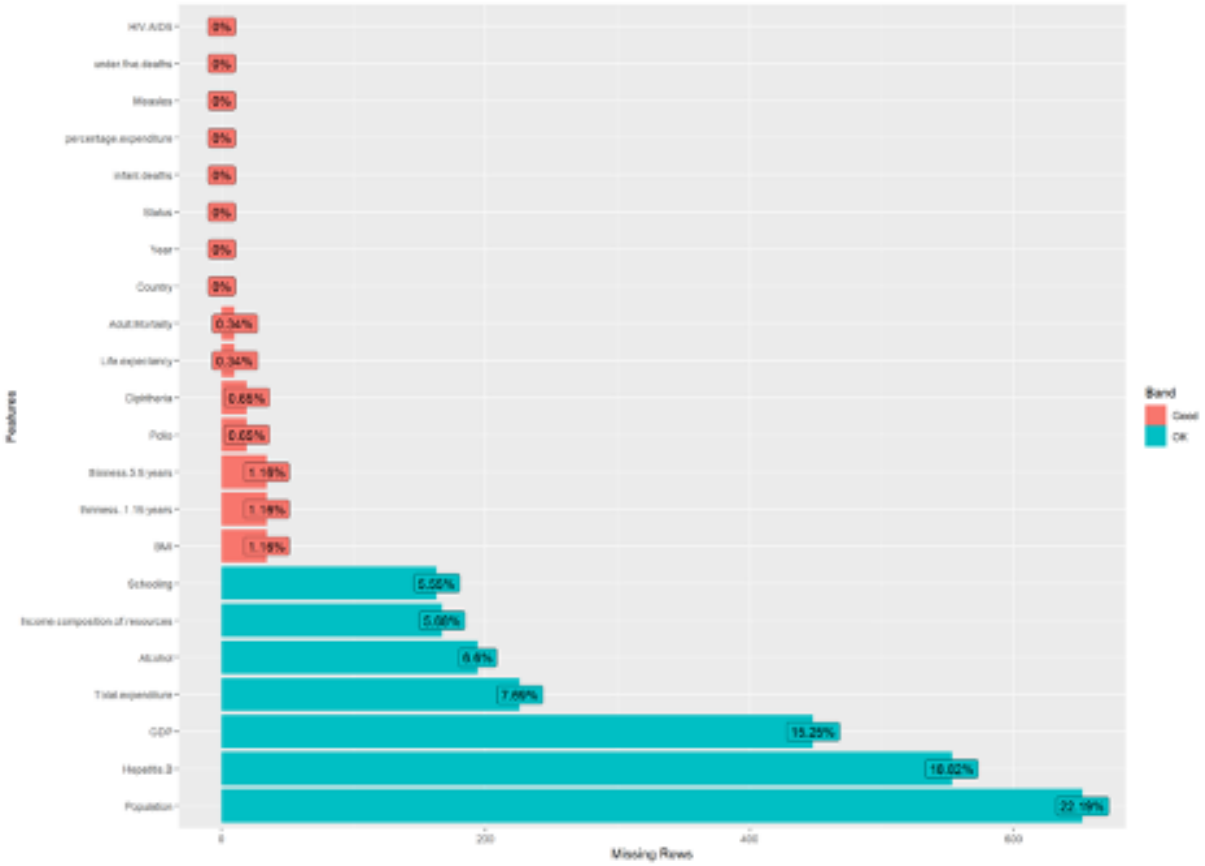
The transformed summary statistics and QQ Plots satisfy normality, equal variance of residuals, and independence. Analyzing the correlations with life expectancy we found log(HIV rate), composition of income, and schooling correlate above 0.7 (p-value <0.001). Outliers were left as part of the data as we know little about the data collection method and can rely on the CLT as the number of observations are > 2,000.

```
##    Country               Year         Status          Life.expectancy
##  Length:2938        Min.   :2000   Length:2938        Min.   :36.30
##  Class :character   1st Qu.:2004   Class :character   1st Qu.:63.10
##  Mode  :character   Median :2008   Mode  :character   Median :72.10
##                     Mean   :2008                      Mean   :69.22
##                     3rd Qu.:2012                      3rd Qu.:75.70
##                     Max.   :2015                      Max.   :89.00
##                                                       NA's   :10
##  Adult.Mortality infant.deaths       Alcohol        percentage.expenditure
##  Min.   :  1.0   Min.   :   0.0   Min.   : 0.0100   Min.   :    0.000
##  1st Qu.: 74.0   1st Qu.:   0.0   1st Qu.: 0.8775   1st Qu.:    4.685
##  Median :144.0   Median :   3.0   Median : 3.7550   Median :   64.913
##  Mean   :164.8   Mean   :  30.3   Mean   : 4.6029   Mean   :  738.251
##  3rd Qu.:228.0   3rd Qu.:  22.0   3rd Qu.: 7.7025   3rd Qu.:  441.534
##  Max.   :723.0   Max.   :1800.0   Max.   :17.8700   Max.   :19479.912
##  NA's   :10                       NA's   :194
##   Hepatitis.B       Measles               BMI        under.five.deaths
##  Min.   : 1.00   Min.   :     0.0   Min.   : 1.00   Min.   :   0.00
##  1st Qu.:77.00   1st Qu.:     0.0   1st Qu.:19.30   1st Qu.:   0.00
##  Median :92.00   Median :    17.0   Median :43.50   Median :   4.00
##  Mean   :80.94   Mean   :  2419.6   Mean   :38.32   Mean   :  42.04
##  3rd Qu.:97.00   3rd Qu.:   360.2   3rd Qu.:56.20   3rd Qu.:  28.00
##  Max.   :99.00   Max.   :212183.0   Max.   :87.30   Max.   :2500.00
##  NA's   :553                        NA's   :34
##      Polio       Total.expenditure   Diphtheria       HIV.AIDS
##  Min.   : 3.00   Min.   : 0.370    Min.   : 2.00   Min.   : 0.100
##  1st Qu.:78.00   1st Qu.: 4.260    1st Qu.:78.00   1st Qu.: 0.100
##  Median :93.00   Median : 5.755    Median :93.00   Median : 0.100
##  Mean   :82.55   Mean   : 5.938    Mean   :82.32   Mean   : 1.742
##  3rd Qu.:97.00   3rd Qu.: 7.492    3rd Qu.:97.00   3rd Qu.: 0.800
##  Max.   :99.00   Max.   :17.600    Max.   :99.00   Max.   :50.600
##  NA's   :19      NA's   :226       NA's   :19
##       GDP              Population        thinness..1.19.years
##  Min.   :     1.68   Min.   :3.400e+01   Min.   : 0.10
##  1st Qu.:   463.94   1st Qu.:1.958e+05   1st Qu.: 1.60
##  Median :  1766.95   Median :1.387e+06   Median : 3.30
##  Mean   :  7483.16   Mean   :1.275e+07   Mean   : 4.84
##  3rd Qu.:  5910.81   3rd Qu.:7.420e+06   3rd Qu.: 7.20
##  Max.   :119172.74   Max.   :1.294e+09   Max.   :27.70
##  NA's   :448         NA's   :652         NA's   :34
##  thinness.5.9.years Income.composition.of.resources   Schooling
##  Min.   : 0.10      Min.   :0.0000                  Min.   : 0.00
##  1st Qu.: 1.50      1st Qu.:0.4930                  1st Qu.:10.10
##  Median : 3.30      Median :0.6770                  Median :12.30
##  Mean   : 4.87      Mean   :0.6276                  Mean   :11.99
##  3rd Qu.: 7.20      3rd Qu.:0.7790                  3rd Qu.:14.30
##  Max.   :28.60      Max.   :0.9480                  Max.   :20.70
##  NA's   :34         NA's   :167                     NA's   :163
```

