RUTGERS
THE STATE UNIVERSITY
OF NEW JERSEY

PROJECT REPORT

Data Analysis on

Life Expectancy

Prepared by

Vishnuram Jatin Bangaru(vb440)

Rishik Salver(rs2116)

Ankeeta Priyam(ap2213)

Prashanth Aripirala(pa500)

Table of Contents

ABSTRACT

Life expectancy data analysis is the process of using statistical methods and other data analysis techniques to predict how long a person is likely to live. This is a complex task that involves considering a wide range of factors that may affect a person's lifespan, including medical history, lifestyle choices, and environmental factors. By analysing data on these and other factors, researchers can create models that can help them make more accurate predictions about a person's life expectancy.

Our project implementation will include three parts:

1. Data pre-processing
2. Data Analysis
3. Building a multiple linear regression to predict life expectancy.

Our main aim of Data analysis would be to find useful insights like finding the factors influencing life expectancy, how each factor contributes to the life expectancy. This part would be extended to each country and we will try to answer a few questions like which factor is influencing the life expectancy in each country, which countries need to improve their health sector to increase the life expectancy etc. We will try to show the results by plotting insightful graphs. We hope to explore factors that affect human life expectancy beyond our current knowledge of it after this analysis. We will be using Hypothesis testing to verify if the data analysis we performed is accurate. Upon feature engineering, we will then build a multiple linear regression model using the filtered set of factors obtained from the above step. We will use a few of the model selection like forward selection & backward elimination, transformations to obtain a better fitted model.

INTRODUCTION

Health analysis is an important domain to focus on for the betterment of any country. The Global Health Observatory (GHO) data repository under World Health Organization (WHO) keeps track of the health status as well as many other related factors for all countries. It has been observed that in the past 15 years, there has been a huge development in health sector resulting in improvement of human mortality rates especially in the developing nations in comparison to the past 30 years. The goal of the project is to do a data analysis on life expectancy and predict life expectancy. Life expectancy can be taken as one of the key factors from which we can analyse the various health factors around the world. The analysis would give us insights on immunizations, diseases, economic factors around different part of our globe. The aim of the analysis will help us impact longevity and improve health, extend life expectancy for people around the world. Predicting life expectancy would be a difficult task because of different factors but the recent trends of improved health factors would make our task less complicated. These predictions would indeed be a wake-up call for countries to improve their health sectors.

PROBLEM FORMULATION

An exploratory data analysis will help us to analyse factors to patterns and trends in data. Each of the four categories:

1. **Immunization factors** like hepatitis, diphtheria, polio can indeed help life expectancy to increase as people as such can be protected from such diseases.

2. **Diseases** like measles, HIV/aids can cause life expectancy to decrease. We can analyse countries which are more prone to such diseases.

3. **Economic factors** - The more they invest on health sector the better the life expectancy. Economic factors help us understand how countries have contributed to their health sector.

4. **Social factors** like amount of alcohol, schooling can by intuition impact life expectancy in a positive and negative way.

All of this can be considered as a regression analysis. We will develop a Multiple Linear Regression to predict life expectancy. Removing of correlated factors will be done as part of analysis and VIFs. Checking for linearity assumptions and improving it will help us achieve the best fit model. Model selections will help us get the filtered set of factors and then transformations will help us improve linearity assumptions. We will remove outliers using Jack knife Residuals. The entire analysis will be helpful for us to understand what factors are crucial and steps can be taken by countries to improve their health sectors.

DATA SET SOURCE

The dataset consists of health factors for 193 countries from the year 2000-2015. The factors have been divided into 4 categories which include immunization factors like Hepatitis B, mortality factors, economic factors, social factors, and other minor health factors. The countries are categorized into developing and developed countries. The data does not contain individual lifestyle factors like diet, exercises, sleep etc.

