LIFE EXPECTANCY DATA STUDY AND ANALYSIS

A PROJECT REPORT

Submitted by

G. SHIVA SAM KUMAR

MOHANRAJ BABU

GOWTHAM DAKSHNAMOORTHY

TANMAY RAUT

in partial fulfillment of the requirements for the Subject

of

LINEAR REGRESSION

ABSTRACT

Life expectancy is of major relevance to every country in order to examine adequate economic and social conditions and assist the population groups in greatest need. This analysis report looks into the impact of several representative criteria on life expectancy. Data from 193 nations are evaluated using visualization techniques to reveal the relationship between several elements that influence life expectancy. Regression techniques have been employed to establish relationships between dependent variables (life expectancy) and several other independent factors. Filtering out potential predictors was accomplished using a coefficient Heat map, ANOVA, and LASSO regression. The model is verified in order to choose the optimum model and to overcome overfitting.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

Denis Witley once said that time and health are two valuable assets that we often take for granted until we lose them. Everyone wants to live a long and happy life, but to do so, we need to make healthy lifestyle choices like maintaining a good diet, exercising regularly, and living in a safe environment. However, factors like health issues, accidents, and diseases can also affect our life expectancy. While it may take years of research to determine the positive and negative factors that affect life expectancy, we can focus on factors that diminish life expectancy.

Previously, studies have been conducted to explore the elements that affect life expectancy. These investigations have considered demographic variables, income composition, and mortality rates. In any case, the effect of vaccination and human development index has not been viewed as in these examinations. Furthermore, a few studies were restricted to involving just a single year of information for all nations and utilized numerous direct relapse investigations. To address these limits, another review will utilize a blended impacts model and numerous straight relapses to dissect information from 2000 to 2015 for all nations. The review will likewise consider significant vaccinations like Hepatitis B, Polio, and Diphtheria, as well as mortality, monetary, social, and other well-being related factors. By breaking down information from various nations, this study plans to recognize the key factors that add to bring down the future, permitting nations to focus on working on those particular regions to expand their population's life expectancy.

Despite efforts to improve public health, many countries continue to struggle with low life expectancies. While previous studies have identified various factors that impact life expectancy, there is still a need for research to identify the most effective interventions that can be implemented to improve life expectancy in specific regions or populations. By identifying the key factors that contribute to low life expectancies and developing targeted interventions, public health officials can improve health outcomes and quality of life for populations around the world. Therefore, there is a need for further research to identify the most effective strategies for improving life expectancy in different regions and populations.

The aim to concentrate on factors influencing the future is significant as it can give important insight into the areas that should be addressed to work on personal satisfaction for individuals. Life Expectancy is a critical sign of a country's general wellbeing and advancement, and by distinguishing the variables that add to a lower life expectancy, nations can foster successful strategies and projects to work on the wellbeing and prosperity of their citizens. Additionally, the review will give a chance to look at the effect of vaccination in the future. Immunization is a basic general wellbeing meditation that has been displayed to decrease the occurrence and seriousness of numerous irresistible infections. Nonetheless, its effect on the future has not been widely examined, and this study will assist with filling this hole in information.

From this study we can determine the impact of alcohol consumption on life expectancy, we can perform a statistical analysis that compares the life expectancy of people who consume alcohol with those who do not. While many scientific tests could be conducted to find the answer to this, we can

instead focus on statistics. We can consider the population of specific countries and their recorded life expectancy, the types of diseases they are affected with, and perform a regression analysis to see if these factors have a positive or negative relationship with life expectancy.

We can also consider a country's GDP, percentage, and total expenditure, as these factors can determine a person's livelihood, which in turn can affect their life expectancy. By comparing the results of tests using regressors from five different factors – diseases, alcohol consumption, dietary maintenance, economic data, and type of deaths – we can determine which factor has the most significant impact on life expectancy. While diet is a challenging metric to measure, we can use BMI data to determine its impact. We will average the results of each subcategory and compare them to see which factor affects life expectancy the most to obtain a conclusion.

Moreover, the review's emphasis on numerous variables, including economic, social, and health-related factors, will provide a comprehensive understanding of the complex interplay between these factors and their effect on life expectancy. This will empower nations to foster a more comprehensive way to deal with further developing wellbeing results and diminishing wellbeing disparities. Overall, the study on factors affecting life expectancy can possibly make a critical commitment to worldwide wellbeing and advancement by distinguishing the regions that should be addressed to work to improve the quality of life around the world.

2.1 DATA SOURCE

The data utilized in our study was accumulated from different sources, including the Global Health Observatory (GHO) data repository maintained by the World Health Organization (WHO), and the United Nations website for economic data. The GHO information archive gathers data on different wellbeing pointers, including future, illness predominance, death rates, and health system performance. The information was gathered over a time of 15 years, from 2000 to 2015, which is a critical period for worldwide wellbeing improvement, particularly in non-industrial nations. The informational collection incorporates data for 193 nations, covering a large number of elements connected with wellbeing and financial matters. The data collection was accumulated by consolidating individual information documents, which were then cleaned and handled utilizing the R programming. The Miss map order was utilized to distinguish and deal with missing information, which for the most part consisted of population, Hepatitis B, and Gross domestic income information, and was principally from less notable nations.

The World Health Organization (WHO) collects data on diseases affecting various countries by utilizing hospital and government records, which they have a shared perspective on. Similarly, information on alcohol consumption is obtained through surveys and interviews. Although the exact method that WHO employs to gather this data is difficult to ascertain due to the numerous sources they rely on, their reputation for providing factual information that is accessible to the public is well established. WHO's overarching goal is to improve the well-being of all people and their approach is guided by science. Therefore, their data collection methods are designed to provide accurate and reliable information that can help in identifying public health trends and designing appropriate interventions.

Missing data were found and handled using the Miss map program in R software to assure the data set's accuracy. This enabled a more comprehensive data collection to be used in the research. However, certain countries were omitted from the final data collection due to lacking data. When working with huge data sets, missing data might have an impact on the accuracy and completeness of the research. In addition, the data collection is divided into four major categories: immunization-related factors, mortality factors, economic factors, and social factors. This categorization enables a more extensive study of the data, allowing researchers to uncover patterns and linkages between various elements that may influence health outcomes.

Overall, the data set used in this project provides a valuable resource for researchers and policymakers in the field of global health. It allows for a comprehensive analysis of health-related factors and their relationship to economic development, providing insights into the complex interplay between these factors and their impact on health outcomes around the world.

CHAPTER 2

In the research study, the independent variable that will remain constant is "life expectancy." The study aims to establish a relationship between life expectancy and various predictors, including alcohol consumption rate, prevalence of diseases such as Hepatitis B, Measles, Polio, Diphtheria, and HIV/AIDS, economic indicators such as Percentage expenditure, Total expenditure, Per Capita Income, and GDP, the total population, and types of death including adult mortality and under 5 deaths. Additionally, the study will consider dietary data and compare the BMI metric with the recommended BMI reading for a healthy individual. The data for each variable has been collected for 15 years, spanning from 2000 to 2015, and is specific to each country.

2.1 Variable Descriptions:

In this Analysis, we have provided a comprehensive description of all the variables and parameters used. Each entry will include the variable name, its type, and a brief description. The values for each variable are specific to the parameter they represent.

❖ Countries (Categorical variable)

The dataset includes 193 different countries, each presenting its unique data for each parameter in the 15-year period. Although this variable will not be used directly in the results, all the data that the countries present will be utilized. Since it is a categorical variable, we will represent it through a frequency distribution chart. However, as there are 193 countries, we will provide a small example. Each country has a frequency of 15, indicating the number of observations per year from 2000 to 2015.

Countries	Afghanistan	Albania	Algeria
Frequency	15	15	15

Chart 1: Frequency Distribution of Countries

❖ Years (Quantitative variable)

The data represents each year in the 15-year timeframe, from 2000 to 2015. We will compute the total data for each year and display them in a histogram to observe any patterns or changes over time.

However, this will not reveal any relationships between variables. As a continuous variable, the results will be presented in the form of a histogram.

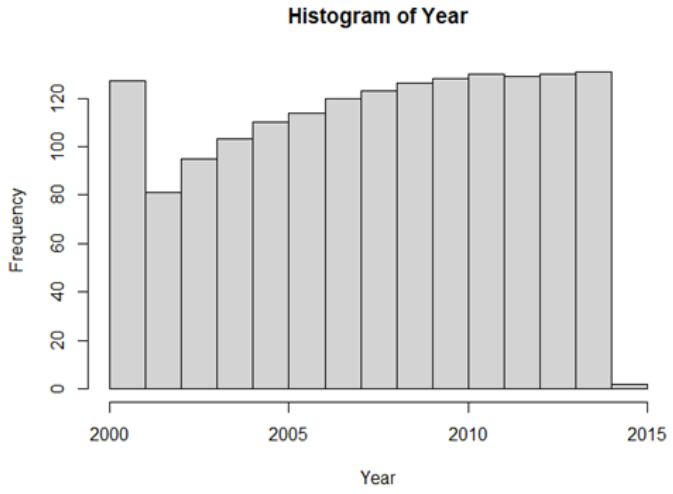

Figure 1.1: Histogram of Years

The variables related to "types of death" are listed below:

❖ Adult Mortality (Dependent Variable)

This variable represents the probability of death for individuals aged between 15 and 60 per 1000 people, regardless of gender. As this age range is considered to be a fundamental range of age for living, this variable could be used as a reliable metric for any deaths that occur accidentally. The data shows the number of adult deaths in this age range per thousand people in each country per year from 2000 to 2015. However, the total number of deaths will be considered for each year, and the results will be presented in the form of a histogram.

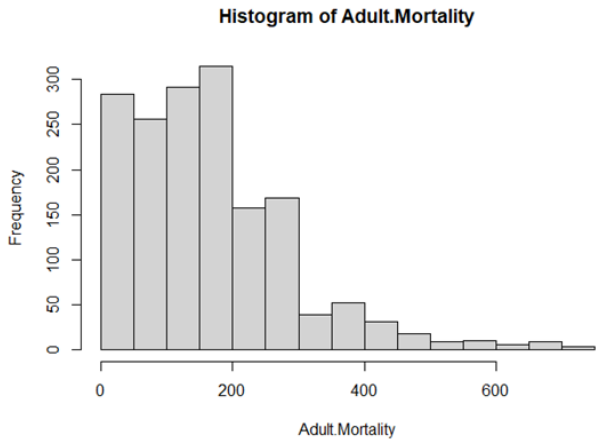

Figure 1.2: Histogram of Adult Mortality

❖ Under 5 deaths (Dependent Variable)

This data presents the number of deaths of children under the age of 5 in each country per year from 2000 to 2015. This variable is independent of the other variables and holds significance in establishing its association with life expectancy. The data has been collected for each year and will be considered as a total. The data is continuous and presented in the form of a histogram.

