# Life Expectancy

Advanced Calculus Group Project


**Report by:**

| | |
|---|---|
| **Prakruti Makwana** | **BS20DMU001** |
| **Manav Agarwal** | **BS20DMU020** |
| **Krish Sakharkar** | **BS20DMU012** |
| **Armaan Dhar** | **BS20DMU009** |

# Table of Contents


1. Introduction ..... 2
2. Dataset Exploration ..... 3
   2.1. Variable Explanation ..... 4
3. Visual Representation of data ..... 6
   3.1. Differences in life expectancy across the world ..... 6
   3.2. Graphs based on significant attributes ..... 7
4. Data Model ..... 10
   4.1. Population Data-model ..... 11
   4.2. Sample Data-Model ..... 14
5. Prediction of Future Values ..... 16
6. Conclusion ..... 18
7. Bibliography ..... 19

## 1. Introduction

Longevity is a key metabolic measure for human health. There is a narrow matrix of infant and child mortality, which focuses only on the mortality of young people, years of life holding death throughout the course of life. It tells us the average age of death in society.

The term “life expectancy” refers to the number of years a person can expect to live. Life expectancy is based on an estimate of the average age that members of a particular population group will be when they die.

Estimates suggest that in the pre-modern, impoverished world, life expectancy was about 30 years in all regions of the earth. Over the last few decades, life expectancy has increased dramatically around the globe. In 1841, a baby girl was expected to live to just 42 years of age, a boy to 40. In 2016, a baby girl could expect to reach 83; a boy, 79. Since the 1900s the average life expectancy on earth has more than doubled and is now over 70 years.

Life expectancy has grown rapidly since the Age of Enlightenment (European movement in the late 17$^{th}$ century). Since the beginning of the 19th century, life expectancy began to increase in the industrialized world while it remained low around the world. The decline of child mortality was important for the increase of life expectancy, but as we explain further, it was certainly not only about falling child mortality; life expectancy increased at all ages due to a range of factors.

Such improvements in life expectancy were a landmark sign of progress. It was the first time in human history that we achieved sustained improvements in health worldwide. After millennia of stagnation in terrible health conditions the seal was finally broken. (ourworldindata, 2019)

Good health in rich countries and persistently poor health in those countries that remained poor. In recent decades, this global inequality has diminished. No country in the world has a lower life expectancy than the countries with the highest life expectancy in the 1800s. Many countries in the recent past have been suffering from malnutrition are catching up rapidly.

## 2. Dataset Exploration

The Global Health Observatory (GHO) data repository under the World Health Organization (WHO) keeps track of the health status as well as many other related factors for all countries. The datasets are made available to the public for the purpose of healthcare analysis. For the dataset of life expectancy, health factors for 193 countries have been collected from the WHO data repository website via Kaggle and its corresponding economic data was collected from United Nations website.

Among all categories of health-related factors, only those critical factors were chosen which are more representative. It has been observed that in the past 15 years, there has been a huge development in the health sector resulting in an improvement in human mortality rates, especially in the developing nations in comparison to the past 30 years. Therefore, in this project we have considered data from the year 2000-2015 for 193 countries for further analysis. The individual data files have been merged into a single dataset.

The final merged file (final dataset) consists of 22 Columns and 2938 rows. Out of these 22 columns, 17 columns are considered as variables to predict the life expectancy. (Rajarshi, 2018)

## 2.1. Variable Explanation

There are 22 columns in the dataset. The following fields are a part of the dataset:

1. **Country**: Name of the countries being recorded for their life expectancy data.
2. **Year**: The years for the data given.
3. **Status**: Developed or developing status for each country.
4. **Life expectancy**: Life expectancy in age
5. **Adult Mortality**: Adult Mortality Rates of both sexes (Number of deaths per 1000 people aging between 15 to 60).
6. **Infant Deaths**: Number of Infant Deaths per 1000 population
7. **Alcohol**: Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol)
8. **Percentage expenditure**: Expenditure on health as a percentage of Gross Domestic Product per capita (%)
9. **Hepatitis B**: Hepatitis B immunization coverage among 1-year-olds (%)
10. **Measles**: number of reported cases per 1000 population
11. **BMI**: Average Body Mass Index of entire population
12. **Under-five deaths**: Number of under-five deaths per 1000 population
13. **Polio**: immunization coverage among 1-year-olds (%)
14. **Total Expenditure**: General government expenditure on health as a percentage of total government expenditure (%)
15. **Diphtheria**: Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)
16. **HIV/AIDS**: Deaths per 1000 live births HIV/AIDS (0-4 years)
17. **GDP:** Gross Domestic Product per capita (in USD)
18. **Population**: Population of the country

