**Life Expectancy Prediction**

By: Thanojkumar Guntupalli and Nataliya Volkova

1. **Introduction**

There are many definitions of life expectancy. For this project, we would use the official definition by World Health Organization. So, life expectancy, which is also called life expectancy at birth, is the average number of years that a newborn could expect to live, if he or she were to pass through life exposed to the sex and age-specific death rates prevailing at the time of his or her birth, for a specific year, in a given country, territory, or geographic area.

This indicator is so important for describing population conditions that, together with the education index and the Gross Domestic Product (GDP) index, it forms the Human Development Index used by the United Nations Development Programme (UNDP). There is no better indicator of a country's social development than having a long and healthy life.

The significance of studying this subject lies in the understanding that life expectancy is the key metric to assess the population’s health, state of welfare, economic conditions, and demographic situation in a population.

In this project, our objectives were to find statistically significant predictors by building a multiple linear regression model, to check the difference in life expectancy between developing and developed countries, and, if there is a difference, we wanted to see which predictors are significant for life expectancy in developed and developing countries.

1. **Dataset**

Originally the data was collected by WHO from 2000 to 2015. It includes data on life expectancy for 193 countries. The dataset consists of 2939 observations and 22 attributes. The table below shows the variables’ names, types, and descriptions.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Variable** | **Type** | **Description** |
| 1 | Country | Character | Country name |
| 2 | Year | Character | Year of the data |
| 3 | Status | Character | Developed or developing |
| 4 | Life Expectancy | Numeric | Life expectancy in ages |
| 5 | Adult Mortality | Numeric | Probability of dying between 15 and 60 years per 1000 population |
| 6 | Infant deaths | Numeric | Number of Infant Deaths per 1000 population |
| 7 | Alcohol | Numeric | Alcohol, recorded per capita (15+) consumption (in liters) |
| 8 | Percentage expenditure | Numeric | Expenditure on health as a percentage of Gross Domestic Product per capita |
| 9 | Hepatitis B | Numeric | Hepatitis B (HepB) immunization coverage among 1-year-olds (%) |
| 10 | Measles | Numeric | Measles - number of reported cases per 1000 population |
| 11 | BMI | Numeric | Average Body Mass Index of entire population |
| 12 | Under five deaths | Numeric | Number of under-five deaths per 1000 population |
| 13 | Polio | Numeric | Polio (Pol3) immunization coverage among 1-year-olds (%) |
| 14 | Total expenditure | Numeric | General government expenditure on health as a percentage of total government expenditure (%) |
| 15 | Diphtheria | Numeric | Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%) |
| 16 | HIV/AIDS | Numeric | Deaths per 1000 live births HIV/AIDS (0-4 years) |
| 17 | GDP | Numeric | Gross Domestic Product per capita (in USD) |
| 18 | Population | Numeric | Population of the country |
| 19 | Thinness 1-19 years | Numeric | Prevalence of thinness among children and adolescents for Age 10 to 19 (% ) |
| 20 | Thinness 5-9 years | Numeric | Prevalence of thinness among children for Age 5 to 9(%) |
| 21 | Income composition of resources | Numeric | Human Development Index in terms of income composition of resources (index ranging from 0 to 1) |
| 22 | Schooling | Numeric | Number of years of Schooling(years) |

We used life expectancy as a response variable and the rest as predictors. To better understand the dataset, we divided predictors into several categories:

* Immunization-related factors (Hepatitis B, Measles, Polio, Diphtheria, HIV/AIDS);
* Mortality factors (Adult Mortality, Infant Deaths, Under five deaths);
* Economic factors (GDP, Total Expenditure on health care, Income Composition of resources, Percentage expenditure);
* Social factors (Schooling, Alcohol, Population, Status);
* Others (Country, Year, BMI, Thinness 1-19 years, Thinness 5-9 years).

This classification is useful when we assess which predictors are significant for predicting life expectancy in developed and developing countries.

1. **Data visualization**

We started exploring our dataset by looking into life expectancy distribution. For this purpose, we built a histogram and saw that the data is slightly skewed left. The median life expectancy is 72, and the mean is 69.22.