Variable	Description
Country	Names of Country
Year	Year of the data; Range from 2000 to 2015 for each country
Status	Status of the Country categorized into Developed/Developing
Life Expectancy	Life Expectancy in years
Adult Mortality	No. of deaths of people of both sexes between the age 15 and 60 per 1000.
Infant Deaths	No. of Infant deaths per 1000 population
Alcohol	Per capita consumption of alcohol(15+) in litres of pure alcohol
Percentage Expenditure	Percentage of GDP spent on health.
Hepatitis B	Percentage of 1-year olds who are Hepatitis B immune
Measles	Number of reported cases per 1000
BMI	Average Body Mass Index of the entire population
under-five deaths	Number of under-five deaths per 1000
Polio	Percentage of 1-year olds who are Polio immune
Total Expenditure	Percentage of amount spent on health to the total expenditure
Diphtheria	Percentage of 1-year olds who are DTP3 immune
HIV/AIDS	Number of Deaths per 1000
GDP	Gross Domestic Product per capita in USD
Population	Population of the Country
Thinness 10-19 years	Thinness percentage of people between the ages 10 and 19
Thinness 5-9 years	Thinness percentage of people between the ages 5 and 9
Income Composition	Human development index in terms of income composition of resourses ranging from 0 to 1.
Schooling	Number of years of schooling

METHODOLOGY

1. Exploratory Data Analysis

1.1 Data Cleaning and Pre-processing

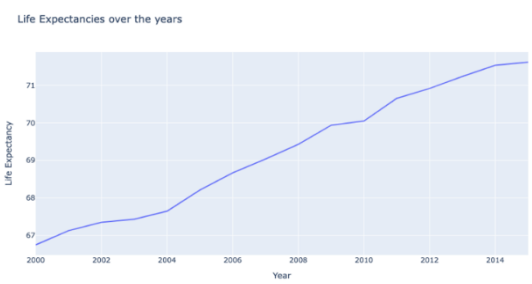
- The data given by Kaggle was already been cleaned using R Missmap but there still are missing values in few of the columns. We finally used SimpleImputer() class to fill the missing values which fills the data by using either mean median or frequent values.
- The data consists of inconsistent values where few countries have data for only one year which we have excluded from the analysis.
- We have added another column continent to expand our analysis.

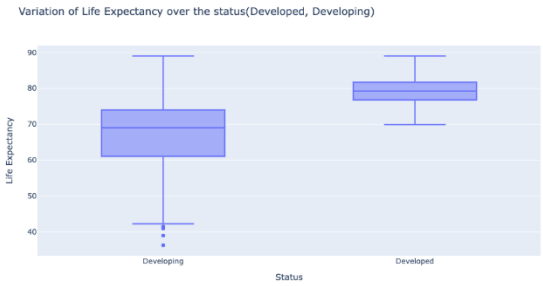
1.2 Data Analysis

Now we will explore few of the important factors which could affect life expectancy in various ways. The analysis will be expanded yearly, country wise and continent wise. We have focussed on factors which we felt would give insights to life expectancy and extended the analysis where we felt we could find few more patterns.

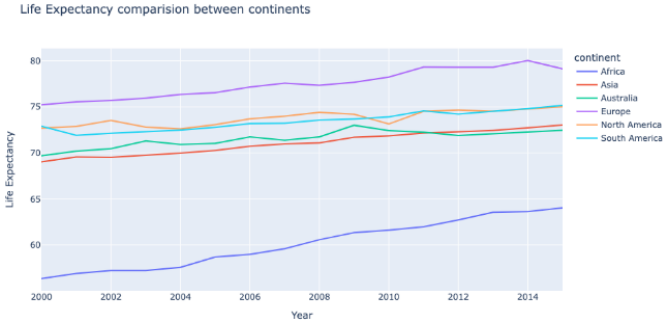
1.2.1 Understanding life expectancy

We have analysed how life expectancy has varied over the years by taking the mean for all the countries. We observed that there is a linearly increasing straight line for the years. As expected developed countries have a better life expectancy than developing countries.