19. **Thinness** 10-19 years: Prevalence of thinness among children and adolescents for Age 10 to 19 (%)

20. **Thinness 5-9 years**: Prevalence of thinness among children for Age 5 to 9(%)

21. **Income composition of resources:** Human Development Index in terms of income composition of resources (index ranging from 0 to 1)

22. **Schooling:** Number of years of Schooling(years)

## 3. Visual Representation of data

This section covers visual representation of the dataset and shows the trendline of some of the most crucial factors affecting life expectancy

### 3.1. Differences in life expectancy across the world

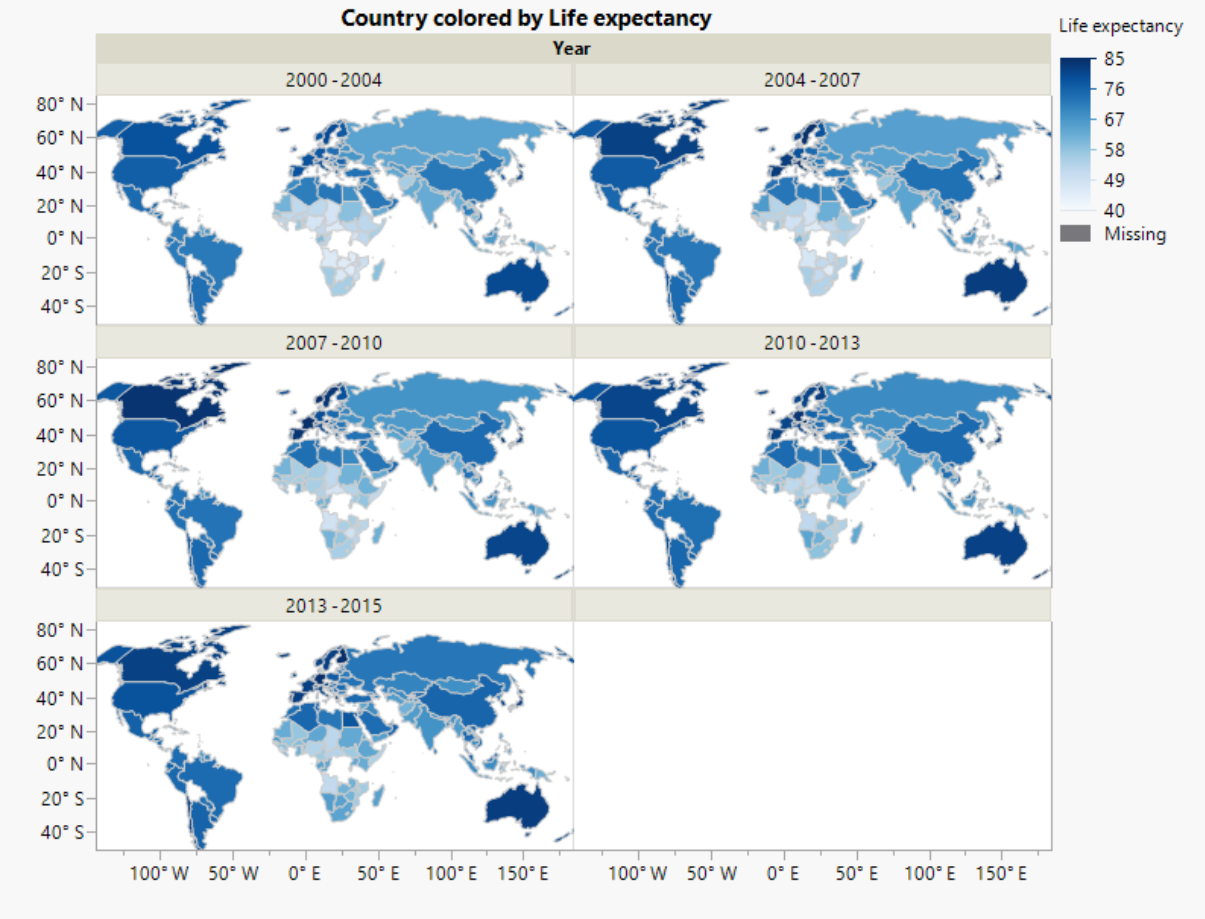


***Fig 3.1: Life Expectancy Country wise***

The graph shows the life expectancy trend from 2000 – 2015 all around the world. As seen in fig 3.1, during 2000-2004 Africa had the lowest life expectancy as a continent. As time passed by, with all the technological developments and advancement in healthcare facilities, the life expectancy of the entire world has increased. North America and Australia have the highest life expectancy since the beginning and have stayed the same.