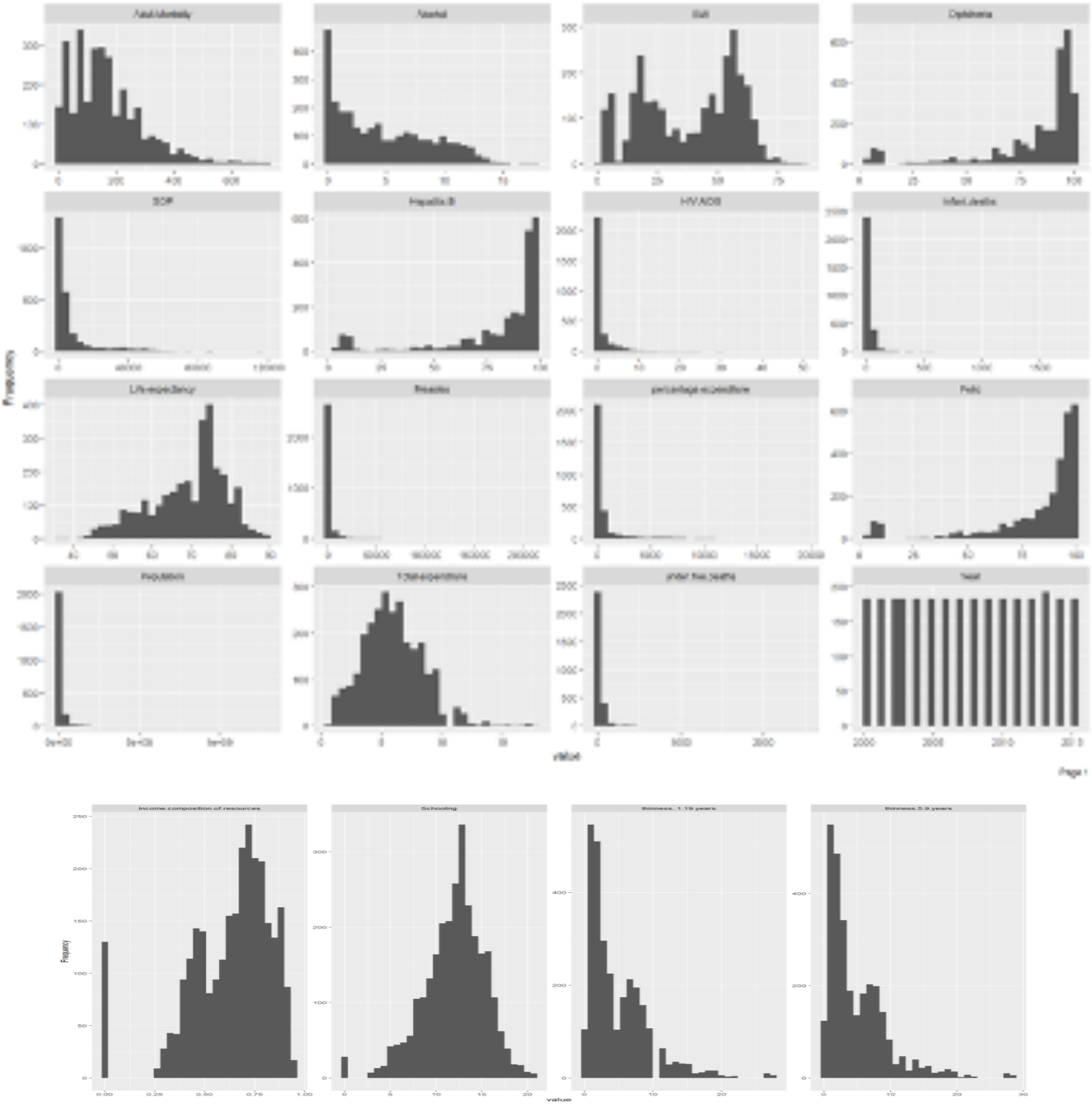
Data Metrics Analysis (Figure 1)

root (Classes 'data.table' and 'data.frame': 2938 obs. of 22 variables:)
- Country (chr)
- Year (int)
- Status (chr)
- Life.expectancy (num)
- Adult.Mortality (int)
- infant.deaths (int)
- Alcohol (num)
- percentage.expenditure (num)
- Hepatitis.B (int)
- Measles (int)
- BMI (num)
- under.five.deaths (int)
- Polio (int)
- Total.expenditure (num)
- Diphtheria (int)
- HIV.AIDS (num)
- GDP (num)
- Population (num)
- thinness..1.19.years (num)
- thinness.5.9.years (num)
- Income.composition.of.resources (num)
- Schooling (num)

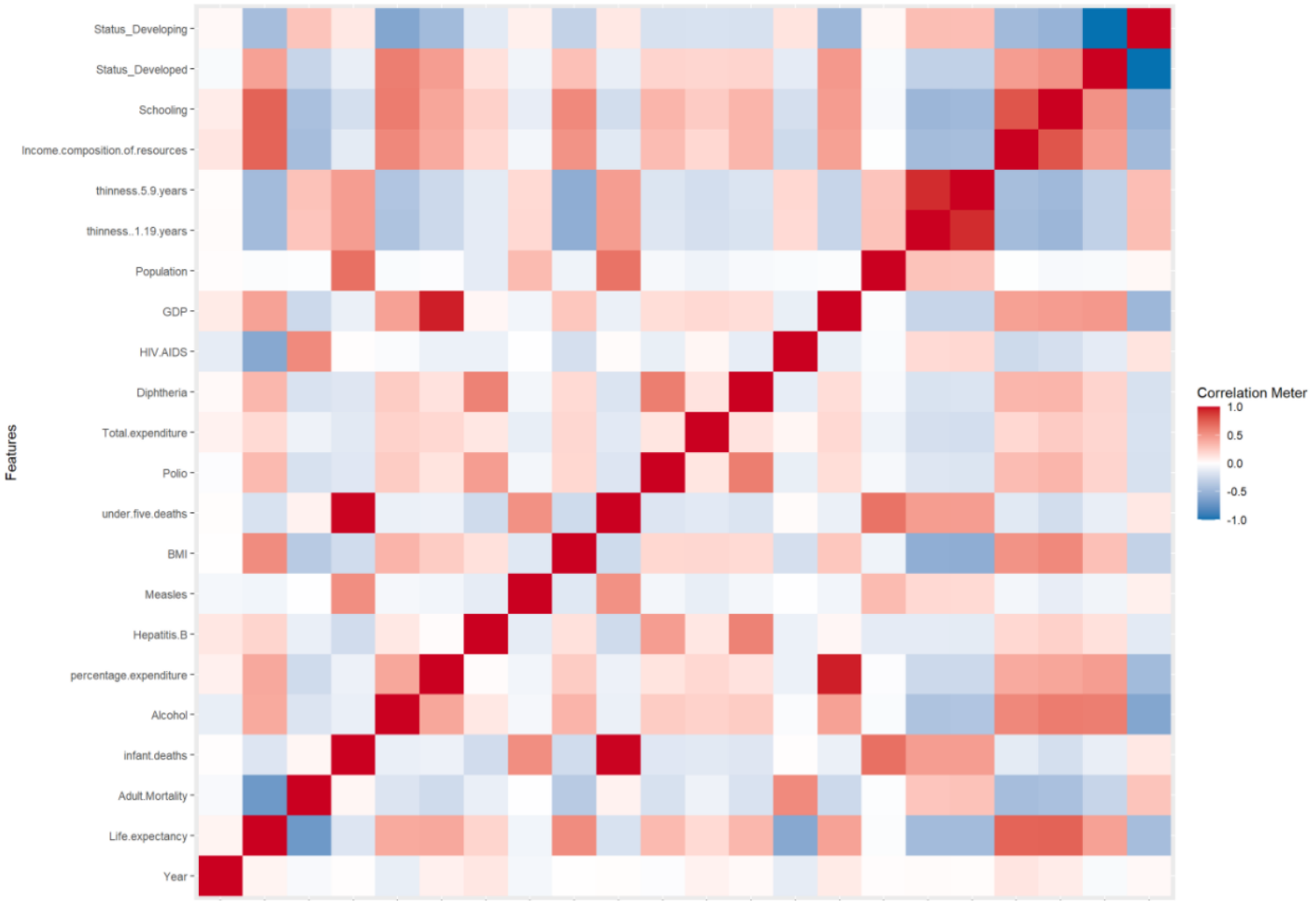
Variable Classification (Figure 2)



Missing Data (Figure 3)



Summary Statistics of Data (Figure 4)



Correlation Analysis (Figure 5)

Objective 1 Prediction Model Vs Interpretation Model

Problem Statement and Approach to Solve

Objective of this study is to attempt to build a model that would identify and point out key relationships and specific factors that affect life expectancy of countries provided in this data set, as are related to the variables provided.