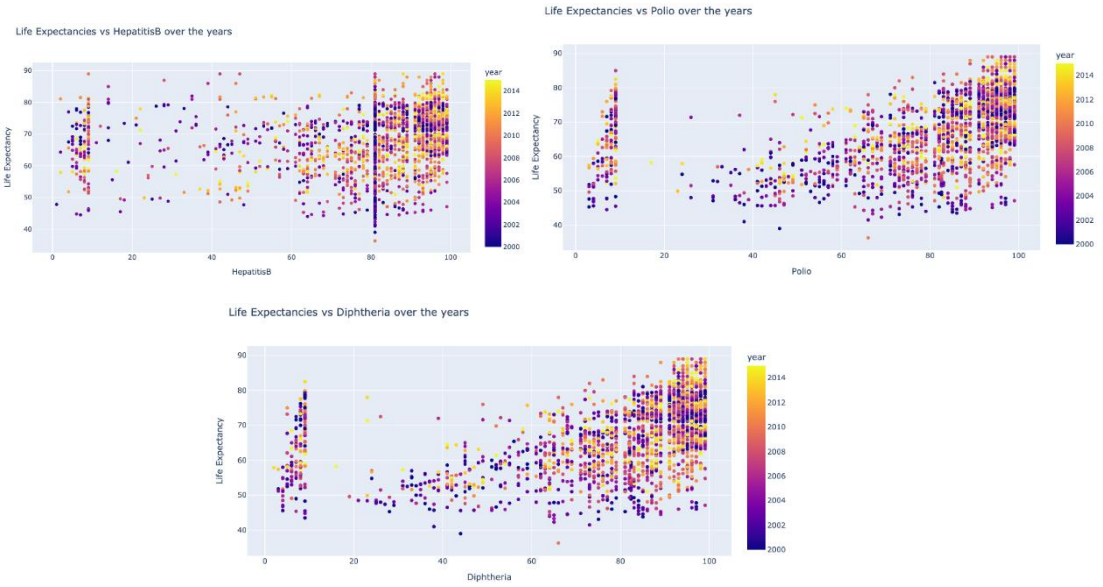


Analysis has been expanded to how the life expectancy has varied over continents and we observed that Africa has the lowest life expectancy whereas Europe has the highest.



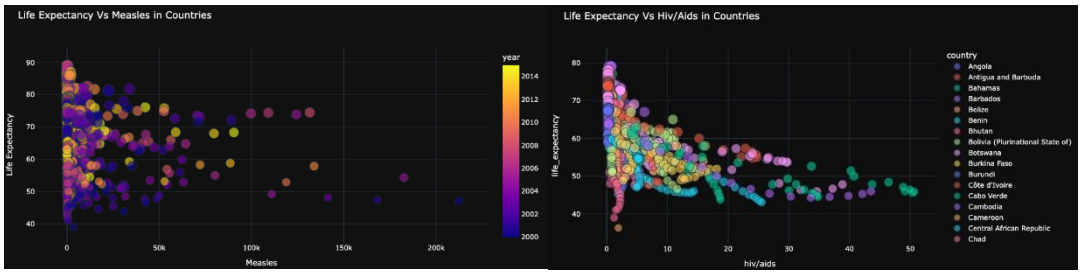
1.2.2 Effect of Immunization factors

We have plotted graphs for understanding the effect of immunization factors of hepatis, polio and diphtheria. None of the immunizations seem to have an impact on life expectancy.



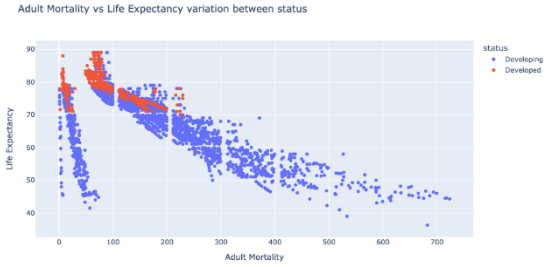
1.2.3 Effect of Diseases

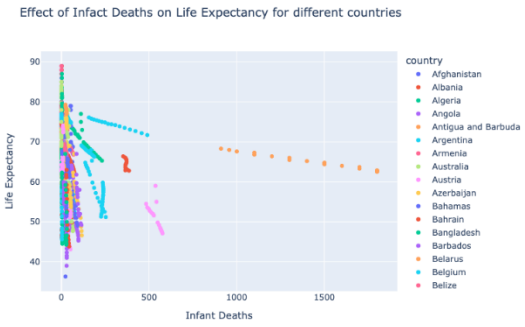
We have taken two important disease measles and HIV/AIDS given in the data to observe its significance on life expectancy. There was no observed significance to life expectancy but however we observed that the number of these diseases were more in the initial years and thereby there was a lower life expectancy.