## 3.2. Graphs based on significant attributes

The following 4 graphs have been plotted to show the trend of attributes such as HIV/AIDS, Adult Mortality, etc. over the years. In section 4.2 of Sample Data Model, it has been shown that these 4 regressors (Adult mortality, HIV, Schooling, and Income Composition of Resources) have a significant impact on life expectancy.

**a) HIV/ AIDS vs Years:**

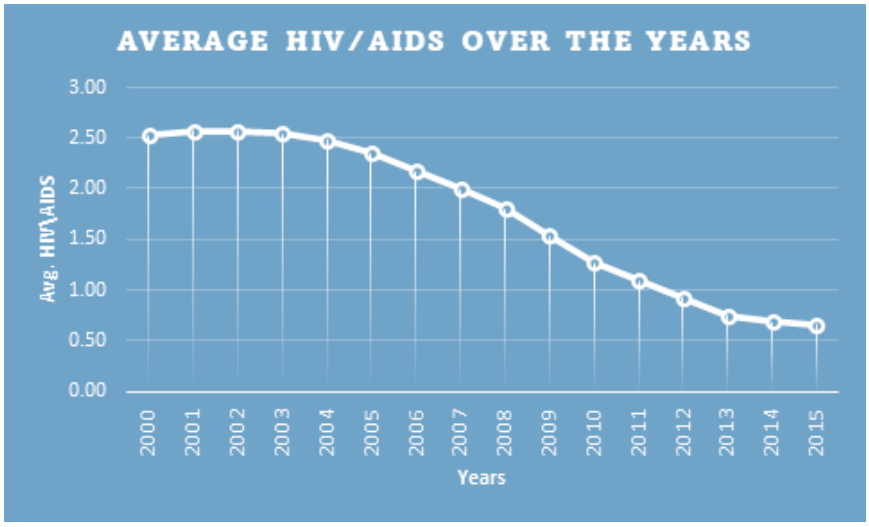


***Fig 3.2***

HIV is a virus that attacks the body's immune system. If HIV is not treated, it can lead to AIDS and since there is no cure for HIV which makes it even deadlier. We see a decline based on the above graph in the number of deaths per 1000 live births due to HIV as it has dropped from 2.5 to approximate 0.6 between the years 2000 to 2015.

**b) Adult mortality vs Years:**

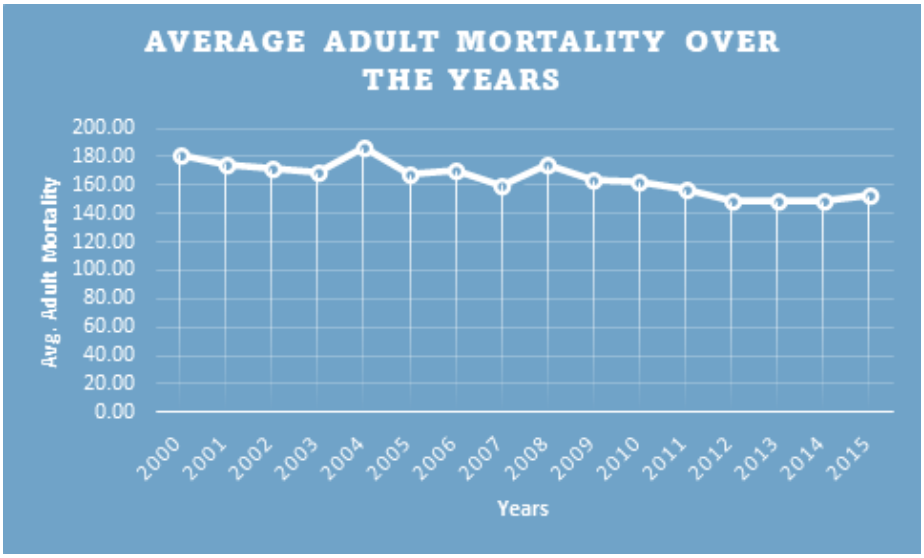


***Fig 3.3: Average Adult Mortality (Number of deaths per 1000 people aged between 15 to 60)***

Figure 3.3 depicts Adult Morality over the years 2000-2015. As seen above, the average adult morality over the years has fluctuated a lot and we can't see any evident trends in the graph. Initially, the mortality of adults decreased (from 2000-2003) and then a sudden hike can be seen in the year 2004 as the peaks in the graph. After that it again moved towards X axis but rose again in 2008 reaching approximately 180 deaths and after then it has been coming down.