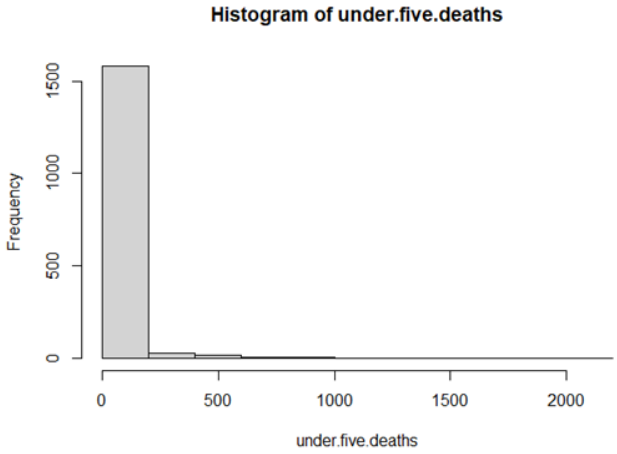

Figure 1.3: Histogram representing deaths of children under the age of 5.

❖ Alcohol Consumption (Dependent variable)

The data on alcohol consumption is crucial to this project as it helps determine if it plays a role in increasing deaths and reducing life expectancy. The variable is continuous and expressed as a percentage of the total population consuming alcohol. The data has been collected for each country from

2000 to 2015 and will be considered as a total for each year. The results will be presented in the form of a histogram.

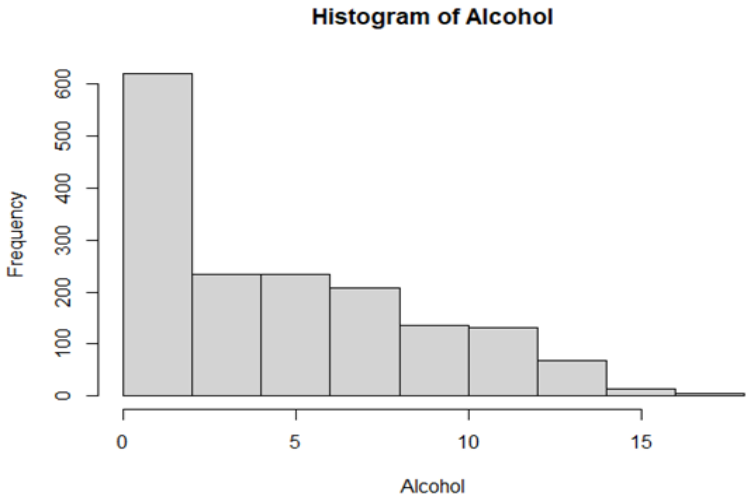

Figure 1.4: Histogram showing alcohol consumption rate.

❖ Hepatitis B (Dependent variable)

This data represents the number of deaths caused by Hepatitis B in each country from 2000 to 2015. The variable is continuous, and the data will be considered as a total for each year. The results will be presented in the form of a histogram.

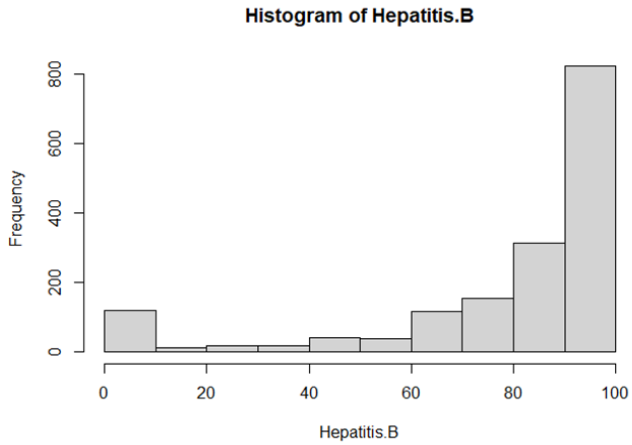

Figure 5: Histogram representing deaths caused by Hepatitis B.

❖ Measles (Dependent variable)

The data on Measles shows the number of deaths caused by Measles in each country from 2000 to 2015. The variable is continuous, and the data will be considered as a total for each year. The results will be presented in the form of a histogram.

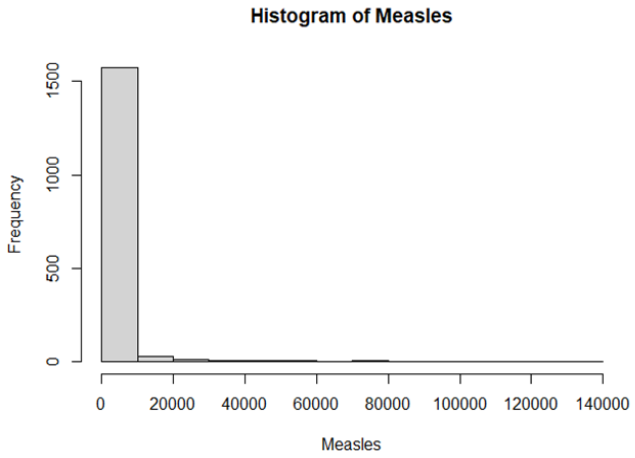

Figure 6: Histogram representing deaths caused by Measles.

❖ Diphtheria (Dependent variable)

The data on Diphtheria shows the number of deaths caused by Diphtheria in each country from 2000 to 2015. The variable is continuous, and the data will be considered as a total for each year. The results will be presented in the form of a histogram.

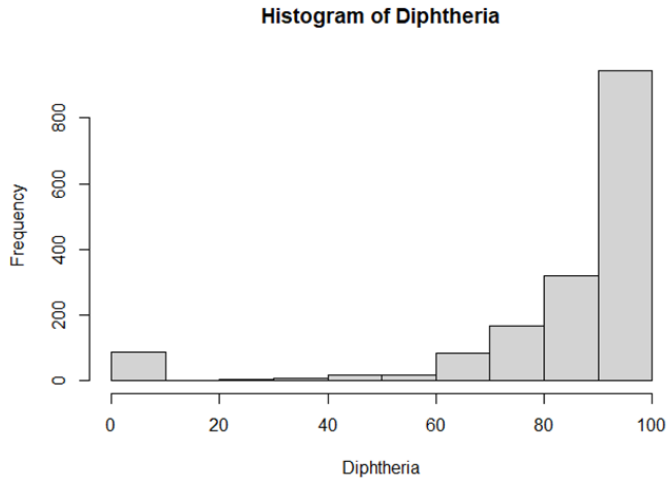

Figure 7: Histogram representing deaths caused by Diphtheria.

❖ Polio (Dependent variable)

The data on Polio shows the number of deaths caused by Polio in each country from 2000 to 2015. The variable is continuous, and the data will be considered as a total for each year. The results will be presented in the form of a histogram.

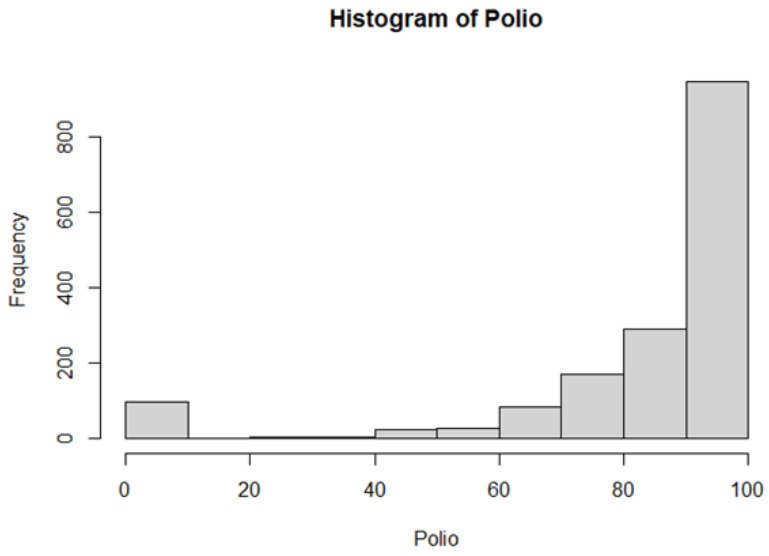

Figure 8: Histogram representing deaths caused by Polio.

❖ HIV/AIDS (Dependent variable)

This data shows the number of deaths caused by HIV/AIDS in each country from 2000 to 2015. The variable is continuous, and the data will be considered as a total for each year. The results are calculated by dividing the estimated number of persons newly infected with HIV/AIDS during the specified time by the number of persons at risk of HIV/AIDS infection. The results will be presented in the form of a histogram.

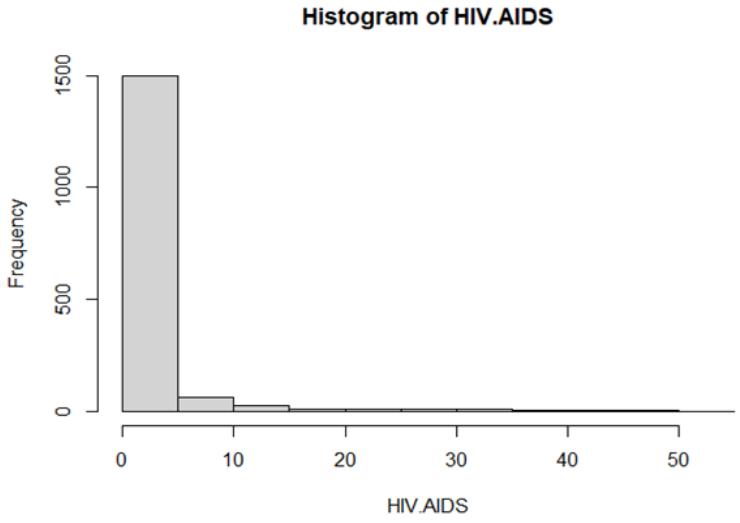

Figure 9: Histogram representing deaths caused by HIV/AIDS.

The Following are subcategories for “Economic data.”

❖ Population (Dependent variable):

This data provides information about the population size of each country for every year from 2000 to 2015. The data will be used to analyze its correlation with life expectancy, and it will be considered from an economic perspective since it determines a country's expenditure and GDP. The total population data for each year will be presented as a continuous variable and in the form of a histogram.