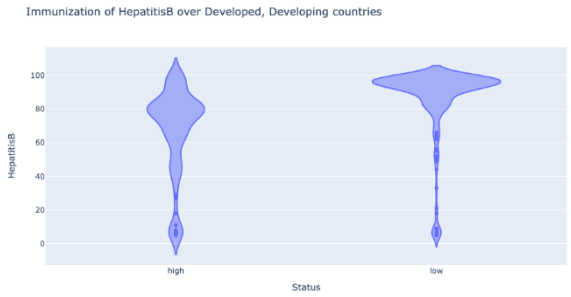
1.2.4 Mortality effects.

Adult mortality is the probability of dying between the ages of 15-60 and we see that there is a clear negative correlation to life expectancy. We have also observed that developing countries have a more adult mortality compared to developed countries.



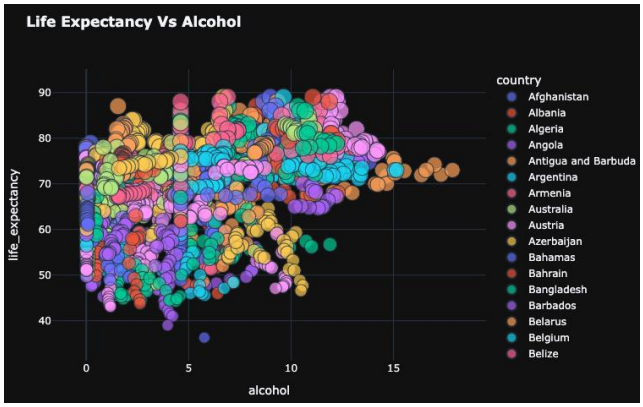


Infant Deaths: Although there is no relation of infant deaths to life expectancy, there were few countries which have larger number of infant deaths.



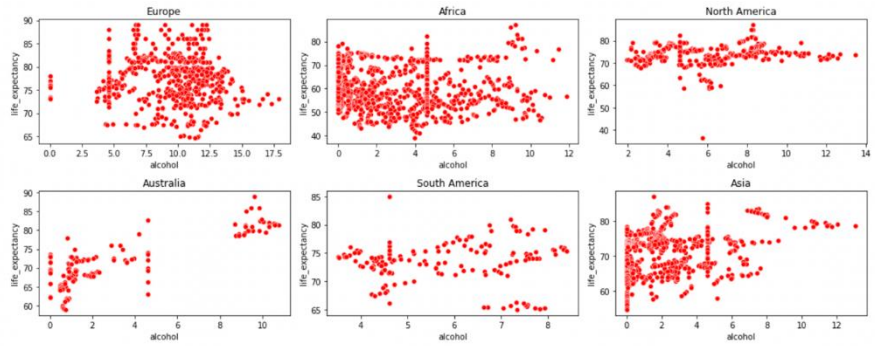
On further analysis after comparing with countries having a lower infant death that, countries having low immunization coverage have larger number of infant deaths.

1.2.5 Alcohol



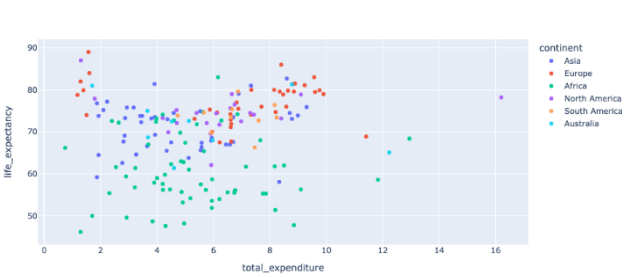
We wanted to analyse life expectancy vs alcohol however surprisingly we did not find much correlation with respect to life expectancy.

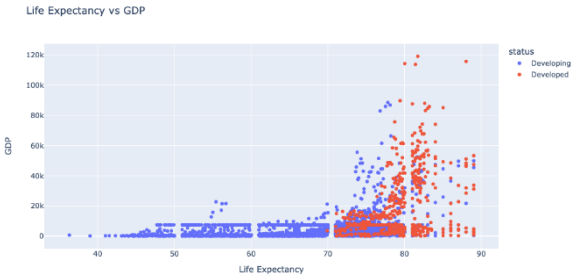
We extended our analysis to continents and we found few insights where Australia has a positive correlation but there is no other significance for other continents.