### c) Income Composition vs Years

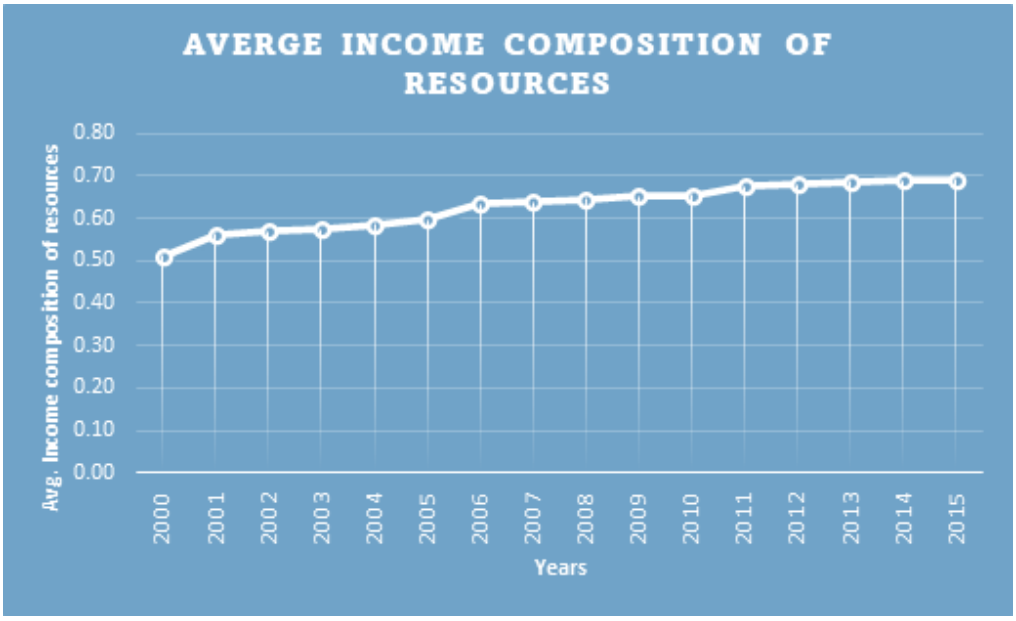


***Fig 3.4: Average Income Composition of Resources Human Development Index in terms of income composition of resources (index ranging from 0 to 1)***

The above graph depicts the average “Income Composition of Resources” over the years 2000-2015. From the initial point, we can observe that the trendline moved away from X axis and reached a point beyond which it remained constant. The lowest point in the dataset was in 2005(initial point) when the value of Income Composition of Resources was 11. The highest value of Income Composition of Resources was the years 2014 and 2015 where the value was 13.

### d) Schooling vs Years:

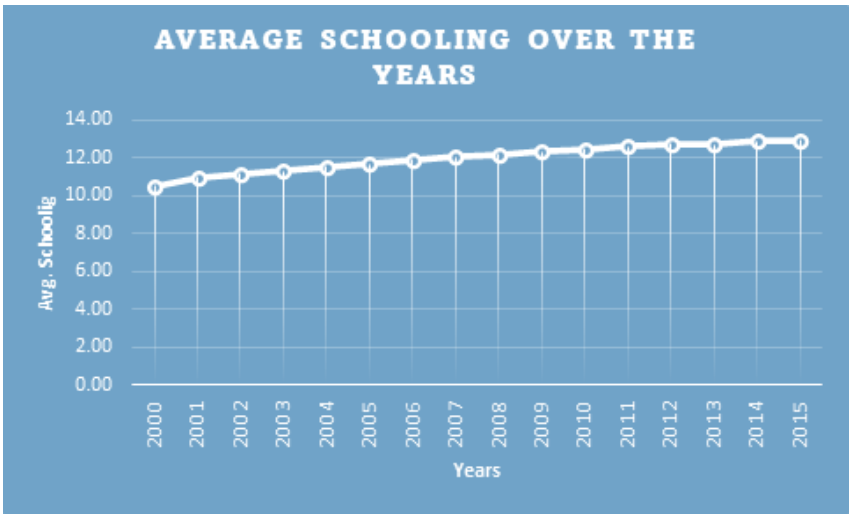


***Fig 3.5: Average Schooling***

The above graph depicts the schooling trend over the years 2000-2015. As seen above, schooling had started off low with 10.5 but now with realizing the importance of schooling it is going up.

## 4. Data Model

Linear regression models are the sum of the products of coefficient parameters and factors. In addition, linear models for continuous responses are usually specified with a normally distributed error term. The parameters are chosen such that their values minimize the sum of squared residuals. This technique is called estimation by least squares. The most important application of least square is data fitting.

Using the linear regression model, we have fitted a mechanism to determine some of its parameters (coefficients). These parameters are determined by the method of least squares, which finds the parameter values that minimize the sum of squared distances from each point to the line of fit.
The line of best fit determined from the least squares' method has a prediction equation that tells the relationship between the data points.