Method to Solve

This study intends to identify key relationships that influence life expectancy produce two models one of which would be readily interpretable, and the other would be for the sole purpose of prediction. Our approach would be

I. Remove null values

II. Visualize the distribution

III. Transform data as required (only log transform for interpretation model)

IV. Check for correlation between response variables

V. Pick variables with highest correlation to the response variable

VI. Check for high correlation between selected predictor variables

VII. Check VIF of selected predictor variables

VIII. Create model with test and train data set

IX. Look at fitted model residuals to ensure assumptions are met

X. Use forward selection and LASSO variable selection method

XI. Check ASE, BIC, AIC and accuracy of final models

XII. Compare Interpretation model to Prediction model

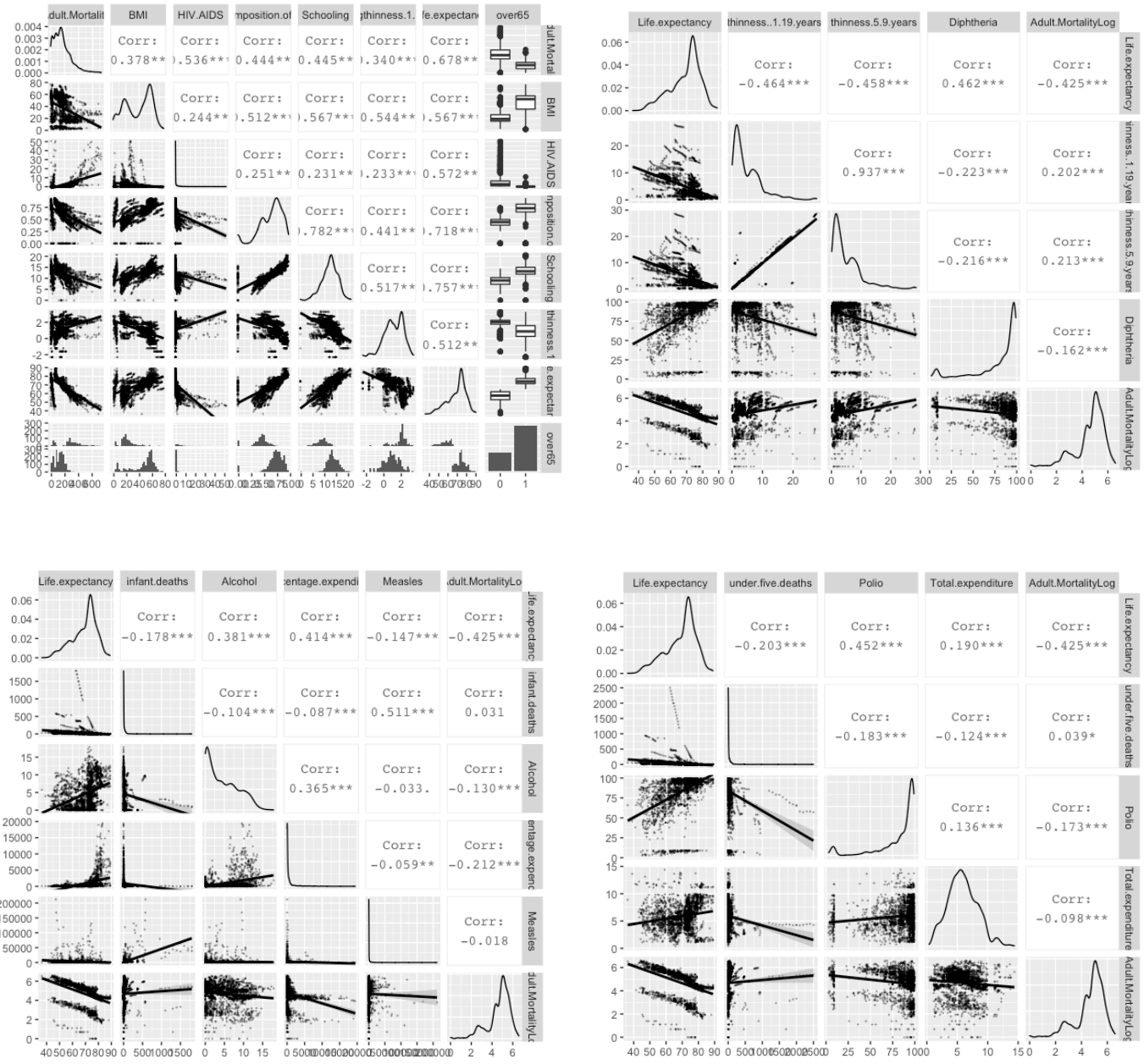
Correlation of full data set:Initial Visualization with Scatter plots

	Country	Year	Status	Life.expectancy	Adult.Mortality	infant.deaths	Alcohol
Country	1.00	0.00	-0.02	-0.02	0.04	-0.03	-0.08
Year	0.00	1.00	0.00	0.16	-0.08	-0.04	-0.06
Status	-0.02	0.00	1.00	-0.47	0.30	0.11	-0.59
Life.expectancy	-0.02	0.16	-0.47	1.00	-0.68	-0.18	0.38
Adult.Mortality	0.04	-0.08	0.30	-0.68	1.00	0.06	-0.17
infant.deaths	-0.03	-0.04	0.11	-0.18	0.06	1.00	-0.10

	percentage.expenditure	Measles	BMI	under.five.deaths	Polio	Total.expenditure
Country	-0.02	0.00	0.02	-0.03	0.04	0.01
Year	0.08	-0.10	0.09	-0.04	0.09	0.08
Status	-0.50	0.07	-0.30	0.11	-0.22	-0.26
Life.expectancy	0.41	-0.15	0.57	-0.20	0.45	0.19
Adult.Mortality	-0.25	0.01	-0.38	0.07	-0.25	-0.09
infant.deaths	-0.09	0.51	-0.22	1.00	-0.17	-0.12

	Diphtheria	HIV.AIDS	thinness..1.19.years	thinness.5.9.years
Country	0.01	0.10	0.02	0.04
Year	0.15	-0.13	-0.04	-0.04
Status	-0.20	0.15	0.36	0.36
Life.expectancy	0.46	-0.57	-0.46	-0.46
Adult.Mortality	-0.25	0.54	0.29	0.30
infant.deaths	-0.17	0.02	0.46	0.47

Scatter plot Visualization



Scatter plot Visualization (Figure 6)

Interpretation Model

Transformations (Interpretation Model):

Log transformed variables for interpretation model are Adult mortality, Polio, Diphtheria, Thinness1-19 and HIV.AIDS

Initial variable reduction (Interpretation Model):

Variables most correlated with life expectancy will be the first set of variables added to the model. As shown below. After these are added a few other variables with lower correlations are added after looking at the visual EDA in sections above. The final model VIF is checked to ensure there is no correlation within the variables in the model. The VIFs are all below 5 so we can use this model as is.

	Life.expectancy
Life.expectancy	1.0000000
Adult.Mortality	-0.6779106
BMI	0.5669106
HIV.AIDS	-0.5717070
Income.composition.of.resources	0.7179532
Schooling	0.7574450
logthinness.1.19	-0.5122769

Most correlated with Life Expectancy (Figure 7)

```
> vif(model)
```

Adult.MortalityLog	PolioLog	DiphtheriaLog
1.203275	1.360561	1.382272
BMI	HIV.AIDSlog	Income.composition.of.resources
1.789369	1.667976	2.740356
Schooling	percentage.expenditure	logthinness.1.19
3.183416	1.297681	1.652174

VIF of Interpretation model (Figure 8)

Assumptions Check for Interpretation Model:

I. Residual Plots

II. Influential point analysis (Cook's D and Leverage)

The residual plot for the model seems fine there seems to be no extreme outlier and the model seems to be normal.