1.2.6 Effect of Expenditure

We have analysed the total expenditure (expenditure spent on health) and the Gross Domestic Product. Developing countries have a higher GDP and total expenditure and they have a higher life expectancy than developing countries. On continent analysis, Africa have a lower expenditure and therefore low life expectancy.





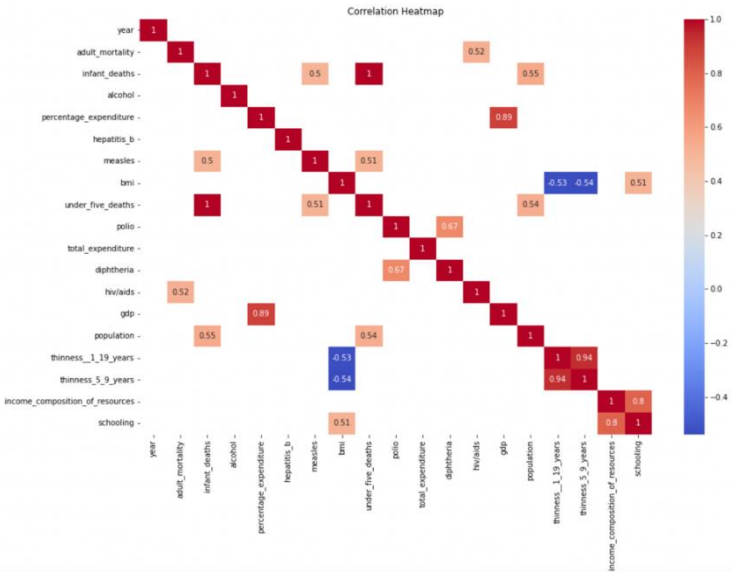
1.2.7 Schooling vs Life Expectancy

Education can be an important factor which could influence life expectancy. There data proves this as life expectancy increases as the number of schooling years are more. Developed countries have a better education and thereby have a better life expectancy.

1.3 Correlations

Finding correlations can be an important part of model building. We have correlation matrix where few of the factors are highly correlated.

1. Infant deaths and deaths under 5
2. Schooling, Income composition of resources
3. GDP, Percentage expenditure
4. Thinness_5_9_years, Thinness_10_19_years



2. Model Building:

In the dataset, for each country for a particular year and status, there were 19 predictors affecting the life expectancy initially. But, this model has high multicollinearity, low adjusted R^2, high AIC and BIC. Through different methods, we arrived at the model that is the best fit for the data. We discuss those steps involved in building the model below.