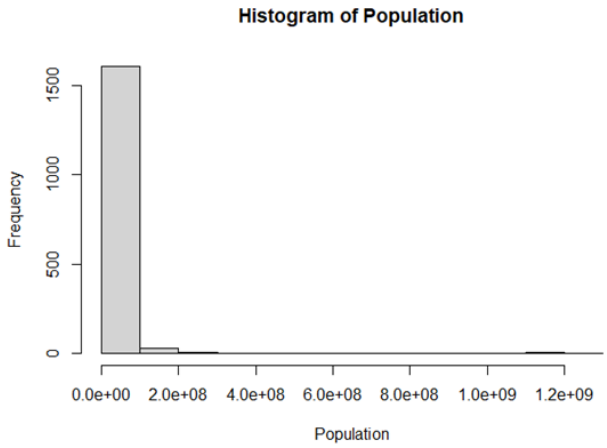

Figure 10: Histogram of Population

❖ Per Capita Income (Dependent variable):

This data presents the per capita income of each country for each year from 2000 to 2015. The data will be calculated in total for each year and is a continuous variable. It will be analyzed to understand its correlation with life expectancy. The per capita income is calculated by dividing a country's national income by its population. The results will be presented as a continuous variable and in the form of a histogram.

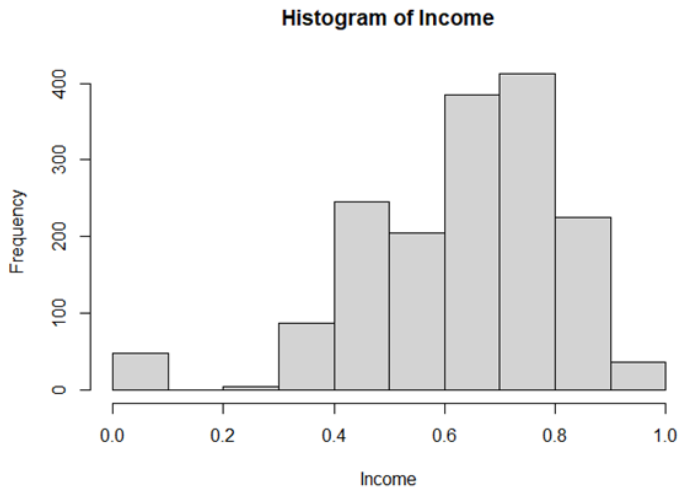

Figure 11: Histogram of Per Capita Income

❖ Percentage Expenditure (Dependent variable):

This data shows the percentage expenditure of each country for every year from 2000 to 2015. The data will be calculated in total for each year and is a continuous variable. The analysis will focus on its correlation with life expectancy. The percentage expenditure is calculated by dividing the average expenditure per household for an item by total expenditure and multiplying it by 100. The results will be presented as a continuous variable and in the form of a histogram.

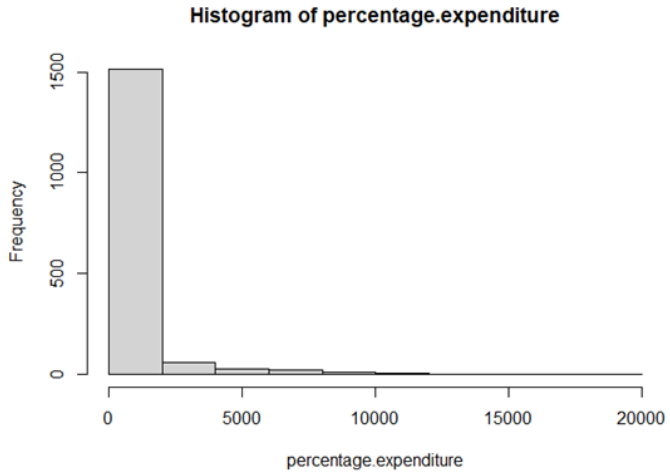

Figure 12: Histogram of Percentage Expenditure

❖ Total Expenditure (Dependent variable):

This data shows the total expenditure of each country for every year from 2000 to 2015. The data will be calculated in total for each year and is a continuous variable. It will be analyzed to understand its correlation with life expectancy. The total expenditure is calculated by summing up the price paid for one or more products by the amount of each item purchased. The results will be presented as a continuous variable and in the form of a histogram.

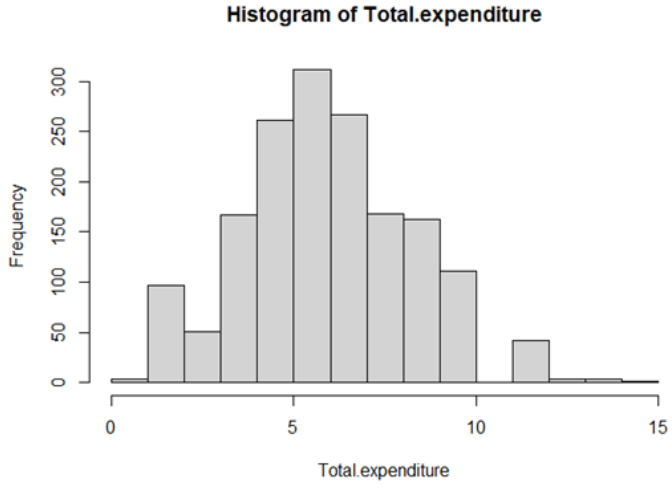

Figure 13: Histogram of Total Expenditure

❖ GDP (Dependent variable):

This data shows the GDP (Gross Domestic Product) of each country for every year from 2000 to 2015. The data will be calculated in total for each year and is a continuous variable. It will be analyzed to understand its correlation with life expectancy. The GDP is calculated by summing up private consumption, gross private investment, government spending, and the difference between exports and imports. The results will be presented as a continuous variable and in the form of a histogram.

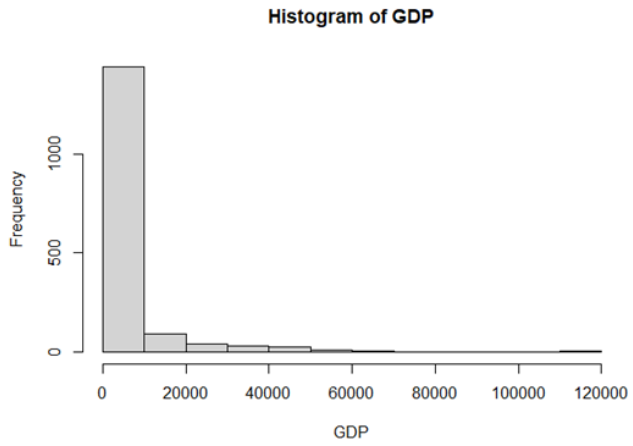

Figure 15: Histogram of GDP

The following are subcategories for "Dietary Maintenance":

❖ Body Mass Index (BMI) (Dependent variable)

This data shows the BMI value for each country from 2000 to 2015. BMI is a continuous variable and is calculated by dividing a person's weight in kilograms by their height in meters squared. This data will be used to examine its relationship with life expectancy, as BMI is a good indicator of a person's overall health and dietary habits. The healthy BMI range of 18.5 to 24.9 will be used as a reference point to see how healthy people are and how it affects life expectancy. The results will be presented in the form of a histogram.

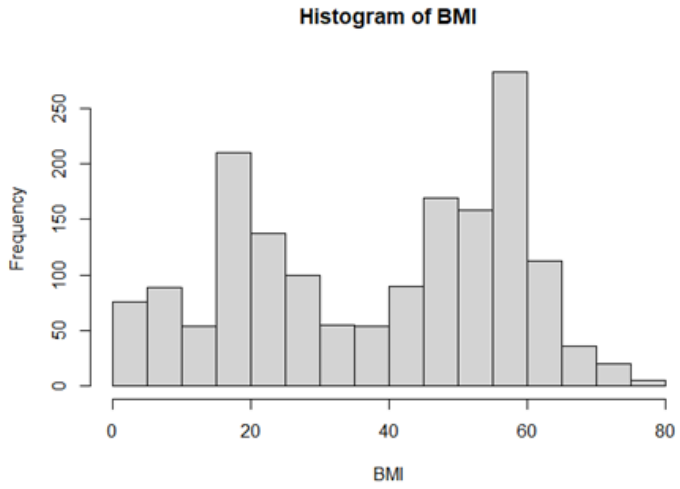

Figure 14: Histogram of BMI

❖ Life Expectancy (Independent variable)

This data is the centerpiece of this project, showing the life expectancy for each country from 2000 to 2015. This is a continuous variable that will be used throughout the project to see how the different variables mentioned before affect this independent variable. The life expectancy values are calculated using the "Farr's death rate equation method," which takes into account the number of deaths at a specific age and the average population at that age. The probability of death from a particular age is also calculated. The results will be presented in the form of a histogram.

Figure 16: Histogram of Life Expectancy

Likewise all the descriptions about the variables are further stated in the table below. Which gives us more information on the variable and the type of variable. It also includes the range of the variable and type of distribution, which is suitable for.

DATA VARIABLE	DATA INFO	TYPE OF VARIABLE	RANGE	DISTRIBUTION TYPE
Countries	Country	Categorical Nominal Variable	193 Unique Values	Categorical Distribution
Years	Year	Numeric discrete Variable	2000 to 2015	Discrete Uniform Distribution
Adult Mortality	Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)	Numeric Continuous variable	1 - 723	Poisson Distribution
Under 5 Deaths	Number of under-five deaths per 1000 population	Numeric discrete Variable	0 - 2500	Poisson Distribution
Alcohol Consumption	Alcohol, recorded per capita (15+) consumption (in liters of pure alcohol)	Numeric Continuous variable	0.01 - 17.9	Gamma Distribution
Hepatitis B	Hepatitis B (HepB) immunization coverage among 1-year-olds (%)	Numeric Continuous variable	1 - 99	Binomial Distribution

Measles	Measles - number of reported cases per 1000 population	Numeric discrete Variable	0 - 212k	Poisson Distribution
Diphtheria	Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)	Numeric discrete Variable	2 - 99	Binomial Distribution
Polio	Polio (Pol3) immunization coverage among 1-year-olds (%)	Numeric Continuous variable	3 - 99	Binomial Distribution
HIV/ AIDS	Deaths per 1 000 live births HIV/AIDS (0-4 years)	Numeric Continuous variable	0.1 - 50.6	Poisson Distribution
Population	Population of the country	Numeric discrete Variable	34 - 1.29B	Normal Distribution
Income Composition	Human Development Index in terms of income composition of resources (index ranging from 0 to 1)	Numeric Continuous variable	0 - 0.95	Normally distributed or skewed distribution
Percentage Expenditure	Expenditure on health as a percentage of Gross Domestic Product per capita(%	Numeric Continuous variable	0 - 19.5k	Normal Distribution

Total Expenditure	General government expenditure on health as a percentage of total government expenditure (%)	Numeric Continuous variable	0.37 - 17.6	Normal Distribution
GDP	Gross Domestic Product per capita (in USD)	Numeric Continuous variable	1.68 - 119k	Normal Distribution
BMI	Average Body Mass Index of entire population	Numeric Continuous variable	1 - 87.3	Normal Distribution
Life Expectancy	Life Expectancy in age	Numeric Continuous variable	36.3 - 89	Normal Distribution
Status	Developed or Developing status	Categorical Nominal Variable	Developing 83% Developed 17%	Categorical Distribution
infant deaths	Number of Infant Deaths per 1000 population	Numeric Discrete variable	0 - 1800	Poisson Distribution
thinness 1-19 years	Prevalence of thinness among children and adolescents for Age 10 to 19 (%)	Numeric Continuous variable	0.1 - 27.7	Normally distributed or skewed distribution
thinness 5-9 years	Prevalence of thinness among children for Age 5	Numeric Continuous variable	0.1 - 28.6	Normally distributed or skewed distribution

	to 9(%)			
Schooling	Number of years of Schooling(years)	Numeric discrete Variable	0 - 20.7	Normally distributed or skewed distribution

Table 3: Summary of the Data

CHAPTER 3

4.1 Preprocessing of Dataset

As discussed above the dataset contains various numerical and categorical variables. It is vast as it has nearly 2900 data points. As this is real time data, it can be expected to be noisy with random variance, unwanted values, errors, and outliers. To ensure consistency and understanding of the dataset, preprocessing is a vital step for that.