Results for Interpretation Model

AIC: 11443.39

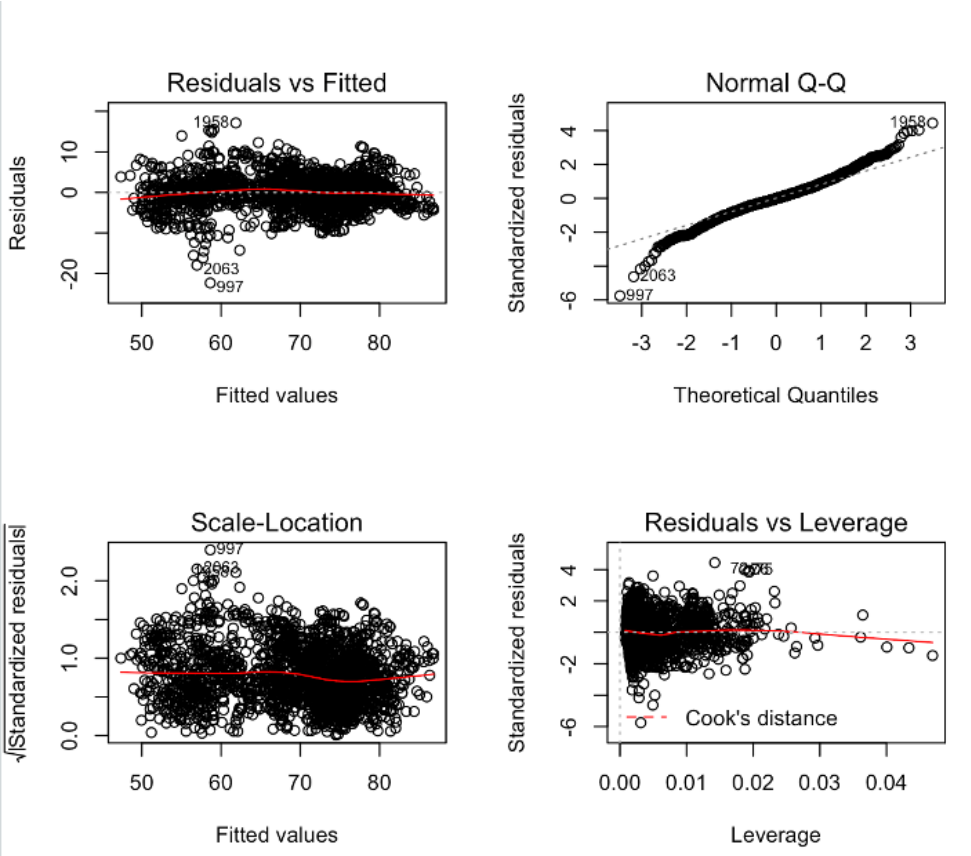
BIC: 11505.23

ASE: 4309.131

Accuracy: 0.92

```
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     50.168518   0.959404  52.291  < 2e-16 ***
Adult.MortalityLog              -0.668576   0.088669  -7.540 7.03e-14 ***
PolioLog                         0.496894   0.168267   2.953 0.003183 **
DiphtheriaLog                    0.349028   0.169192   2.063 0.039249 *
BMI                              0.008798   0.005957   1.477 0.139828
HIV.AIDSlog                     -2.837572   0.070834 -40.059  < 2e-16 ***
Income.composition.of.resources  8.095105   0.666760  12.141  < 2e-16 ***
Schooling                        0.811875   0.049565  16.380  < 2e-16 ***
Total.expenditure                0.060918   0.039743   1.533 0.125480
logthinness.1.19                -0.364800   0.107659  -3.388 0.000716 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.965 on 2034 degrees of freedom
Multiple R-squared:  0.8202,    Adjusted R-squared:  0.8194
F-statistic:  1031 on 9 and 2034 DF,  p-value: < 2.2e-16
```

Summary Interpretation model (Figure 9)



Residuals Interpretation model (Figure 10)

Prediction Model

Transformations Prediction Model:

For the prediction model the same variables used in the interpretation model were log transformed (are Adult mortality, Polio, Diphtheria, Thinness1-19) in addition also these variables were log transformed HIV AIDS, Alcohol and Diphtheria was squared.

Initial variable reduction Prediction Model:

There was no initial variable reduction method after the variables were transformed and added in the dataset. The transformed variables old columns were removed and also variables year, country and status were not used in the model. After the initial model is created, we check VIF and remove variables with high VIFs. After this we use stepwise forward selection to select the best variables with R.

```
> vif(model_pre)
                Adult.Mortality                  infant.deaths                        Alcohol
                       7.074409                     211.536445                       4.667235
         percentage.expenditure                    Hepatitis.B                        Measles
                      11.171866                       1.834237                       1.500785
                            BMI                  under.five.deaths                          Polio
                       1.919613                     202.936314                      50.783752
           Total.expenditure                      Diphtheria                       HIV.AIDS
                       1.157878                      40.601358                       2.935623
                            GDP                     Population          thinness..1.19.years
                      11.883130                       1.907795                      11.945005
          thinness.5.9.years Income.composition.of.resources                       Schooling
                       8.217386                       3.214968                       4.080329
          Adult.MortalityLog                     AlcoholLog                         Over65
                       4.351594                       2.916126                       3.354950
                  Developing                     Diphtheria2                        PolioLog
                       2.013532                      59.769559                      36.664304
       thinness..1.19.yearsLog                    HIV.AIDSLog
                       5.042322                       4.817976
```

VIF Predictions Model (Figure 11)

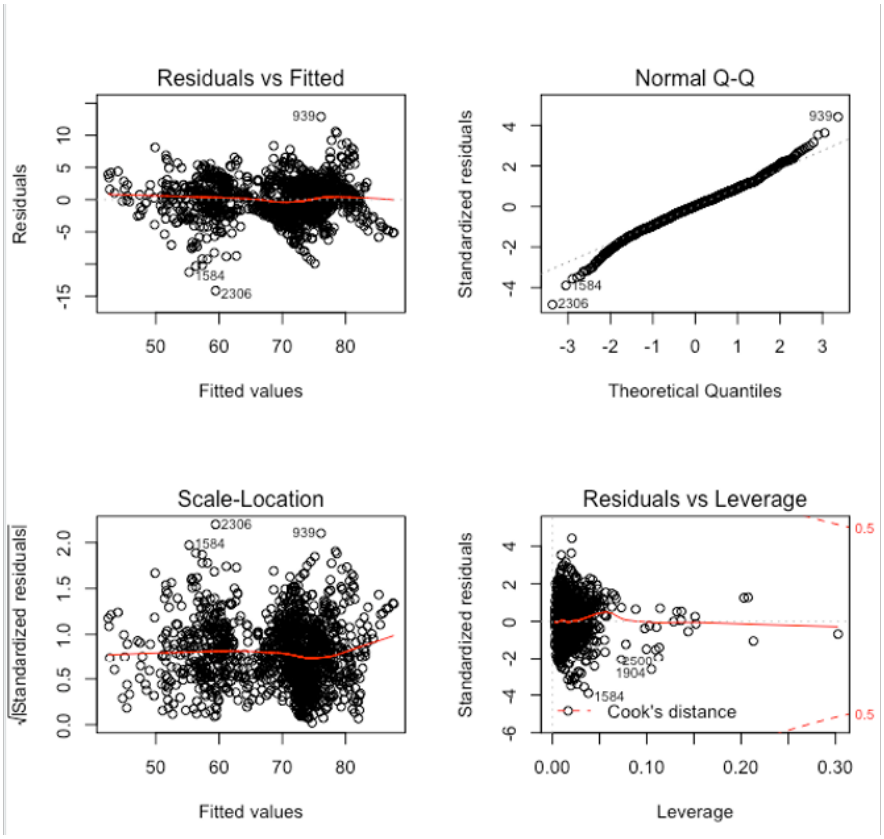
Assumptions Check for Prediction Model:

I. Residual Plots

II. Influential point analysis (Cook's D and Leverage)

```
Residuals:
     Min       1Q   Median       3Q      Max
-14.1891  -1.7770   0.0678   1.8452  12.7954

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     5.835e+01  9.774e-01  59.704  < 2e-16 ***
Over65Yes                      -4.867e+00  3.310e-01 -14.701  < 2e-16 ***
HIV.AIDSLog                    -9.797e-01  1.051e-01  -9.319  < 2e-16 ***
Schooling                       4.807e-01  5.715e-02   8.411  < 2e-16 ***
Adult.Mortality                -2.037e-02  1.713e-03 -11.892  < 2e-16 ***
percentage.expenditure          5.430e-04  5.752e-05   9.440  < 2e-16 ***
Adult.MortalityLog              1.161e+00  1.568e-01   7.405 2.36e-13 ***
Income.composition.of.resources 6.414e+00  7.807e-01   8.216 5.03e-16 ***
HIV.AIDS                       -1.759e-01  2.257e-02  -7.792 1.34e-14 ***
Total.expenditure               1.690e-01  3.736e-02   4.523 6.64e-06 ***
thinness..1.19.yearsLog        -2.603e-01  1.067e-01  -2.440 0.014824 *
AlcoholLog                      3.525e-01  6.258e-02   5.633 2.17e-08 ***
DevelopingYes                  -1.714e+00  3.260e-01  -5.258 1.70e-07 ***
Alcohol                        -2.078e-01  4.324e-02  -4.807 1.71e-06 ***
Diphtheria2                     1.646e-04  4.526e-05   3.637 0.000287 ***
Hepatitis.B                    -1.326e-02  4.254e-03  -3.116 0.001875 **
BMI                             1.384e-02  5.618e-03   2.464 0.013865 *
infant.deaths                  -1.637e-03  7.344e-04  -2.230 0.025941 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.947 on 1301 degrees of freedom
Multiple R-squared:  0.8892,    Adjusted R-squared:  0.8878
F-statistic: 614.2 on 17 and 1301 DF,  p-value: < 2.2e-16
```

Summary Prediction Model (Figure 12)



Residuals Predictions Model (Figure 13)

The residual plot for the model seems fine there seems to be no extreme outlier and the model seems to be normal.

Results for Prediction Model

AIC: 6614.178

BIC: 6712.686

ASE: 4345.182

Accuracy: 0.94

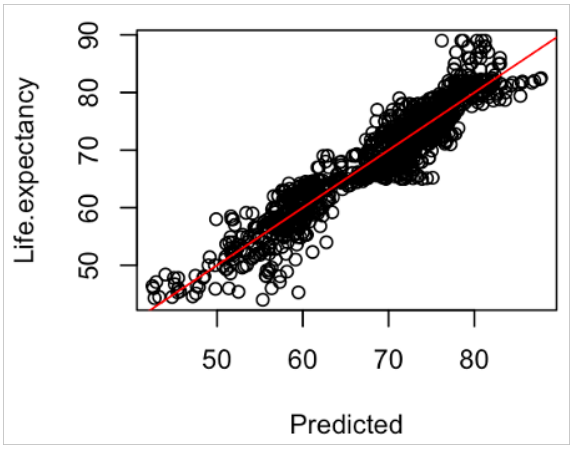
Model Comparison Prediction Model Vs Interpretation Model

Results	
Prediction Model	**Interpretation Model**
AIC: 6614.178	AIC: 11723.26
BIC: 6712.686	BIC: 11785.11
ASE: 4345.182	ASE: 4304.671
Accuracy: 0.94	Accuracy: 0.89

Confidence Intervals (First 6 values)

Confidence Intervals	
Prediction Model	**Interpretation Model**

Fit	Lower Int	Upr Int	Fit	Lower Int	Upr Int
74.37818	73.40472	75.35163	77.10461	76.7308	77.47843
81.77939	81.23486	82.32391	73.12583	72.80243	73.44922
72.96912	72.57272	73.36551	56.30392	55.91417	56.69368
73.19794	72.83856	73.55732	77.45352	76.93968	77.96736
73.08371	72.6444	73.52301	69.65364	68.94969	70.35759
72.39095	72.02512	72.75678	79.52539	79.04395	80.00683



Life expectancy Vs Predicted Values (Figure 14)

Model Comparison Summary

The ASE tells is the reproducibility of our model on a future data set, we see that both the models are pretty close however our interpretable model (ASE: 4304.671) gives us the most reproducible with 40 points lower than the prediction model (ASE: 4345.182) for the ASE. However, In every other metric (AIC, BIC, and accuracy) prediction model is better than interpretation model.

In terms of predictions the prediction model is better since the accuracy is considerably higher it has a value of 94% while the interpretation model has an accuracy of 84%. In conclusion we can say that the prediction model is mor4e accurate and hence better than the interpretation model.

Interpretation of coefficients

Variable	Estimate	Log Transform
Adult.MortalityLog	-0.668576	YES
PolioLog	0.496894	YES
HIV.AIDSlog	-2.837572	YES
DiphtheriaLog	0.349028	YES
Income.composition.of.resources	8.095105	NO
Schooling	0.811875	NO
BMI	0.008798	NO
Total.expenditure	0.060918	NO

The Intercept had an estimate of 50.16. means that when all other factors are at 0 the average life expectancy is approximately 50 years old.

All logs are computed to the base of 10 as default in R.

1) Holding all other factors constant it is estimated that a 10 fold increase in the probability of death between the ages 15 and 60 per 1000 individuals is associated with ($10^{-0.66875}$=0.214) decrease in mean life expectancy by 78% of 1 year.
2) Holding all other factors constant it is estimated that a 10 fold increase in polio immunizations per 1 year old is associated with ($10^{0.496894}$=3.13) the increase in life expectancy by 3.13 years
3) Holding all other factors constant it is estimated that a 10 fold increase in deaths per lives births due to HIV/AIDS is associated with the ($10^{-2.837572}$=0.001) decrease in the mean life expectancy by 0.99 years.
4) Holding all other factors constant it is estimated that a 10 fold increase in Diphtheria tetanus toxoid and pertussis (DPT3) immunization coverage for 1 year-olds is associated with ($10^{0.349028}$=0.001)in crease in mean life expectancy by 2.2 years.

5) Holding all other factors constant it is estimated for each additional percent recorded in the human development aspect of income composition of resources it increases the average life expectancy by 8years.

6) Holding all other factors constant it is estimated for each additional year of schooling an individual has its resources it increases the average life expectancy by 0.8 years.

7) Holding all other factors constant it is estimated for each additional 1 point increase in BMI of the average body mass in a population 0.009 years.

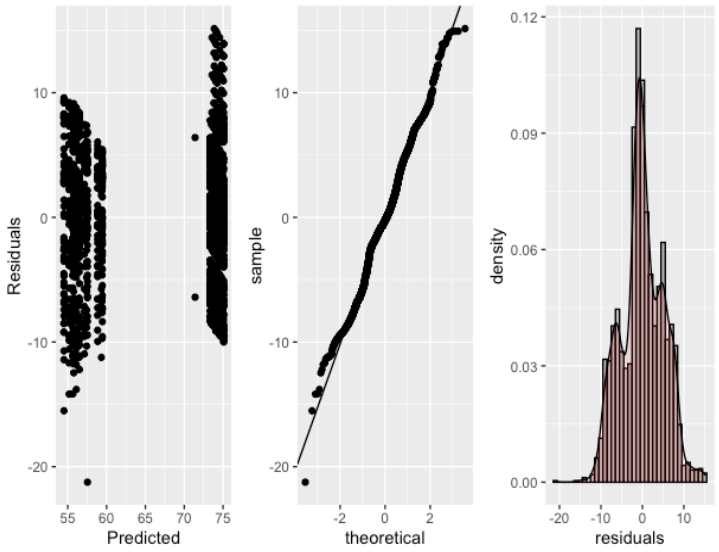
8) Holding all other factors constant it is estimated for each additional percent spent on the health from the government total expenditure budget increases life expectancy by 0.06 years.