### 4.1. Multiple Regression

Multiple regression is used to estimate the relationship between two or more independent variables and one dependent variable. The coefficient of determination (R-squared) measures how much of the variation in outcome can be explained.$R^2$ always increases as more predictors are added to the model.$R^2$ lies between 0 and 1.

**Formula and Calculation of Multiple Linear Regression**

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon$$

**where, for $i = n$ observations:**

$y_i$ = dependent variable

$x_i$ = explanatory variables

$\beta_0$ = y-intercept (constant term)

$\beta_p$ = slope coefficients for each explanatory variable

$\epsilon$ = the model's error term (also known as the residuals)

## 4.2. Population Data-model

### a) Actual By Predicted Plot:

The Actual by Predicted plot is a type of residual plot that shows the actual ratio values plotted against the predicted ratio values. It enables us to see unexplained patterns in the data. There might be some improvement with the least square fit, but there still appears to be a pattern in the residuals. (John Sall, 2017)

For a good fit, the points should be close to the fitted line and the same is observed in fig 4.1.

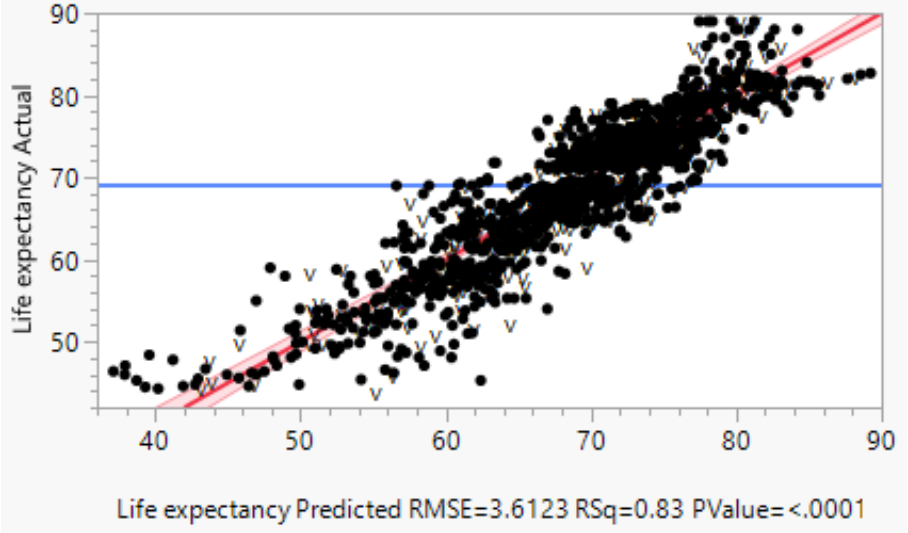


***Fig 4.1: Actual by prediction plot***

### b) Effect Summary:

Effect shows how each regressor contributes to the fit. It sorts the effects in descending order of significance. By observing Table 4.1, we can see that there are 9 variables having a P Value of <0.05. This indicates that these 10 variables have a significant impact on the regression equation. Since the P Value of the remaining variables is much greater than 0.05, they have negligible effect on the regression equation, therefore, they can be excluded in the sample model.

| Source | Log Worth | P Value |
|---|---|---|
| HIV/AIDS | 87.273 | 0.00000 |
| Adult Mortality | 51.187 | 0.00000 |
| Schooling | 39.996 | 0.00000 |
| Income composition of resources | 23.247 | 0.00000 |
| Under-five deaths | 15.832 | 0.00000 |

| | | |
|---|---|---|
| Infant deaths | 14.775 | 0.00000 |
| Percentage expenditure | 10.389 | 0.00000 |
| BMI | 4.138 | 0.00007 |
| Diphtheria | 2.005 | 0.00990 |
| Alcohol | 1.058 | 0.08747 |
| Total expenditure | 1.044 | 0.09044 |
| Hepatitis B | 0.976 | 0.10575 |
| Polio | 0.912 | 0.12238 |
| Thinness 5-9 years | 0.883 | 0.13102 |
| Measles | 0.514 | 0.30597 |
| Population | 0.276 | 0.53005 |
| Thinness 1-19 years | 0.153 | 0.70372 |

***Table 4.1***

### c) Summary of Fit:

This table gives an overall summary of how well the model fits. For a good model, the RSquare value should be close to 1 as it describes the variation explained by the model. Since our model has a RSquare value of 83.16%, it can be considered a good fit.

| | |
|---|---|
| R Square | 0.83163 |
| R Square Adj | 0.829456 |
| Root Mean Square Error | 3.612332 |
| Mean of Response | 69.22639 |
| Observations (or Sum Weights) | 1334 |

***Table 4.2***

### d) Parameter Estimates:

Both the graph and the Summary of Fit indicate a strong linear relationship between the two parts of these sonatas. Table 4.7 shows the parameter estimates panel from the results.