Data preprocessing is an essential step in any machine learning project, including linear regression. It involves transforming raw data into an understandable format by cleaning the data and making it suitable for a machine learning model. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. It helps increase the accuracy and efficiency of a machine learning model. It also ensures that the performance of the dataset is optimal, and it makes knowledge discovery faster. With respect to the dataset in this project the following issues must be addressed and eliminated.

- Missing values of the dataset
- Skewness of data for numerical variables.

4.2 Missing values of the dataset

Usually missing values in a dataset are due to human error, and glitch in the data collection system and problems as such. This leads to loss of significant information of the dataset and sampling bias. Missing values affects the statistical power of the analysis and it will in turn affect the validity of the results. Therefore, it is imperative to resolve this issue. There are different methods to handle this issue which could leverage the performance of the dataset. But as numerical data is considered in this dataset, three methods were implemented to fill out the missing values in the dataset.

- Mean Imputation
- Median Imputation
- Interpolation

a) Mean Imputation

Imputation is a simple method to fill out the missing data in the dataset with respect to the data which is available in the dataset. One of the simplest imputation is mean imputation where the missing data for the variables are being filled with the mean of the data under the variable. This method ensures the sample size and it is convenient to use. But it tends to reduce the variability of data. And another factor needs to be considered in this, is the randomness in location of missing

data. If the missing data is spread in a random manner, then mean imputation will not lead to a biased analysis. But, if the spread of missing data is skewed, then it will lead to biased analysis, where the mean imputation method could not be used. It also needs to be noted that the mean imputation does not preserve relationships among variables. First to have an idea of the amount of missing data, it is needed to visualize the amount of data missing in the dataset, and the following plot represents that.

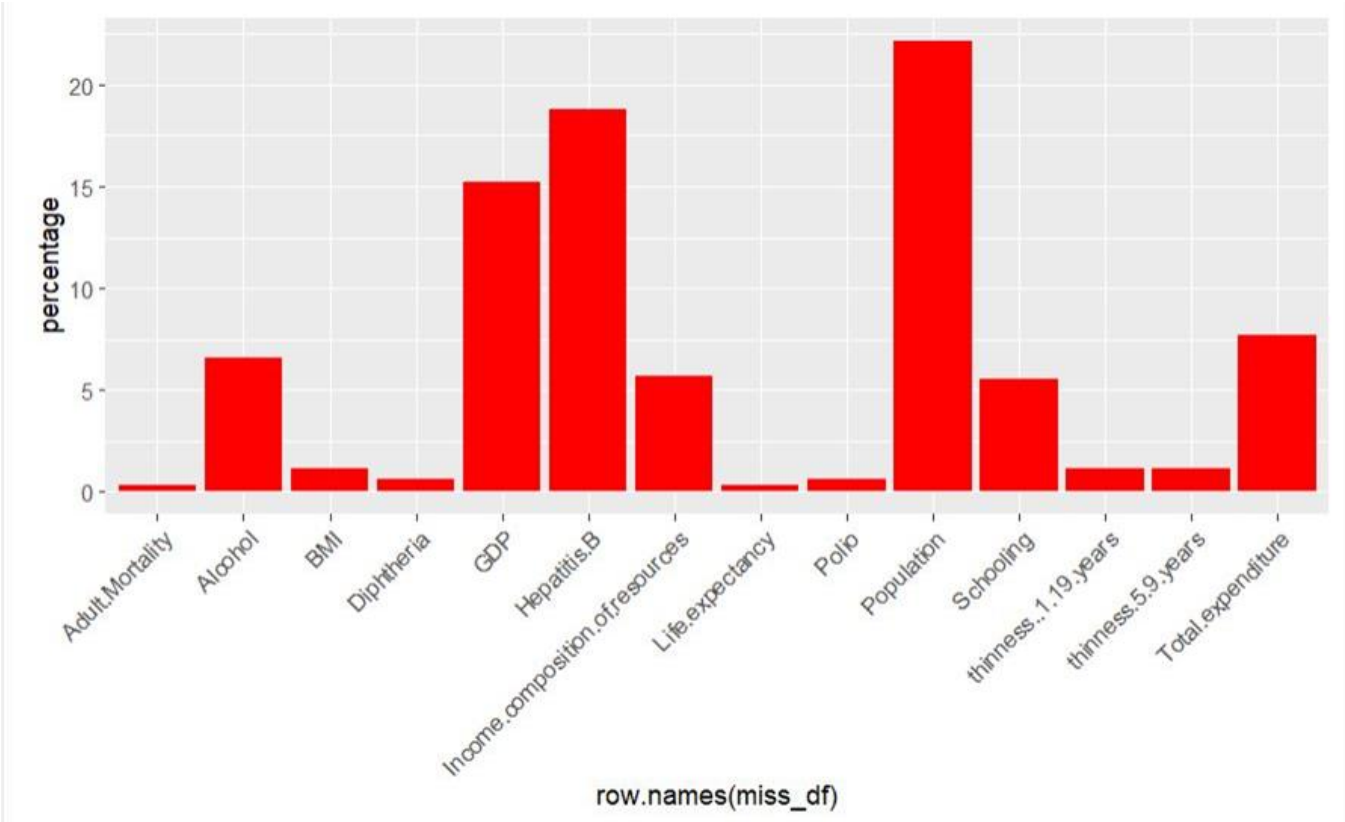

Fig 4.2 Missing Values Graph

Then by means of mean imputation, the missing data is being filled out and the RMSE score is being calculated for this.

b) Median Imputation

The drawbacks of the mean imputation could be overcomed by median imputation. Median imputation is a simple method, similar to mean imputation where the missing values for the variables are being filled out with the median of the corresponding variables calculated with the existing values. The median imputation method does not tamper with the standard deviation of the dataset, and it will ensure the variance of the dataset. Unlike mean imputation, this technique can be used in

cases where the spread of missing values is skewed, which makes it less viable for biased analysis. It ensures the relationship between the variables and it is robust in nature for outliers unlike mean imputation. But on the other hand, median imputation might lead to loss of data.

c) Interpolation

Interpolation is a method where new points are being generated between the existing points without the loss of pattern existing between the points. This eliminates the possibility of outliers being generated. In this method, with the help of data present under every variable, a function is being generated which explains the trend of data variance for that variable. Then this function is used to extrapolate the missing data. This is how the pattern is ensured for every variable and it does not account for the uneven spread of the missing data, as the missing points are being interpolated with respect to the nearest points from it. Even if there are several outliers, the errors caused due to outliers would be very much less compared to the other previous methods.

Then by means of interpolation, the missing data is being filled out and the RMSE score is being calculated for this. From the RMSE scores of the corresponding datasets, it is evident that interpolation method is the optimal method to fill out missing data as it has lower RMSE score.

3.3 Skewness of data for numerical variables

Skewness could be defined as a statistical measure which would explain the asymmetrical distribution of data over the dataset points. This makes it essential to understand the shape of the data spread. In other words, skewness tells how much the data is deviating from a gaussian distribution of the given dataset. This is also to be considered as one of the crucial problems as skewed data would lead to the degradation of the model's performance. Skewness can be classified as two types i.e., left and right skewed based on the direction in which the value plots are high. This will change the comparison between mean, median and mode of a dataset. Usually the skewness is calculated by a term known as Pearson's First and Second Coefficient which is dependent upon the mean and mode of the dataset. But in this case, the skewness is calculated by the most commonly used formula as below

$$\tilde{\mu}_3 = \frac{\sum_i^N (X_i - \bar{X})^3}{(N-1) * \sigma^3}$$

$\tilde{\mu}_3$ = skewness
N = number of variables in the distribution
X_i = random variable
$\bar{X}$ = mean of the distribution
σ = standard deviation

With the help of the above formula, the skewness is being calculated for the numerical variables and they are as follows:

```
[1] "Life.expectancy -: -0.635856669158128"
[1] "Adult.Mortality -: 1.17172589519657"
[1] "infant.deaths -: 9.78030062030323"
[1] "Alcohol -: 0.609030000411964"
[1] "percentage.expenditure -: 4.64888453275825"
[1] "Hepatitis.B -: -1.58984428227641"
[1] "Measles -: 9.43490490012689"
[1] "BMI -: -0.217067070312807"
[1] "under.five.deaths -: 9.48860103228038"
[1] "Polio -: -2.08628016732132"
[1] "Total.expenditure -: 0.619390409871752"
[1] "Diphtheria -: -2.06115278591113"
[1] "HIV.AIDS -: 5.39243871874848"
[1] "GDP -: 3.30244457730235"
[1] "Population -: 17.22294470789"
[1] "thinness..1.19.years -: 1.67634269718188"
[1] "thinness.5.9.years -: 1.7384013998031"
[1] "Income.composition.of.resources -: -1.16698627027287"
[1] "Schooling -: -0.616449433074087"
```

From the above R Output, it is evident that the skewness values are varying between -2 to 17.5. This shows that the data is being highly skewed and it needs to be resolved. This is done by a method called normalization. This method helps the most in skewed data as it aids in reducing the impacts of the outliers or extremities which are caused due to the skewness. Various transformations could be considered to normalize the data such as log transformation, box cox transformation etc., but the problem is that these transformations are being applied aggressively over the dataset which makes it complex to be applied on big datasets like this. Therefore it is optimal to go for a simpler transformation, where in this case square root transformation is being considered. This is a statistical method where square root is being taken for the data under the variable for transformation purpose. The advantage of using this method is that it is easy to implement and it does not factor in the size of the dataset, which makes it much more versatile. In this method the scale of the data is being compressed and the larger values are reduced in magnitude. This helps in normalizing the data and making the data more symmetrical. So, the dataset is being transformed by square root transformation and then again the skewness is being calculated using the formula discussed above. The calculated skewness coefficients are as follows.

```
[1] "Life.expectancy is: -0.635856669158128"
[1] "Adult.Mortality is: 1.17172589519657"
[1] "infant.deaths is: 0.679958962571999"
[1] "Alcohol is: 0.609030000411964"
[1] "percentage.expenditure is: 0.729238115659287"
[1] "Hepatitis.B is: -1.58984428227641"
[1] "Measles is: 1.65910196085473"
[1] "BMI is: -0.217067070312807"
[1] "under.five.deaths is: 0.727920277782387"
[1] "Polio is: -2.95866101548017"
[1] "Total.expenditure is: 0.619390409871752"
[1] "Diphtheria is: -2.06115278591113"
[1] "HIV.AIDS is: 1.88517466390931"
[1] "GDP is: 0.747794352536623"
[1] "Population is: 1.19313265642152"
[1] "thinness..1.19.years is: 1.67634269718188"
[1] "thinness.5.9.years is: 1.7384013998031"
[1] "Income.composition.of.resources is: -1.16698627027287"
[1] "Schooling is: -0.616449433074087"
```

From the above R -Output it is evident that there is a drastic difference in the skewness of the variables before and after the transformation. As the range of skewness is between -3 to +2, it is possible to assert that the skewness is moderate and the dataset is now optimal to be subjected for analysis.