Key Relationships: Total Expenditure Vs Life Expectancy

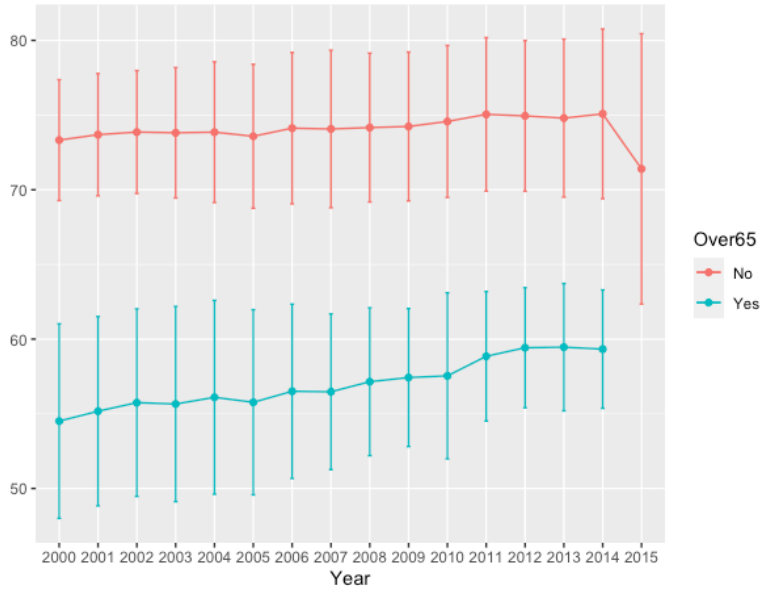
We decided to look at the effect of healthcare expenditure on life expectancy. Our question of interest is should a country having a lower life expectancy value(<65) increase its healthcare expenditure in order to improve its average lifespan? Analysis includes first using 2-way Anova to check if the independent variable Year has an interaction with life expectancy value of 65years old. Using multiple linear regression to check if the expenditure affects if life expectancy is above or below 65 years.

Our first goal is to see if three is an effect of time (2000 - 2015) by Year on the life expectancy, Particularly we are looking at if over the course of 15 years if the life expectancy of above 65 is more or less likely.

We can see from the means plot there seems to be little to no interaction between year and above or below 65 years of age. After examining the residual plot there do not seem to be any concerns about the assumptions, hence no need for transformations or such. Below looking at the summary of year Vs above or below 65 we see that the interaction term is not significant, hence there is no need to run further test or apply correction factors



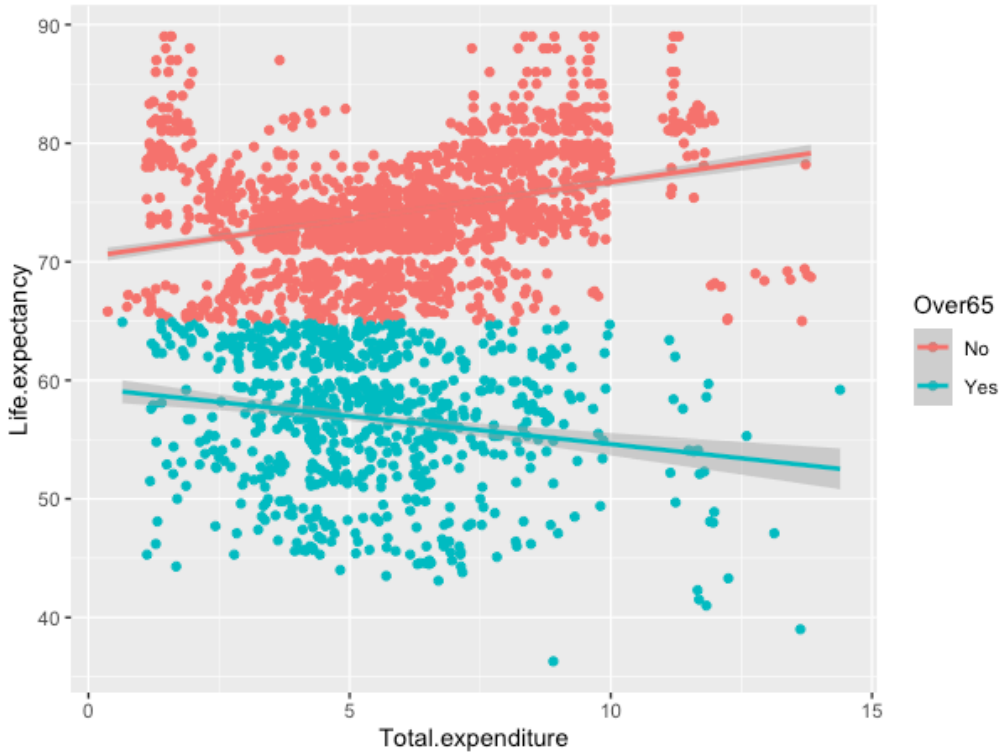
Residual Plots (Figure 15)



Means Plot Years Vs Over65 life Expectancy (Figure 16)

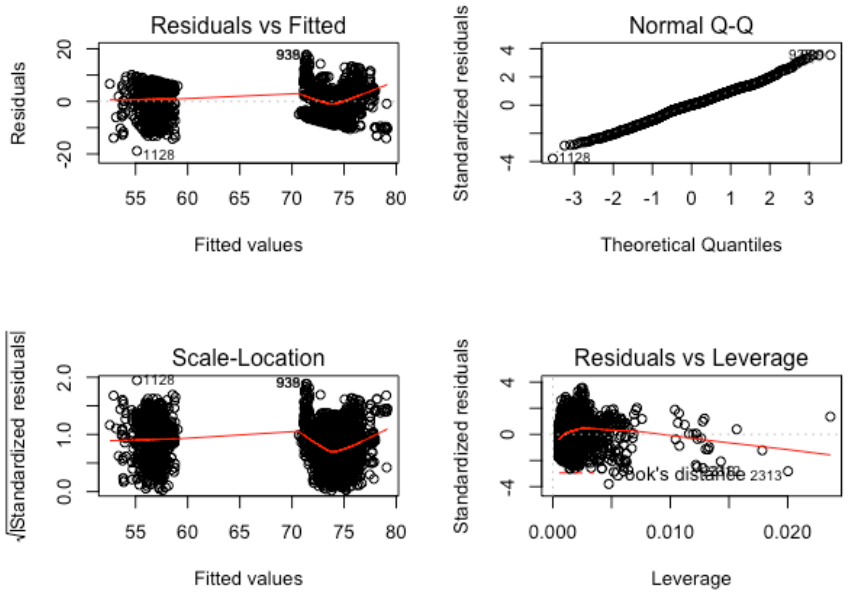
```
Anova Table (Type III tests)

Response: Life.expectancy
            Sum Sq   Df F values Pr(>F)
Over65       13214    1 505.5624 <2e-16 ***
Year           554   15   1.4121 0.1322
Over65:Year    551   14   1.5048 0.1009
Residuals    65996 2525
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Summary Year Vs Over65 Model (Figure 17)



Life Expectancy Vs Total expenditure (Figure 18)

From the above there seems to be some interaction between total expenditure and life expectancy as regards to below or above 65 years. We examine the Residual plot below and we see that the assumptions are met hence no need for interactions. Also below is the summary of the model for total expenditure vs life expectancy and we see there is significant relationship between these two variables and its interaction term. Conclusion: We can state that there is evidence to suggest that total healthcare expenditure will have an significant effect on life expectancy being above or below 65 years of age.



Life Expectancy Vs Total expenditure (Figure 19)

```
Call:
lm(formula = Life.expectancy ~ Over65 + Total.expenditure + Total.expenditure:Over65,
    data = Age65_cost)

Residuals:
     Min       1Q   Median       3Q      Max
-18.8334  -3.1847   0.1583   3.3216  17.6658