| Term | Estimate | Std Error | t Ratio | Prob>\|t\| |
|---|---|---|---|---|
| Intercept | 53.08014 | 0.821612 | 64.60 | <.0001* |
| Adult Mortality | -0.016706 | 0.001054 | -15.85 | <.0001* |
| Infant deaths | 0.0989545 | 0.012275 | 8.06 | <.0001* |
| Alcohol | -0.059694 | 0.034906 | -1.71 | 0.0875 |
| Percentage expenditure | 0.0004575 | 6.873e-5 | 6.66 | <.0001* |
| Hepatitis B | -0.00838 | 0.005177 | -1.62 | 0.1058 |
| Measles | -0.000012 | 1.166e-5 | -1.02 | 0.3060 |
| BMI | 0.0268543 | 0.006748 | 3.98 | <.0001* |
| Under-five deaths | -0.074058 | 0.00885 | -8.37 | <.0001* |
| Polio | 0.0089489 | 0.005789 | 1.55 | 0.1224 |
| Total expenditure | 0.077788 | 0.045911 | 1.69 | 0.0904 |
| Diphtheria | 0.0172734 | 0.006687 | 2.58 | 0.0099* |

| | | | | |
|---|---|---|---|---|
| HIV/AIDS | -0.425205 | 0.019793 | -21.48 | <.0001* |
| Population | -1.321e-9 | 2.103e-9 | -0.63 | 0.5301 |
| Thinness 1-19 years | 0.0232825 | 0.061208 | 0.38 | 0.7037 |
| Thinness 5-9 years | -0.090625 | 0.059976 | -1.51 | 0.1310 |
| Income composition of resources | 9.5230704 | 0.92472 | 10.30 | <.0001* |
| Schooling | 0.9267275 | 0.067004 | 13.83 | <.0001* |

***Table 4.3***

### e) Cross validation:

Model validation is the process of using a separate set of data to determine the necessary model complexity. The data used to build the model is called the "Training Data" and the data used to check the reasonableness of the data model is called "Validation Data." We apply the models built with the training data to the validation data and see how well the different models predict these new values. In our model, we considered 80% of the rows as "Training Data" and 20% of the rows as "Validation Data". The data model gave a R Square value of 83.16% on the training set and a value of 83.70% on the validation set. From this we can conclude that our model can be used to effectively predict the life expectancy.

| Source | R Square | RASE | Freq |
|---|---|---|---|
| Training Set | 0.8316 | 3.5879 | 1334 |
| Validation Set | 0.8370 | 3.5989 | 323 |

***Table 4.4***

### f) Prediction Expression:

Based on the "Fit least Squares" method, the following equation was formulated and used to predict the value of Life Expectancy.

***Life Expectancy*** = 53.08 + (-0.01* Adult Mortality) +
(+0.09 * Infant deaths) + (-0.05 * Alcohol) +
(+0.0004 * Percentage expenditure) + (-0.008 * Hepatitis B) +
(-0.00001 * Measles) + (+0.026 * BMI) +
(-0.07 * Under-five deaths) +(+0.008* Polio) +
(+0.07 * Total expenditure) +(+0.01 * Diphtheria) +
(-1.322e-9 * Population) + (+0.02 * Thinness 10-19 years) +
(-0.09 * Thinness 5-9 years) + (+9.52 * Income composition of resource) +
(-0.42 * HIV/AIDS) + (+0.92 * Schooling)

## 4.3. Sample Data-Model

### a) Actual by Predicted Plot:

For a good fit, the points in the Actual by predicted plot should be close to the fitted line and the same is observed in fig 4.2. This verifies that the least square fit model is the most accurate fit.

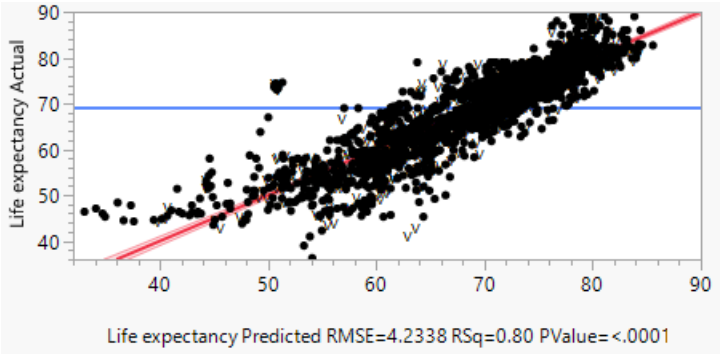


***Fig 4.2: Actual by Predicted Plot***

### b) Effect Summary:

As we saw in table 4.1, there were several significant terms among 17 attributes. Out of these, the four regressors shown in table 4.5 were the most significant and have been selected for the prediction expression owing to their low p-value.