CHAPTER 4

4.1 Feature Handling

A crucial stage in machine learning programs is feature handling. It entails choosing and extracting the most pertinent elements from the raw data in order to convert it into a comprehensible format. By utilizing fewer features, feature handling seeks to increase the precision and effectiveness of machine learning models.

Faster training times, decreased overfitting, and increased accuracy are all advantages of feature handling. Feature handling techniques include feature extraction, feature scaling, and feature selection. The process of feature selection entails choosing the dataset's most crucial attributes. Feature extraction entails building new features from ones that already exist. In order to ensure that all features are given equal weight in the data, feature scaling entails normalizing the data.

Methods used for Feature selection :

- Correlation
- Lasso regression
- Polynomial Regression

4.2 Correlation

Correlation is a statistical measure that describes the relationship between two variables. It is used in feature selection to identify the most important variables that are related to the target variable . The logic behind using correlation for feature selection is that good variables correlate highly with the target. Furthermore, variables should be correlated with the target but uncorrelated among themselves. If two variables are correlated, we can predict one from the other .

The benefits of correlation-based feature selection include decreased overfitting, reduced training time, and improved accuracy . Correlation-based feature selection can reduce model complexity, enhance learning efficiency, and even increase predictive power by reducing noise . Correlation-based feature selection is also less computationally demanding than other feature selection methods .

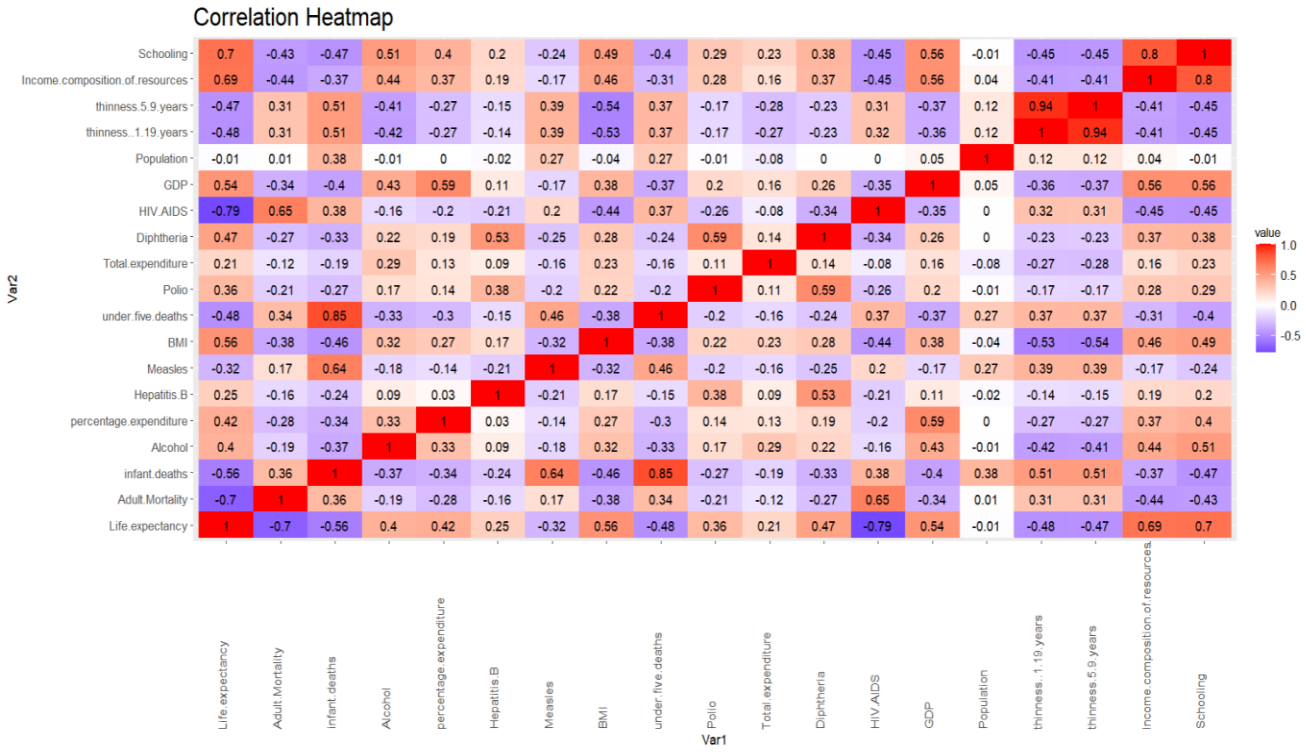

Fig 4.1 : Correlation Heat map

From the Correlation we can drop out few columns such as :

- ❖ Under five deaths
- ❖ Total expenditure
- ❖ Adult Morality
- ❖ Hepatis B

4.3 Lasso Regression

Lasso regression is a kind of linear regression in which some of the coefficients are shrunk to zero using L1 regularization to lessen the model's complexity. Because lasso regression can automatically choose the most crucial features and weed out the unimportant ones, it is used for feature selection.

Lasso regression has the advantage of making linear models more accurate and comprehensible. Less collinear feature subsets are preferred by lasso regression, which can enhance prediction metrics. Additionally, Lasso regression has the option to set the coefficients for uninteresting features to zero, which means that it automatically selects which features should be included and which ones shouldn't.

Lasso regression has several benefits over other feature selection techniques, including the capacity to manage multicollinearity and the ability to automatically pick features.

```
(Intercept)                     65.808456827
Adult.Mortality                 -0.012814688
infant.deaths                    .
Alcohol                          0.082674418
percentage.expenditure           0.233564199
 Hepatitis.B                     .
Measles                          .
BMI                              0.013396145
under.five.deaths               -0.937336375
Polio                            0.126472664
Total.expenditure                0.059211576
Diphtheria                       0.030325391
HIV.AIDS                        -8.829941192
GDP                              0.097495520
Population                       .
thinness..1.19.years            -0.000283757
thinness.5.9.years              -0.021850339
Income.composition.of.resources  6.418558508
Schooling                        0.378772575
```

4.4 Anova Feature selection

A statistical technique called ANOVA (Analysis of Variance) is used to compare two or more means. Anova feature selection is a subset of feature selection that chooses the most crucial features using ANOVA.

ANOVA feature selection has advantages such as lowering overfitting, raising accuracy, and shortening training times. The number of input factors can be decreased using ANOVA feature selection to only include those that are most helpful in predicting the target variable.

ANOVA has the ability to handle huge datasets with numerous features and its capacity to detect interactions between characteristics gives it an edge over other feature selection techniques.

```
[1] "The feature scores generated using ANOVA method are: "
[1] "Feature 1: HIV.AIDS - 4777.22665632284"
[1] "Feature 2: Schooling - 2889.74889319136"
[1] "Feature 3: Adult.Mortality - 2754.37051311568"
[1] "Feature 4: Income.composition.of.resources - 2681.19684779012"
[1] "Feature 5: under.five.deaths - 1528.09447948579"
[1] "Feature 6: infant.deaths - 1338.24216681776"
[1] "Feature 7: BMI - 1316.71733393934"
[1] "Feature 8: GDP - 1237.7757047412"
[1] "Feature 9: thinness..1.19.years - 863.891726115104"
[1] "Feature 10: Diphtheria - 851.194967910277"
[1] "Feature 11: thinness.5.9.years - 843.812169338012"
[1] "Feature 12: percentage.expenditure - 642.652842179943"
[1] "Feature 13: Alcohol - 560.300534606921"
[1] "Feature 14: Polio - 433.226702343313"
[1] "Feature 15: Measles - 324.594622322818"
[1] "Feature 16: Hepatitis.B - 192.531415987745"
[1] "Feature 17: Total.expenditure - 136.457263899143"
[1] "Feature 18: Population - 0.586613561295334"
```

4.5 Selected Features

From all the above methods we have selected 10 features, which are calculated as the ones with the best scores in all the above methods:

```
'data.frame':   2938 obs. of  11 variables:
 $ Life.expectancy                : num  65 59.9 59.9 59.5 59.2 58.8 58.6 58.1 57.5 57.3 ...
 $ Schooling                      : num  10.1 10 9.9 9.8 9.5 9.2 8.9 8.7 8.4 8.1 ...
 $ HIV.AIDS                       : num  0.562 0.562 0.562 0.562 0.562 ...
 $ Adult.Mortality                : num  263 271 268 272 275 279 281 287 295 295 ...
 $ infant.deaths                  : num  2.81 2.83 2.85 2.88 2.9 ...
 $ Alcohol                        : num  0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.03 0.02 0.03 ...
 $ Measles                        : num  5.83 4.71 4.55 7.27 7.41 ...
 $ Total.expenditure              : num  8.16 8.18 8.13 8.52 7.87 9.2 9.42 8.33 6.73 7.43 ...
 $ BMI                            : num  19.1 18.6 18.1 17.6 17.2 16.7 16.2 15.7 15.2 14.7 ...
 $ GDP                            : num  4.92 4.98 5.01 5.09 2.82 ...
 $ Income.composition.of.resources: num  0.479 0.476 0.47 0.463 0.454 0.448 0.434 0.433 0.415 0.405 ...
```