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                  70.41913    0.31547  223.22   <2e-16 ***
Over65Yes                   -11.07023    0.57673  -19.20   <2e-16 ***
Total.expenditure             0.63111    0.04847   13.02   <2e-16 ***
Over65Yes:Total.expenditure  -1.10475    0.09573  -11.54   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.977 on 2552 degrees of freedom
Multiple R-squared:  0.7198,    Adjusted R-squared:  0.7195
F-statistic:  2186 on 3 and 2552 DF,  p-value: < 2.2e-16
```

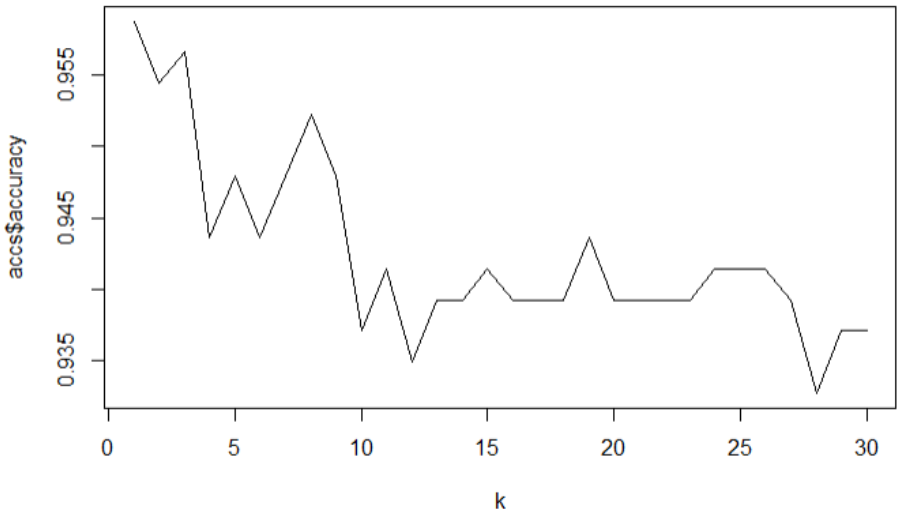
Life Expectancy Vs Total expenditure (Figure 20)

Objective 2: Non-Parametric Model

(K-Nearest Neighbor KNN)

KNN, K Nearest Neighbor is a Supervised Machine Learning algorithm that classifies a new data point into the target class, depending on the features of its neighboring data points. In theory, we’d like to

use the Bayesian Classifier, but with real-world data all assumptions aren't met easily. As we saw in regression, k-nearest neighbors have no such model parameters. Instead, it has a tuning parameter, k. This is a parameter which determines *how* the model is trained, instead of a parameter that is *learned* through training. Noting that tuning parameters are not used exclusively with non-parametric methods. There are four fundamental challenges when applying nonparametric regression: definition of an appropriate state space, definition of a distance metric to determine nearness of historical observation to the current conditions, selection of a forecast generation method given a collection of nearest neighbors, and management of the potential neighbors' database. Because the KNN classifier predicts the class of a given test observation by identifying the observations that are nearest to it, the scale of the variables matters. Any variables that are on a large scale will have a much larger effect on the distance between the observations, and hence on the KNN classifier, than variables that are on a small scale and suffer this way. Using KNN cannot predict if the model built for the particular research can be used with any other datasets, lacking reproducibility. The WHO may not be able to pin-point life expectancy exactly, but more general factors such as Over 65 years life expectancy or under may be more achievable to predict a countries placement given data and changes. The accuracy of the model produced was 96% a k = 1 groups.



Best Determined k-groups for KNN (Figure 21)

```
Confusion Matrix and Statistics

                 classifications
                  -0.638454030606267 1.56575036648094
 -0.638454030606267                315                13
 1.56575036648094                    6               127

               Accuracy : 0.9588
                 95% CI : (0.9364, 0.975)
    No Information Rate : 0.6963
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9012

 Mcnemar's Test P-Value : 0.1687

            Sensitivity : 0.9813
            Specificity : 0.9071
         Pos Pred Value : 0.9604
         Neg Pred Value : 0.9549
             Prevalence : 0.6963
         Detection Rate : 0.6833
   Detection Prevalence : 0.7115
      Balanced Accuracy : 0.9442

       'Positive' Class : -0.638454030606267
```

KNN Statistics (Figure 22)

Conclusion

Life expectancy, magic pill, or fountain of youth, no matter what we're all looking for ways to lengthen our lives. The point is because we don't have magic we must use the skills we have, data and statistics. In our analysis question of interest we found the answer may be different for those developed and underdeveloped or countries with greater or lesser than 65 years expected life expectancy. As expected immunizations and HIV rates have an overall effect on life expectancy, but not health expenditure or BMI as we would expect based on current belief. Instead, we recommend a country seeking to increase its life expectancy would focus on increasing its investment in Health Development Index, schooling, and its adult mortality rates. When comparing the models the more complex the better accuracy, AIC, and BIC, though regarding ASE our most basic model prevailed as our recommended model for practicality/interpretability. Otherwise use of an advanced complex method, such as KNN, Bayesian Categorical, and/or SVM can be utilized. We recommend further investigation of possible confounding/interactions across, especially amongst smaller countries with less collectable data, and focusing on a particular question and methodology at hand on a country by country basis over time.

References

Smith, Brian & Williams, Billy & Oswald, R.. (2002). Comparison of parametric and nonparametric models for traffic flow forecasting. Transportation Research Part C: Emerging Technologies. 10. 303-321. 10.1016/S0968-090X(02)00009-8.

Appendix

Code

Code for Interpretation Model Vs Prediction Model

```
library(dplyr)
#IMPORT MODEL DROP SOME COLUMNS
Life.Expectancy <- read.csv("~/Library/Mobile Documents/com~apple~CloudDocs/R-SMU/ds7362 project/Life Expectancy Data 2.csv")

dim(Life.Expectancy)
sapply(Life.Expectancy, function(x) sum(is.na(x)))

life_exp_NA<-Life.Expectancy %>%
   mutate_all(~ifelse(. %in% c("N/A", "null", ""), NA, .)) %>%
   na.omit()

dim(life_exp_NA)

drop<- c('Hepatitis.B','Population','GDP')
life_exp = Life.Expectancy[,!(names(Life.Expectancy) %in% drop)]
```

```
library(dplyr)
#IMPORT MODEL DROP SOME COLUMNS
Life.Expectancy <- read.csv("~/Library/Mobile Documents/com~apple~CloudDocs/R-SMU/ds7362 project/Life Expectancy Data 2.csv")
```

```
dim(Life.Expectancy)
sapply(Life.Expectancy, function(x) sum(is.na(x)))

life_exp_NA<-Life.Expectancy %>%
   mutate_all(~ifelse(. %in% c("N/A", "null", ""), NA, .)) %>%
   na.omit()

dim(life_exp_NA)

drop<- c('Hepatitis.B','Population','GDP')
life_exp = Life.Expectancy[,!(names(Life.Expectancy) %in% drop)]
```

```{r}
#SUMMARY HAND SELECTED MODEL
summary(model)
```

```{r}
library(olsrr)
Life_Exp_Mod<- ols_step_forward_aic(model)
Life_Exp_Mod$model
```

```{r}
plot((model$fitted.values),train$Life.expectancy,xlab="Predicted Int",ylab="Life.expectancy")
lines(c(0,400000),c(0,400000),col="red")

head(predict(model,train,interval="confidence"))
```

```{r}
#PREDICTION MODEL CREATION
Life_Ex = Life.Expectancy  %>% mutate(Adult.MortalityLog = log(Adult.Mortality),
#                infant.deathsLog = log(infant.deaths),
                AlcoholLog = log(Alcohol),
                Over65 = as.factor(case_when(
                Life.expectancy < 65.0 ~ 'Yes',
```