1. First, we tried to remove the unwanted predictors using both forward selection and backward elimination. We observed the model obtained through forward selection has a higher Ra^2, lower AIC and BIC compared to the backward elimination model. Hence, we decided to stick with the forward selection model which had 11 predictors.

```
Call:
lm(formula = life_expectancy ~ adult_mortality + alcohol + bmi +
    polio + total_expenditure + diphtheria + hiv.aids + gdp +
    thinness__1_19_years + income_composition_of_resources, data = dat)

Residuals:
     Min       1Q   Median       3Q      Max
-23.6995  -2.3965  -0.1097   2.3730  18.5689

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     5.522e+01  5.394e-01 102.378  < 2e-16 ***
adult_mortality                -2.168e-02  8.436e-04 -25.695  < 2e-16 ***
alcohol                         1.748e-01  2.407e-02   7.263 4.83e-13 ***
bmi                             5.565e-02  5.196e-03  10.710  < 2e-16 ***
polio                           3.536e-02  4.747e-03   7.450 1.22e-13 ***
total_expenditure               1.406e-01  3.507e-02   4.008 6.27e-05 ***
diphtheria                      4.553e-02  4.717e-03   9.652  < 2e-16 ***
hiv.aids                       -4.735e-01  1.881e-02 -25.172  < 2e-16 ***
gdp                             6.199e-05  6.984e-06   8.877  < 2e-16 ***
thinness__1_19_years           -1.105e-01  2.328e-02  -4.746 2.18e-06 ***
income_composition_of_resources  1.282e+01  5.343e-01  24.001  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.363 on 2927 degrees of freedom
Multiple R-squared:  0.7901,    Adjusted R-squared:  0.7894
F-statistic:  1102 on 10 and 2927 DF,  p-value: < 2.2e-16

[1] 17007.43
[1] 17079.25
```

```
Call:
lm(formula = life_expectancy ~ schooling + adult_mortality +
    hiv.aids + diphtheria + bmi + income_composition_of_resources +
    percentage_expenditure + polio + thinness__1_19_years + measles +
    alcohol, data = dat)

Residuals:
     Min       1Q   Median       3Q      Max
-22.3041  -2.2752  -0.0866   2.3523  16.6517

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     5.330e+01  5.220e-01 102.125  < 2e-16 ***
schooling                       6.746e-01  4.322e-02  15.609  < 2e-16 ***
adult_mortality                -2.083e-02  8.152e-04 -25.552  < 2e-16 ***
hiv.aids                       -4.804e-01  1.803e-02 -26.647  < 2e-16 ***
diphtheria                      4.112e-02  4.538e-03   9.060  < 2e-16 ***
bmi                             4.686e-02  5.024e-03   9.328  < 2e-16 ***
income_composition_of_resources  6.702e+00  6.448e-01  10.393  < 2e-16 ***
percentage_expenditure          3.397e-04  4.333e-05   7.840 6.25e-15 ***
polio                           2.996e-02  4.576e-03   6.549 6.84e-11 ***
thinness__1_19_years           -8.362e-02  2.254e-02  -3.710 0.000211 ***
measles                        -3.263e-05  7.002e-06  -4.660 3.30e-06 ***
alcohol                         1.044e-01  2.359e-02   4.427 9.93e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.189 on 2926 degrees of freedom
Multiple R-squared:  0.8066,    Adjusted R-squared:  0.8058
F-statistic:  1109 on 11 and 2926 DF,  p-value: < 2.2e-16
```

Fig: Model summary from backward elimination and forward selection

2. We checked for multicollinearity in the forward selection model, there was no significant multicollinearity between the predictors.

```
```{r}
vif(out.forward)
```
```

```
                  schooling                adult_mortality                       hiv.aids                   diphtheria
                   3.330788                      1.712109                       1.402114                     1.945302
                        bmi  income_composition_of_resources         percentage_expenditure                      polio
                   1.677436                      2.919145                       1.241491                     1.931136
      thinness__1_19_years                      measles                        alcohol
                   1.641504                      1.078763                       1.514769
```

Fig: VIF for the forward selection model

3. We plotted the normality, residual plots for the model. We found the distribution to be light tailed. Hence, we applied a box-cox transform to improve the normality.

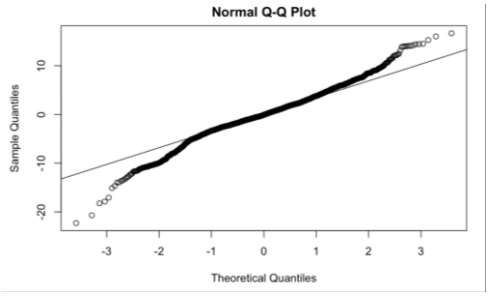
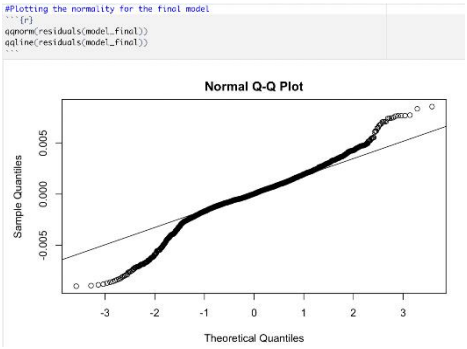


Fig: Normality plot of the forward selection model

4. After applying the box-cox transformation, the distribution was closer to the normal line.



5. Finally, we found the outliers using jack knife residuals and influential points. A total of 11 outliers were detected. We removed the outliers from the dataset and fit the final model.