Fig 4.1 Selected Features

CHAPTER 5

5.1 Multiple Regression

This may be seen as an expanded form of a simple linear regression, where a link between a dependent and independent variable is established using a function. The same is carried out with multiple regression, but with more than one independent variable. One of the key benefits of this regression is that complex predictions may be handled since more independent variables are taken into account when forming the regression function. It will be possible to understand how each independent variable contributed to the dependent variables by looking at the created regression function. The issue with this approach is that the dependent and independent variables are expected to have a linear relationship. If this presumption turns out to be incorrect, the findings will be deceptive. Similar to this, another assumption, the normal distribution of residuals over the dataset, is taken into account and, if incorrect, will again lead to misinformation. It should be emphasized that this approach is outlier sensitive. Unless the assumptions taken into account are met, this method can be used effectively.

```
Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                     67.2125383  0.5509402 121.996  < 2e-16 ***
Schooling                        0.4371944  0.0435650  10.035  < 2e-16 ***
HIV.AIDS                        -9.6886793  0.2671545 -36.266  < 2e-16 ***
Adult.Mortality                 -0.0142390  0.0008612 -16.534  < 2e-16 ***
infant.deaths                   -0.8687122  0.1055072  -8.234 3.02e-16 ***
Alcohol                          0.1190488  0.0244746   4.864 1.23e-06 ***
Measles                         -0.0980182  0.0302741  -3.238  0.00122 **
Total.expenditure                0.1726707  0.0347810   4.965 7.40e-07 ***
BMI                              0.0156336  0.0049469   3.160  0.00160 **
GDP                              0.1729316  0.0316670   5.461 5.25e-08 ***
Income.composition.of.resources  7.7471197  0.6601621  11.735  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.774 on 2280 degrees of freedom
Multiple R-squared:  0.844,  Adjusted R-squared:  0.8433
F-statistic:  1234 on 10 and 2280 DF, p-value: < 2.2e-16
```

As we can infer from the above code The most significant variables are highlighted in the multiple regression model. The model's coefficients, standard error, t value, and P values are calculated, and the P values produced are less than the assumed conditions. i.e., 95% confidence interval. The coefficients vary according to their significance with regard to the dependent variable, and income composition of resources appears to have a very high relationship with respect to life expectancy, followed by schooling. Residual standard error, Degree of freedom, F-statistics are also estimated, and the Adjusted R-squared value generated here is also safe to proceed.

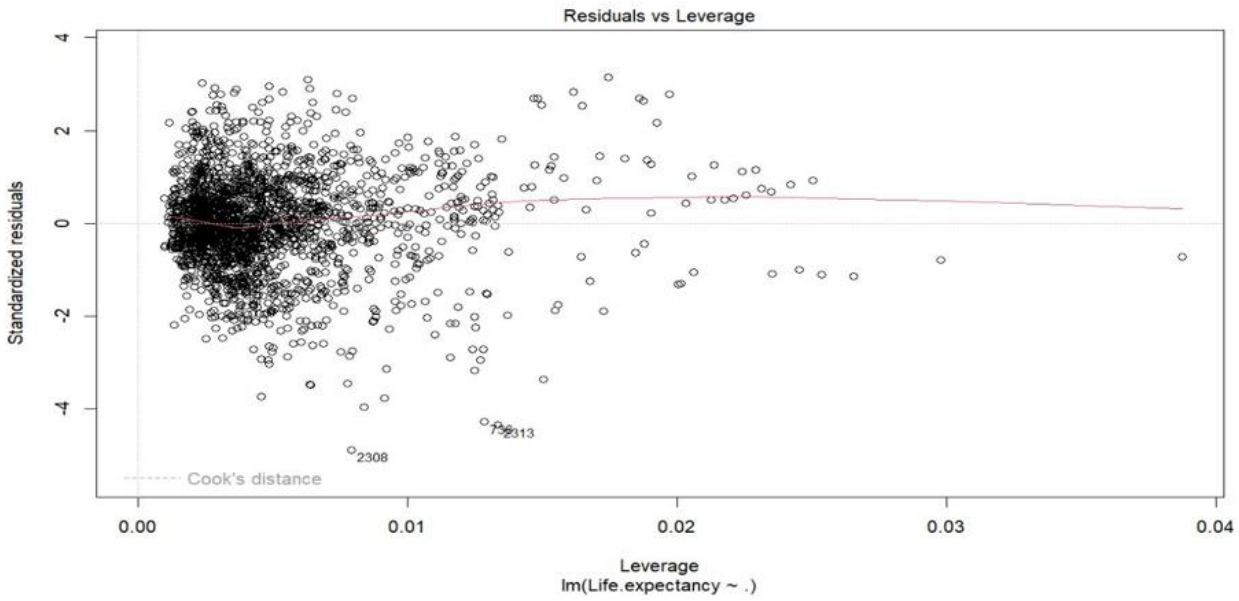

Figure 5.2: Multiple regression Residuals vs Leverage

Leverage values are displayed on the x-axis, and residuals are displayed on the y-axis. In this plot, the observations that may have a significant impact are situated in the upper- or lower-right corners of the plot and have high leverage and big residuals.

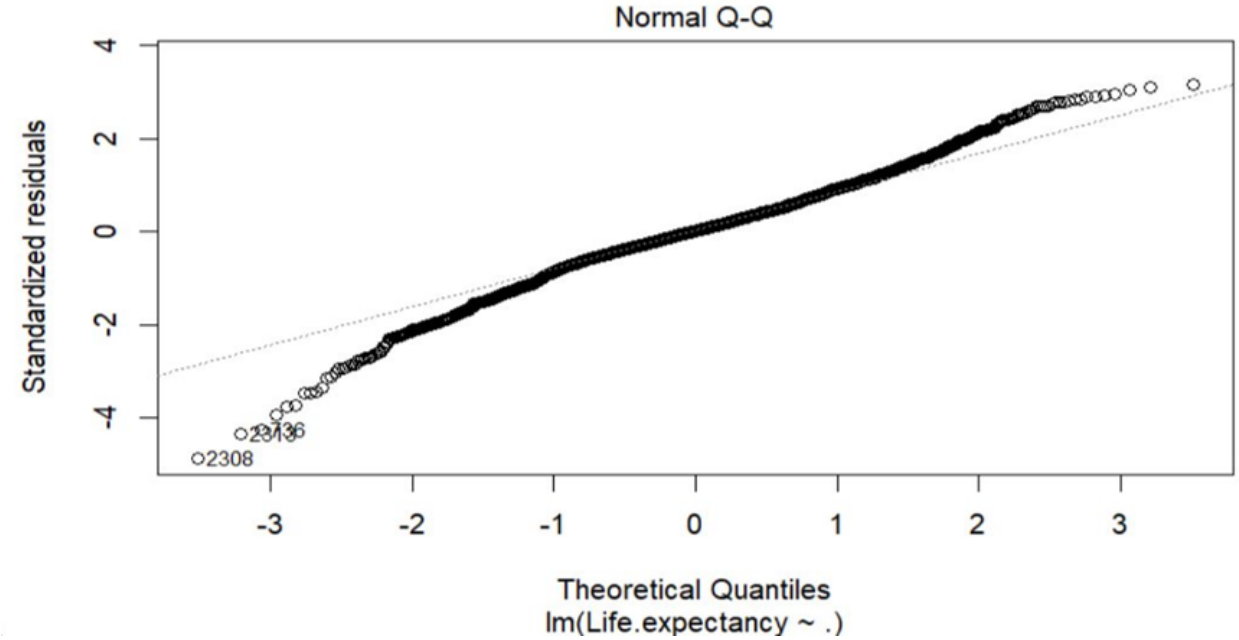

Figure 5.3: Multiple regression Q-Q plot

A graphical method for determining whether the residuals of a multiple regression model are normally distributed is the normal Q-Q (Quantile-Quantile) plot. A normal Q-Q plot compares the theoretical

quantiles of a normal distribution to the standardized residuals, i.e., calculated using the residuals divided by their estimated standard deviation. In Figure 5.3 The plotted points are almost lined up in a straight line, so we can say that the residuals are regularly distributed. Deviations from a straight line signify that the residuals are not typical.

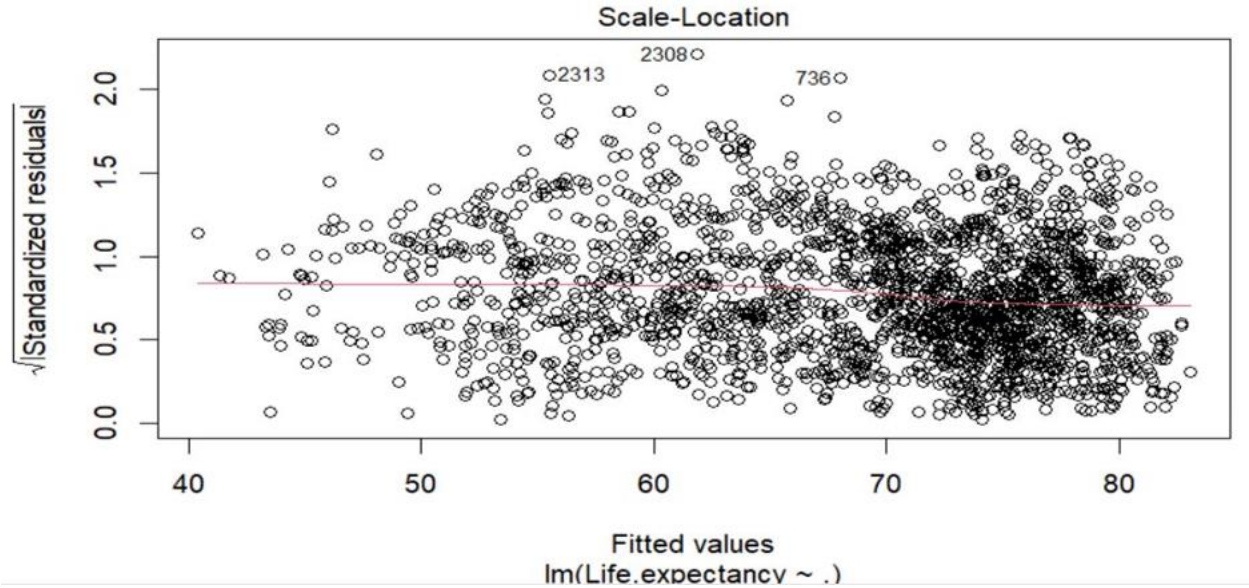

Figure 5.4 : root standard residuals vs Fitted values

The root standardized residuals vs fitted values (or anticipated values) plot in multiple regression is frequently used to examine the model's assumptions. It aids in the detection of any heteroscedasticity or nonlinear connections in the data. The model's residuals are divided by the square root of the residual variance to get the root standardized residuals. The fitted values are the predicted values of the dependent variable based on the independent variables in the model.