| Source | Log Worth | p- Value |
|---|---|---|
| HIV/AIDS | 117.504 | 0.00001 |
| Schooling | 112.575 | 0.00001 |
| Adult Mortality | 80.257 | 0.00001 |
| Income composition of resources | 41.463 | 0.00001 |

***Table 4.5***

### c) Summary of Fit:

The Summary of Fit table shows an $R^2$ of 79.57%, which makes the regression model look like a good fit.

| | |
|---|---|
| R Square | 0.795762 |
| R Square Adj | 0.795391 |
| Root Mean Square Error | 4.233838 |
| Mean of Response | 69.25853 |
| Observations (or Sum Weights) | 2211 |

***Table 4.6***

### d) Parameter Estimates:

Both the graph and the Summary of Fit indicate a strong linear relationship between the two parts of these sonatas. Table 4.7 shows the parameter estimates panel from the results.

| Term | Estimate | Std Error | t Ratio | Prob>\|t\| |
|---|---|---|---|---|
| Intercept | 53.569691 | 0.459179 | 116.66 | <.0001* |
| Adult Mortality | -0.018869 | 0.000949 | -19.87 | <.0001* |
| Income composition of resources | 10.194728 | 0.733251 | 13.90 | <.0001* |
| HIV/AIDS | -0.49636 | 0.020183 | -24.59 | <.0001* |
| Schooling | 1.1086515 | 0.04619 | 24.00 | <.0001* |

***Table 4.7***

### e) Cross Validation:

Cross validation ensures that the model is robust. A portion of the data is held back, and the holdout sample is used to test the model. In this case, 80% of the dataset is used for training the model and 20% is part of the validation set which tests the model. This is different from the usual method of model testing, which uses the entire dataset to test the model.

| Source | R Square | RASE | Freq |
|---|---|---|---|
| Training Set | 0.7958 | 4.2290 | 2211 |
| Validation Set | 0.7768 | 4.4199 | 557 |

***Table 4.8***

### f) Prediction Expression:

Based on the least square fit model, the following expression is used to predict the value of Life Expectancy.

***Life Expectancy*** = 53.56 + (-0.01 * Adult Mortality) + (10.19 * Income composition of resource) + (-0.49 * HIV/AIDS) + (+1.10 * Schooling)

## 5. Prediction of Future Values

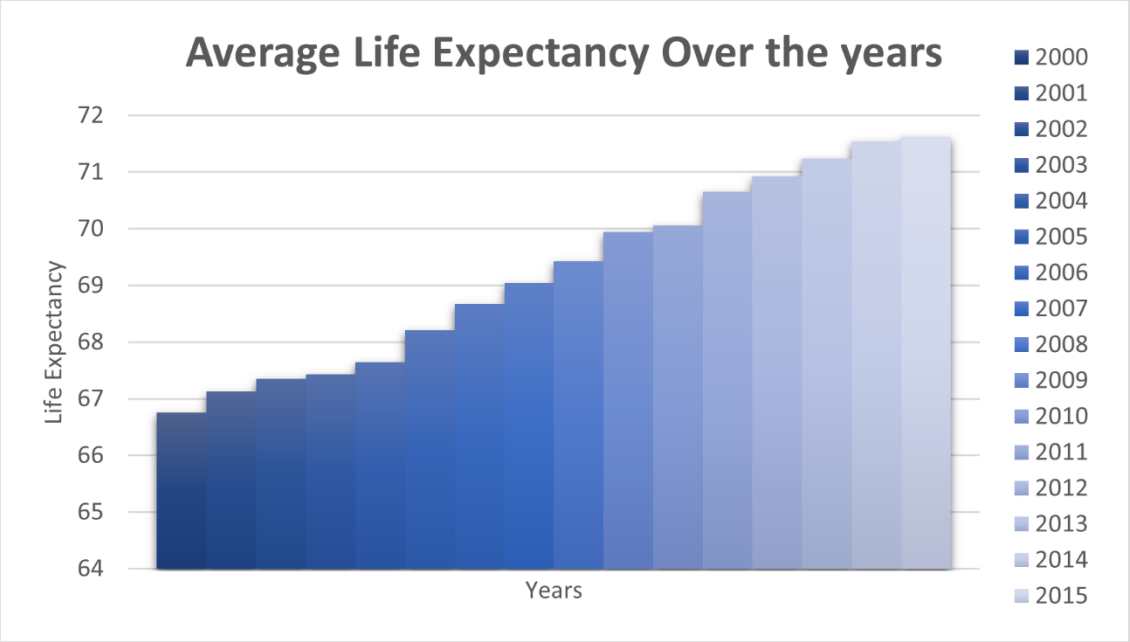


***Fig 5.1: Life expectancy over the years***

As we can already establish that over the years life expectancy has increased gradually with an exponential growth. Exponential growth is a growth whose rate becomes more rapid in proportion to the growing total number. Exponential growth is given by the formula:

$$f(x) = a(1 + r)^2$$

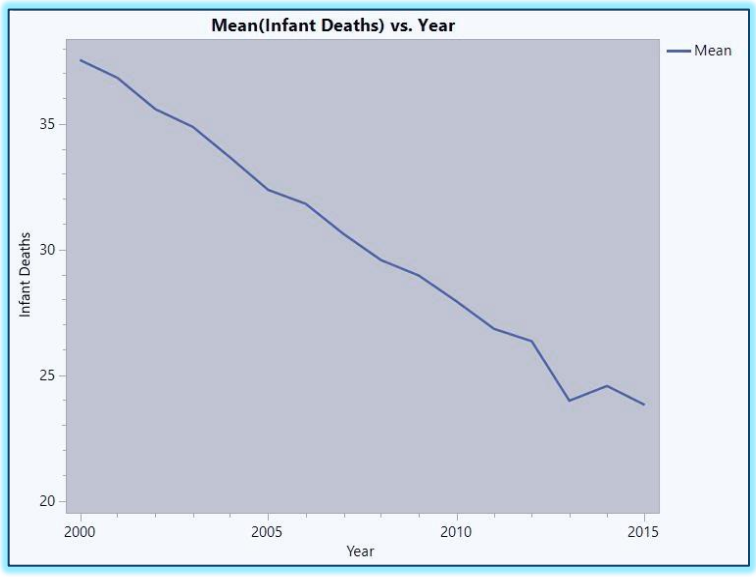


***Fig 5.2: Mean Infant Deaths over the years***

As in figure 5.2 we can deduce that the mean infant deaths have been dropped by 40% in the past 15 years with an increasing rate, from 38 deaths to 23. The data given is also perfectly accurate to determine the future readings as the r-square value is 0.8370 which is really close to 1 and can considered as an exactly accurate. From figure 3.3 we could analyse that the mean adult mortality rate decreases by about 16% in a span of 15 years, from 182 to 154. Understanding that adult death rates are decreasing rapidly, hence boosting the life expectancy of the world by a great margin. Life expectancy at a predicted rate in 69.3 years by calculating from the prediction expression calculated from the model.

**Life Expectancy = 53.569691471 +(-0.018869196 * Adult Mortality (164.796448)) + (10.194728117 * Income composition of resource (0.62755106)) +(-0.496360403 * HIV/AIDS (1.74210347)) +(+1.1086515054 * Schooling (11.9927928))**

**Hence, Life expectancy= 69.3 years**

However, the actual prediction is above 71 years, which makes the actual expectancy greater than the predicted expectancy. Concluding that life expectancy will increase over the period and the data is extremely accurate.

**Short term prediction:** To conclude, the next 16% decrease of adult mortality would happen much sooner and the estimate for that to happen would be by the next 10 years. Same as for the mean infant deaths, it will decrease by the same margin quicker. Life expectancy will boost higher thus maintaining the rules of exponential growth in the short term.

**Long term prediction:** However, in the long term the life expectancy would come to a halt as 2 major factors for determining it would reach its peak. The mean infant deaths will stop dropping down by a huge margin, but it wouldn't increase by a big margin as well, hence being stable. The same can be said for the adult mortality rate as well. Because the value will be dropped to such a level, that it wouldn't be able decrease further as it would reach towards the minimum value for each factor. Making it to stop the growth or the decay for the factors and hence, keeping the life expectancy constant by proving the logistic growth method.

## 6. Conclusion

Life expectancy estimates in this case describe averages and provide a complementary view to help us understand how the inequality of life lengths has changed over time. It is a measure of premature death, and it shows significant health effects worldwide.

It is important to note that the prediction equation formulated here is limited to its use since there are several other social, economic, geographical factors and many other factors which can be used to determine Life Expectancy and have not been taken into consideration here.

This paper gives only part of a greater work, by focusing on the prediction model which has been based on least square regression estimates to predict the life expectancy.

## 7. Bibliography

John Sall, A. L. (2017). Start Statistics: A Guide to Statistics and Data Analysis Using . In A. L. John Sall, *JMP Start Statictics.* NC, USA.

*ourworldindata*. (2019). Retrieved from ourworldindata: https://ourworldindata.org/life-expectancy

Rajarshi, K. (2018). *Statistical Analysis on factors influencing Life Expectancy*. Retrieved from Kaggle: https://www.kaggle.com/kumarajarshi/life-expectancy-who