```
            TRUE ~ 'No',)),
            Developing = as.factor(case_when(
            Status == 'Developing' ~ 'Yes',
            TRUE ~ 'No',)),
            Diphtheria2 = Diphtheria^(2),
#             Hepatitis.B3 = Hepatitis.B^(1/3),
            PolioLog = log(Polio),
            thinness..1.19.yearsLog = log(thinness..1.19.years),
#             thinness.5.9.yearsLog = log(thinness.5.9.years),
#             under.five.deathsLog = log(under.five.deaths),
#             percentage.expenditureLog = log(percentage.expenditure),
#             MeaslesLog = log(Measles),
            HIV.AIDSLog = log(HIV.AIDS),
#             GDPLog = log(GDP),
            Adult.MortalityLog = log(Adult.Mortality)
            )
Life_Ex = Life_Ex %>% na.omit()
```

```{r}
#REMOVE REDUNDANT VARIABLES
drop<- c('Country','Year','Status')
Life_Ex = Life_Ex[,!(names(Life_Ex) %in% drop)]

set.seed(100)  # setting seed to reproduce results of random sampling
trainingRowIndex <- sample(1:nrow(Life_Ex), 0.8*nrow(Life_Ex))  # row indices for training data
train0 <- Life_Ex[trainingRowIndex, ]  # model training data
test0  <- Life_Ex[-trainingRowIndex, ]
```

```{r}
#VIF MODEL
model_pre = lm(Life.expectancy ~ ., data = train0)
vif(model_pre)
```

```
```{r}
#REMOVING REDUNDANT VARIABLES CREATING TEST AND TRIAN SET
drop<- c('Diphtheria','under.five.deaths','Polio','Country','Year','Status','thinness..1.19.years')
Life_Ex0 = Life_Ex[,!(names(Life_Ex) %in% drop)]

set.seed(100)  # setting seed to reproduce results of random sampling
trainingRowIndex <- sample(1:nrow(Life_Ex0), 0.8*nrow(Life_Ex0))  # row indices for training data
train0 <- Life_Ex0[trainingRowIndex, ]  # model training data
test0  <- Life_Ex0[-trainingRowIndex, ]
```

```{r}
#PRDICTION MODEL
model_pre = lm(Life.expectancy ~ ., data = train0)
vif(model_pre)
```

```{r}
#RESIDUALS FOR PREDICTION MODEL
par(mfrow=c(2,2))
plot(model_pre)
```

``{r}
#AIC BIC MODEL
life.expectancy.pred0 <- predict(model_pre, test0)
Life_Exp_Mod0<- ols_step_forward_aic(model_pre)
full.model<- Life_Exp_Mod0$model
AIC(full.model)
BIC(full.model)
ASE<- mean((log(test0$Life.expectancy)-life.expectancy.pred0)^2)
ASE
```

```{r}
seed(124)
```

#Accuracy

actuals_preds0 = data.frame(cbind(actuals=test0$Life.expectancy, predicteds=life.expectancy.pred0))

correlation_accuracy0 = cor(actuals_preds0)

correlation_accuracy0

```

```{r}

#final model summary

summary(Life_Exp_Mod0$model)


plot((full.model$fitted.values),train0$Life.expectancy,xlab="Predicted",ylab="Life.expectancy")

lines(c(0,400000),c(0,400000),col="red")


```

```{r}

prd_c = predict(full.model,train0,interval="confidence", type = c("response"), level = 0.95)


```

```{r setup, include=FALSE}

knitr::opts_chunk$set(echo = TRUE)

```

Load the Packages

```{r load-packages, include=FALSE}

install.packages("pacman")

library(pacman)

pacman::p_load(naivebayes,
data.table,DT,forecast,ggplot2,plotly,scales,shiny,stringr,dplyr,rsconnect,caret,car,GGally,tidyr,ggpubr,readr,knitr,RCurl,skimr,DataExplorer,usmap,maps,statebins)

pacman::p_load(e1071,tidyverse,leaps,neuralnet,corrplot,nnet,randomForest,xgboost,Sleuth3,eemeans,lsmeans,gridExtra,anytime)

```

Load Data, Data Manipulation, & EDA Exploration

```{r}

Life <- read.csv("C:/Users/Dawson/Desktop/Data Science Program SMU/2020 03 FALL/6372 Applied Statistics Inference & Modeling/Zip Folder/ProjectDetails_2_2_2_2_2/LifeExpectancyProject/Kaggle/Life Expectancy Data.csv")

```
skim(Life)

str(Life)

create_report(Life)

LifeMod = Life %>% mutate(
            Over65 = as.factor(case_when(
            Life.expectancy < 65.0 ~ 'Yes',
            TRUE ~ 'No',)),
            Developing = as.factor(case_when(
            Status == 'Developing' ~ 'Yes',
            TRUE ~ 'No',)),
            Diphtheria3 = 1/log(Diphtheria),
            Hepatitis.B3 = 1/log(Hepatitis.B),
            Polio3 = 1/log(Polio),
            thinness..1.19.yearsLog = log(thinness..1.19.years),
            thinness.5.9.yearsLog = log(thinness.5.9.years),
            Adult.MortalityLog = log(Adult.Mortality),
#   infant.deathsLog = log(infant.deaths),
            AlcoholLog = log(Alcohol),
#   under.five.deathsLog = log(under.five.deaths),
#   percentage.expenditureLog = log(percentage.expenditure),
#   MeaslesLog = log(Measles),
            HIV.AIDSLog = log(HIV.AIDS),
            GDPLog = log(GDP),
            Adult.MortalityLog = log(Adult.Mortality)
)
LifeOmit = LifeMod %>% na.omit()
skim(LifeOmit)
create_report(Life)
create_report(LifeMod, y = "Over65")
create_report(LifeOmit, y = "Over65")
skim(LifeMod)
skim(LifeOmit)
```

Plotting Correlations

```{r}
plot_correlation(na.omit(LifeMod), maxcat = 5L)
p_load(corrplot)
LifeNum = select_if(LifeMod, is.numeric) %>% na.omit()
rquery.cormat(LifeNum)
res <- cor(LifeNum)
round(res, 2)
symnum(res, abbr.colnames = FALSE)
str(LifeNum)
p_load("PerformanceAnalytics")
chart.Correlation(LifeNum[,c(2,1:5)], histogram=TRUE, pch=19)
chart.Correlation(LifeNum[,c(2,6:10)], histogram=TRUE, pch=19)
chart.Correlation(LifeNum[,c(2,11:15)], histogram=TRUE, pch=19)
chart.Correlation(LifeNum[,c(2,16:20)], histogram=TRUE, pch=19)
chart.Correlation(LifeNum[,c(2,21:25)], histogram=TRUE, pch=19)
chart.Correlation(LifeNum[,c(2,26:29)], histogram=TRUE, pch=19)
```

KNN MODELING For Over65

```{r}
pacman::p_load(class, caret, e1071)
ModelData2 = LifeMod  %>% mutate(Over65 = as.numeric(as.factor(Over65)), Developing = as.numeric(as.factor(Developing))) %>% dplyr::select(Over65, percentage.expenditure, Developing, GDPLog, HIV.AIDS, under.five.deaths, infant.deaths, thinness.5.9.years, Adult.Mortality, Year, Schooling, Polio3, BMI, AlcoholLog, Income.composition.of.resources) %>% mutate_if(is.numeric, scale) %>% na.omit()
set.seed(100)  # setting seed to reproduce results of random sampling
trainingRowIndex <- sample(1:nrow(ModelData2), 0.8*nrow(ModelData2))  # row indices for training data
train <- ModelData2[trainingRowIndex, ]  # model training data
test  <- ModelData2[-trainingRowIndex, ]   # test data
## Loop for many k and one training / test partition
accs = data.frame(accuracy = numeric(30), k = numeric(30))
for(i in 1:30)
{
  classifications = knn(train[,-1],test[,-1],train$Over65, prob = TRUE, k = i)
  table(test$Over65,classifications)
```

```
  CM = confusionMatrix(table(test$Over65,classifications))
  accs$accuracy[i] = CM$overall[1]
  accs$k[i] = i
}

plot(accs$k,accs$accuracy, type = "l", xlab = "k")
# k = 1
classifications = knn(train[,-1],test[,-1],train$Over65, prob = TRUE, k = 1)
table(test$Over65,classifications)
CM = confusionMatrix(table(test$Over65,classifications))
CM$overall[1]
CM
```