In Figure 5.4 We can see a random distribution of dots surrounding a horizontal line at zero on the plot, so the model's presumptions are true. If the plot shows a pattern, the model assumptions may not be true and the model may not be suitable for the data.

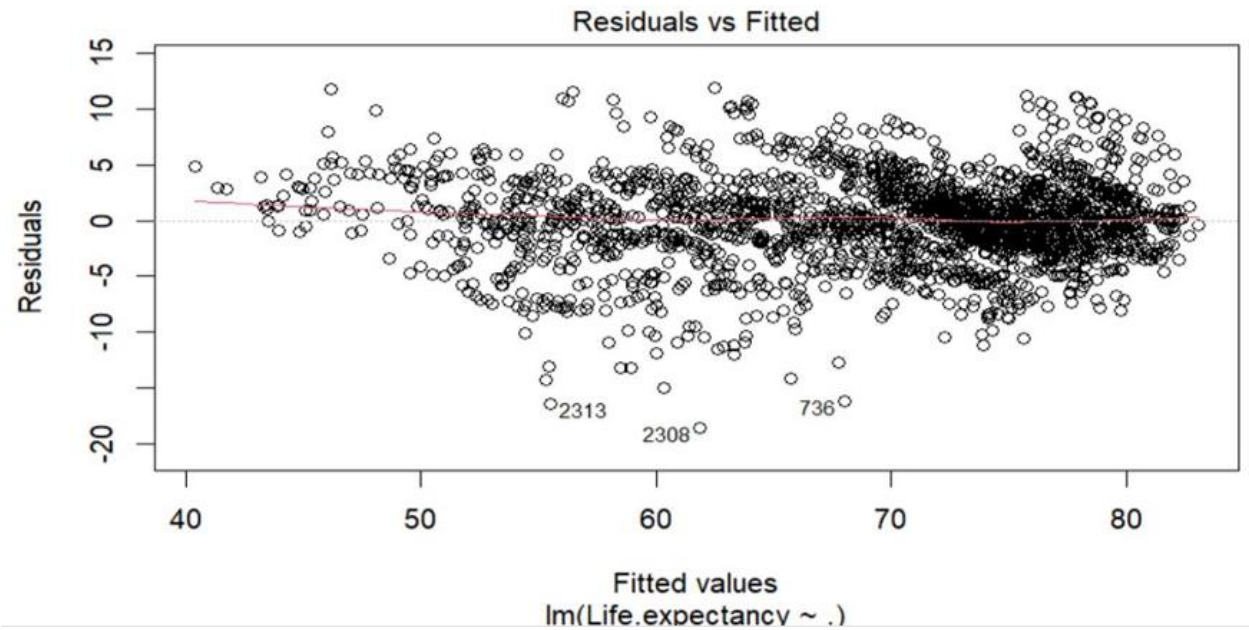

Figure 5.5: residuals vs Fitted values

Similar to the previous plot, In Figure 5.5 we are determining the accuracy of the model. The plots are densely populated near the horizontal zero line, which makes it more accurate model with the right set of Independent variables for predicting the Life expectancy (dependent variable).

5.2 Polynomial Regression

After the feature selection procedure where the number of variables are being reduced from 21 to 10, in order to increase the accuracy of the existing regression model, polynomial regression is being employed. This is the method in which a polynomial equation is being brought into play between the dependent and independent variables of a model. This in turn would establish a curvilinear relationship between them. Here the polynomial function is being characterized by the nth degree of the function. In this case, the order of the variables are being increased to the power of 3. The degree of the polynomial will decide the curve shape, where in this case it results in a curve with two bends. It is also possible to increase the order of the polynomial but it will lead to overfitting of the model. Working with an overfitted model is quite tedious as it will lose its ability to perform generalized prediction and the model will get much more complicated to work with. Third degree is being selected for this, based upon analyzing the AIC score for the cubic regression model. It is noted that as the degree is being increased more than 3, it does increase the AIC score as well. This should be avoided and so that, 3rd degree is finalized.

Based upon the following conclusion, the model is fitted with independent variables raised to the power of 3. Then the summary of the refitted mode is generated as follows

```
Call:
lm(formula = Life.expectancy ~ Schooling + Schooling^2 + Schooling^3 +
    HIV.AIDS + HIV.AIDS^2 + HIV.AIDS^3 + Adult.Mortality + Adult.Mortality^2 +
    Adult.Mortality^3 + infant.deaths + infant.deaths^2 + infant.deaths^3 +
    Alcohol + Alcohol^2 + Alcohol^3 + Measles + Measles^2 + Measles^3 +
    Total.expenditure + Total.expenditure^2 + Total.expenditure^3 +
    BMI + BMI^2 + BMI^3 + GDP + GDP^2 + GDP^3 + Income.composition.of.resources +
    Income.composition.of.resources^2 + Income.composition.of.resources^3,
    data = poly_df)

Residuals:
     Min       1Q   Median       3Q      Max
-20.0988  -1.9763   0.0518   2.2318  14.9758

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                     67.5842926  0.4867453 138.849  < 2e-16 ***
Schooling                        0.4169463  0.0389141  10.715  < 2e-16 ***
HIV.AIDS                        -9.6659788  0.2371008 -40.767  < 2e-16 ***
Adult.Mortality                 -0.0144805  0.0007674 -18.870  < 2e-16 ***
infant.deaths                   -0.9453013  0.0942088 -10.034  < 2e-16 ***
Alcohol                          0.1233136  0.0218041   5.656 1.70e-08 ***
Measles                         -0.0668474  0.0268716  -2.488   0.0129 *
Total.expenditure                0.1352579  0.0306180   4.418 1.03e-05 ***
BMI                              0.0177496  0.0044492   3.989 6.79e-05 ***
GDP                              0.1895129  0.0280834   6.748 1.80e-11 ***
Income.composition.of.resources  7.6730577  0.5933148  12.933  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.809 on 2927 degrees of freedom
Multiple R-squared:  0.8401, Adjusted R-squared:  0.8396
F-statistic:  1538 on 10 and 2927 DF,  p-value: < 2.2e-16
```

From the above code we can interpret that the Income composition and schooling majorly affects the life expectancy, further, to examine the fit of this model, model fit statistics such as Adjusted R-Squared and Multiple R-Squared values are being determined. The former statistic defines how good the fit is done between the regression model and the data. The latter is the same as the former, but it does not factor in the number of independent variables as in the former statistic. So, it is always better to measure both of the statistics to have an idea about the data fit. Adjusted R-Squared value is much more trustworthy as it also penalizes unwanted variables and emphasizes the model’s explanatory power. The range of values is between 0 to 1 and the closer the value of both the statistics to 1, the better fit it is. From the derived values, it is evident that the values are close to 1 which shows that the final polynomial regression model will be a good fit for the dataset, since all the coefficients are significant, which we can see from the R summary output as the P-value for the coefficients are lower than the 95 % confidence interval.

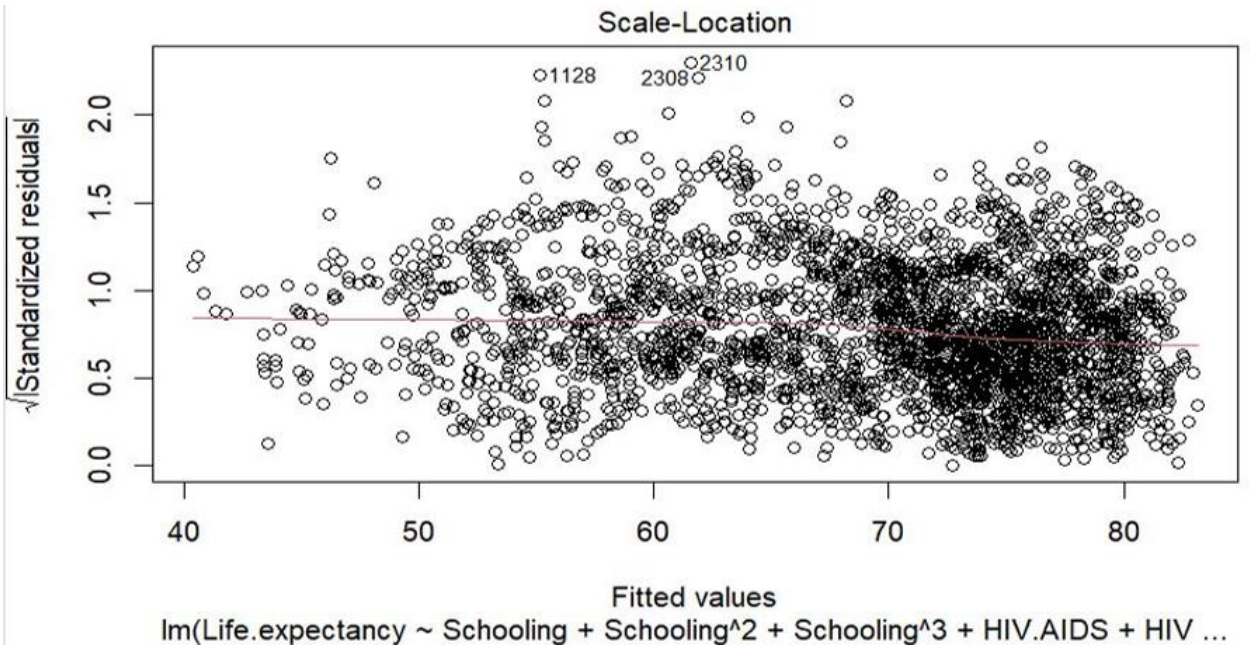

Figure 5.7: Scale location of fitted values in polynomial model

The scale location plot which is also known as spread-location plot has scattered the values almost evenly with constant variance and there are only few outliers which can be omitted without considering.