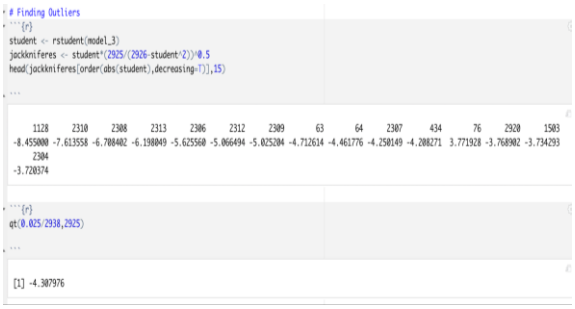


Fig: Outliers of the model

6. For the final model, the Ra^2 for 0.8124 and the distribution was very close to the normal distribution and the residuals had almost constant variance. Hence, we concluded that this model is the best fit for the data

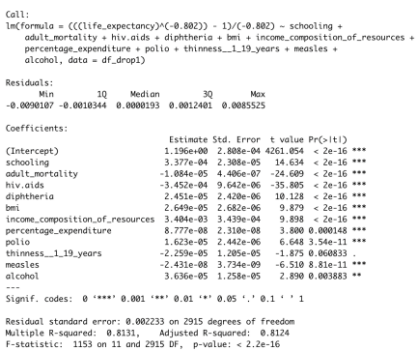


Fig: Final model summary

7. For this model, we performed different hypothesis tests and stated some important conclusions:

2.1 Conclusion 1

Immunization of diphtheria has more effect on life expectancy than the immunization of polio.

```
Linear hypothesis test

Hypothesis:
diphtheria - polio = 0

Model 1: restricted model
Model 2: (((life_expectancy)^(-0.802)) - 1)/(-0.802) ~ schooling + adult_mortality +
    hiv.aids + diphtheria + bmi + income_composition_of_resources +
    percentage_expenditure + polio + thinness__1_19_years + measles +
    alcohol

  Res.Df      RSS Df  Sum of Sq      F  Pr(>F)
1   2916 0.014554
2   2915 0.014536  1 1.8072e-05 3.6241 0.05705 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2.2 Conclusion 2

HIV has more impact on life expectancy than measles

```
Linear hypothesis test

Hypothesis:
- hiv.aids  + measles = 0

Model 1: restricted model
Model 2: (((life_expectancy)^(-0.802)) - 1)/(-0.802) ~ schooling + adult_mortality +
    hiv.aids + diphtheria + bmi + income_composition_of_resources +
    percentage_expenditure + polio + thinness__1_19_years + measles +
    alcohol

  Res.Df      RSS Df Sum of Sq      F    Pr(>F)
1   2916 0.020928
2   2915 0.014536  1 0.0063919 1281.8 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2.3 Conclusion 3

Consumption of alcohol does not have much impact on life expectancy.

```
Linear hypothesis test

Hypothesis:
alcohol = 0

Model 1: restricted model
Model 2: (((life_expectancy)^(-0.802)) - 1)/(-0.802) ~ schooling + adult_mortality +
    hiv.aids + diphtheria + bmi + income_composition_of_resources +
    percentage_expenditure + polio + thinness__1_19_years + measles +
    alcohol

  Res.Df      RSS Df  Sum of Sq      F   Pr(>F)
1   2916 0.014578
2   2915 0.014536  1 4.1644e-05 8.3511 0.003883 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

CONCLUSION

Upon a detailed analysis we can arrive at following conclusions:

1. Developed Countries have a better life expectancy compared to developing countries.

2. Schooling has a great impact on life expectancy i.e. a greater number of schooling years have a greater life expectancy. This indeed makes sense as people who are highly educated have a better understanding on health.

3. Immunization factors and alcohol do not seem to impact life expectancy as one might expect them to affect.

4. Countries having low life expectancy when compared to the ones which have high really need to improve all of the immunization factors and economic factors. Countries belonging in Africa seem to have the lowest life expectancy out of all.

REFERENCES

1. https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who?datasetId=12603&searchQuery=R
2. https://people.umass.edu/biep640w/pdf/diagnosticsStudentNotes.pdf
3. Linear Models in R by Julian Faraway