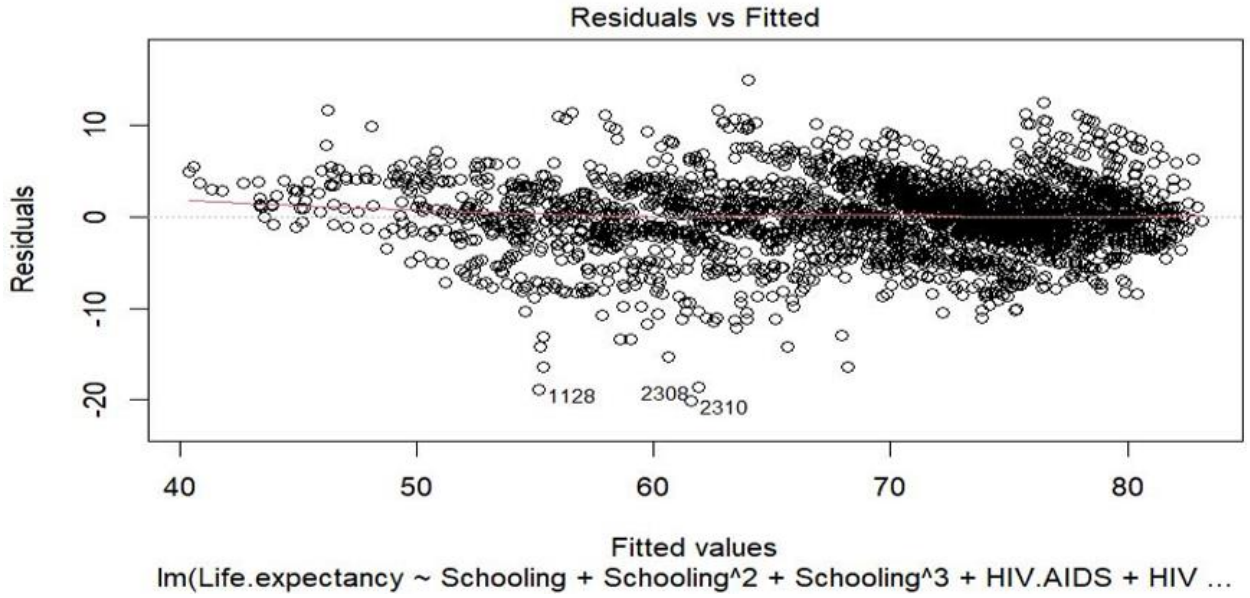

Figure 5.8: Polynomial regression residuals vs Fitted values

From the diagnostic plots of the final model, it is evident that the residuals are nearly zero when compared to the fitted values, which shows that the prediction efficiency of the model is optimal.

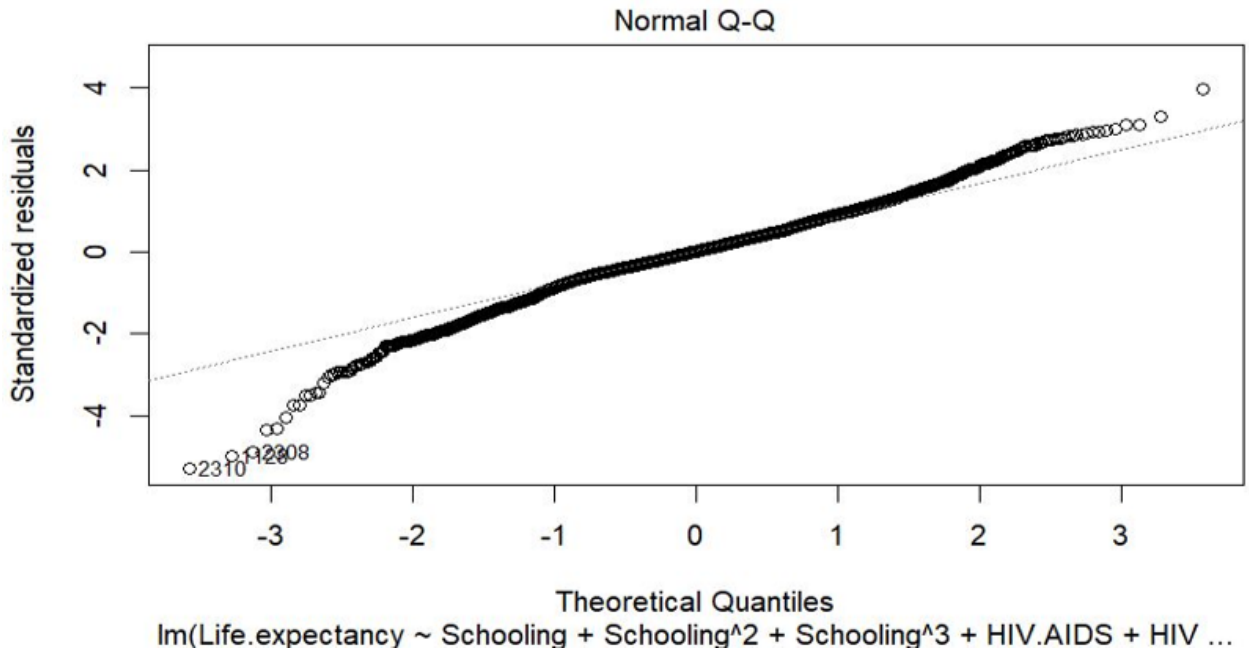

Figure 5.9: polynomial regression Q-Q plot

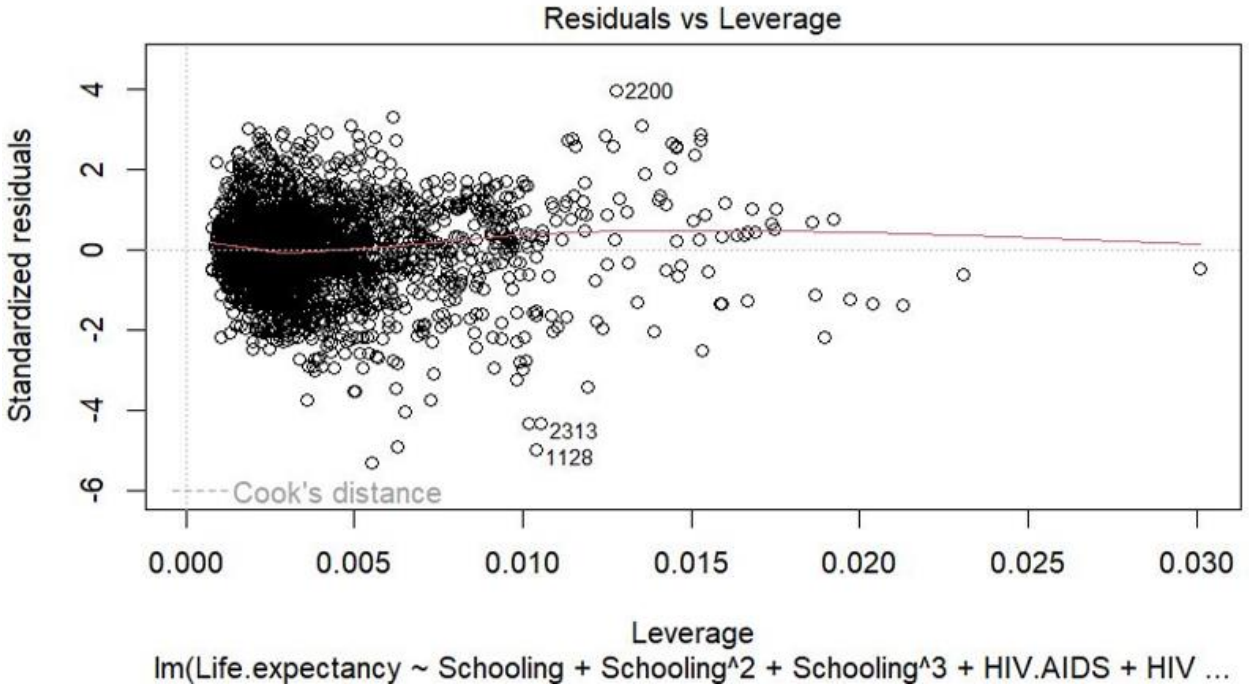

Figure 5.10: Residuals vs Leverage

And also from the Figure 5.10 it could be concluded that the residuals are distributed normally in the dataset. Yes, from the Cook's plot it is identified that there are several outliers, but it does not affect the

model’s performance. The scale location plot depicts nearly a constant spread of residuals across the fitted data.

5.3 AIC (Akaike Information Criterion):

AIC is a metric used in statistics and regression analysis to choose models. The balance between a model's quality of fit and its complexity in terms of the number of parameters is measured by this criterion. AIC penalizes models with more parameters; therefore, the best model has the lowest AIC score.

The AIC score is calculated as follows:

AIC = 2k - 2ln(L)

were,

K - number of estimated parameters in the model

L - maximum value of the likelihood function of the model.

```
> ply_aic = AIC(poly_reg)
> ply_aic
[1] 16209.43
> lm_aic = AIC(lm.fit_mod)
> lm_aic
[1] 16985.04
```

And to validate the polynomial regression model works better than the multiple regression model, the AIC scores are being calculated for both the models and it is inferred that the polynomial regression model works better than the multiple regression as it has the lowest AIC score.

CHAPTER 6

CONCLUSION

This report compares and examines the association between different health, social and economic parameters on life expectancy. Various visualization techniques have been applied to highlight the associations. Missing values are filled using Mean, Median and Interpolation Imputations. ANOVA, LASSO regression, Heat map is applied to identify best parameters for the regression model. Different regression techniques, like Multiple and third-degree polynomial regression are applied on this dataset to develop accurate predictive models, furthermore from model summary predictors like ‘Income composition of resources', ‘Schooling’, ‘GDP’ ranks as the most influencing predictors. ‘HIV.AIDS’ can be referred to as having the most negative influence, since it has the least coefficient value, we suspect that due to the least number of the population being affected by HIV.AIDS and considered as a biased result. Results from the regression model are validated using lowest AIC scores to finalize the model. We can conclude that Polynomial regression is best suited for predicting life expectancy based on the AIC scores. To increase the model accuracy, we suspect including subjective or qualitative data can be included as a potential predictor.

CHAPTER 7

REFRENCES

[1]. Rajarshi, K. (n.d.). Life Expectancy Data. Kaggle. https://www.kaggle.com/datasets/just249/lifeexpectancydatacsv

[2]. Wardle, P. (2021). Life-expectancy-linear-regression-project---Rstudio. GitHub. https://github.com/paddywardle/Life-expectancy-linear-regression-project---Rstudio

[3]. Enjoy Algorithms. (n.d.). Life Expectancy Prediction Using Machine Learning. Enjoy Algorithms. https://www.enjoyalgorithms.com/blog/life-expectancy-prediction-using-linear-model

[4]. Rajarshi, K. (n.d.). Life Expectancy (WHO). Kaggle. https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

[5]. Pandey, A., & Chikara, R. (2020). Analysis of Life Expectancy using various Regression Techniques. 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN). https://doi.org/10.1109/icacccn51052.2020.9362914

[6]. World Health Organization (WHO). (n.d.). Life Expectancy Data.

[7]. Gupta, S. (n.d.). Life Expectancy and GDP. Kaggle.

[8]. Beeksma, M., Verberne, S., van den Bosch, A. et al. Predicting life expectancy with a long short-term memory recurrent neural network using electronic medical records. BMC Med Inform Decis Mak 19, 36 (2019). https://doi.org/10.1186/s12911-019-0775-2

[9]. Bali, V., Aggarwal, D., & Singh, S. (2021). Life Expectancy: Prediction & Analysis using ML. In 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO) (pp. 1-6). IEEE. https://doi.org/10.1109/ICRITO51393.2021.9596123