Chart, histogram

Description automatically generated

Fig. Life Expectancy Distribution

The next step was to look at the life expectancy in developed and developing countries. As we expected, there is a difference in life expectancy among those two categories. If we look at the median life expectancy values in developed and developing countries, we see a more than 10 years difference. Below is the boxplot of life expectancy versus the status of the country as well as the corresponding statistical data

|  |  |
| --- | --- |
| Chart, box and whisker chart  Description automatically generated  Fig. Boxplot of Life Expectancy versus Status | Table  Description automatically generated with low confidence |

We also plotted boxplots of all the predictors to see how the data is distributed and to identify the outliers. Below is just an example of some boxplots.

Chart, box and whisker chart

Description automatically generated

Fig: Boxplots of Alcohol, Hepatitis B, BMI, Polio, Total Expenditure predictors

We built a scatterplot of life expectancy versus adult mortality. It can be seen from the plot that life expectancy and adult mortality are negatively correlated. We also see two distinct clusters in the plot, one going down steeper than the other.

Chart, scatter chart

Description automatically generated

Fig: Scatter Plot of Life Expectancy versus Adult Mortality

Initially, we thought this division was related to the status of the country (developed and developing), but it was not confirmed when we looked at the data points in the scatterplot. Both clusters have data points from developed and developing countries. We believe the clusters are present because of other factors or might be incorrect data collection.

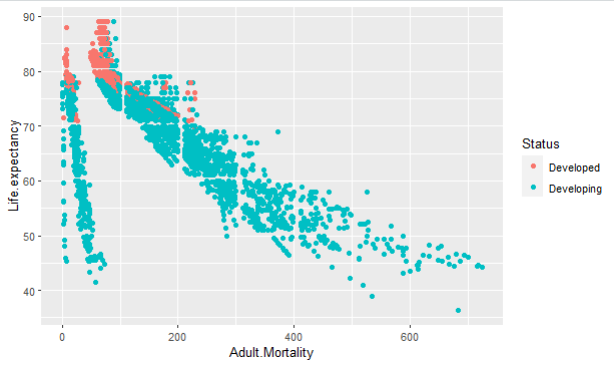


Fig: Plot of Life Expectancy versus Adult Mortality of developed and developing countries

1. **Methods**

As preliminary steps, we loaded the dataset into R, removed country and year predictors, which we thought would be redundant for our model, and converted the categorical variable status to factors.

After, we checked for missing data in our dataset. We found that 14 variables had missing data, and the variable population had the highest proportion of missing data. The picture below shows the information on the missing data. To impute the missing data, the mean imputation was used for predictors without or with a minimal number of outliers present in the distribution, and the median imputation was used for predictors with more outliers.

Table

Description automatically generated

Fig: Null value count of each predictor in dataset

Also, we looked at the correlation among predictors. For this purpose, we built a correlation plot and used the variance inflation factor (VIF) to find the correlation between predictors. Out of the 18 predictors, 3 pairs of predictors were correlated. They are (under.five.deaths, Infant.deaths), (GDP, Percentage.expenditure) and (thinness..1.19 years , thinness 5.9 years).

Chart, scatter chart

Description automatically generated

Fig: Correlation plot between predictors

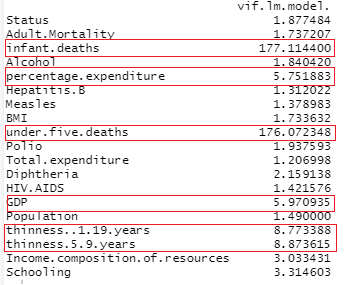


Fig: VIF values of predictors

To eliminate the multicollinearity issue, we removed under.five.deaths, percentage. Expenditure, and thinness. 5.9. years predictors and built a multiple linear regression model with noncollinear predictors. Later we used the step function to find significant predictors, 15 predictors were selected. We performed 7-fold cross-validation on the dataset, fitted the linear model on the training sets, and predicted life expectancy on the test set. We assessed the model based on R2, MSE, RMSE, and NRMSE values.

1. **Results**

The model has a good fit with R-squared 0.8112 and adjusted R-squared 0.81. The RMSE of results from 7 fold cross validation is 4.168045. And based on the NRMSE = 0.05981204, the model has good accuracy, the linear model gives a 5.98% error for life expectancy prediction.

We assessed the model based on the assumptions of linearity, normality, and constant variance. We plotted the residue versus yhat and saw that variance is constant and there no pattern.

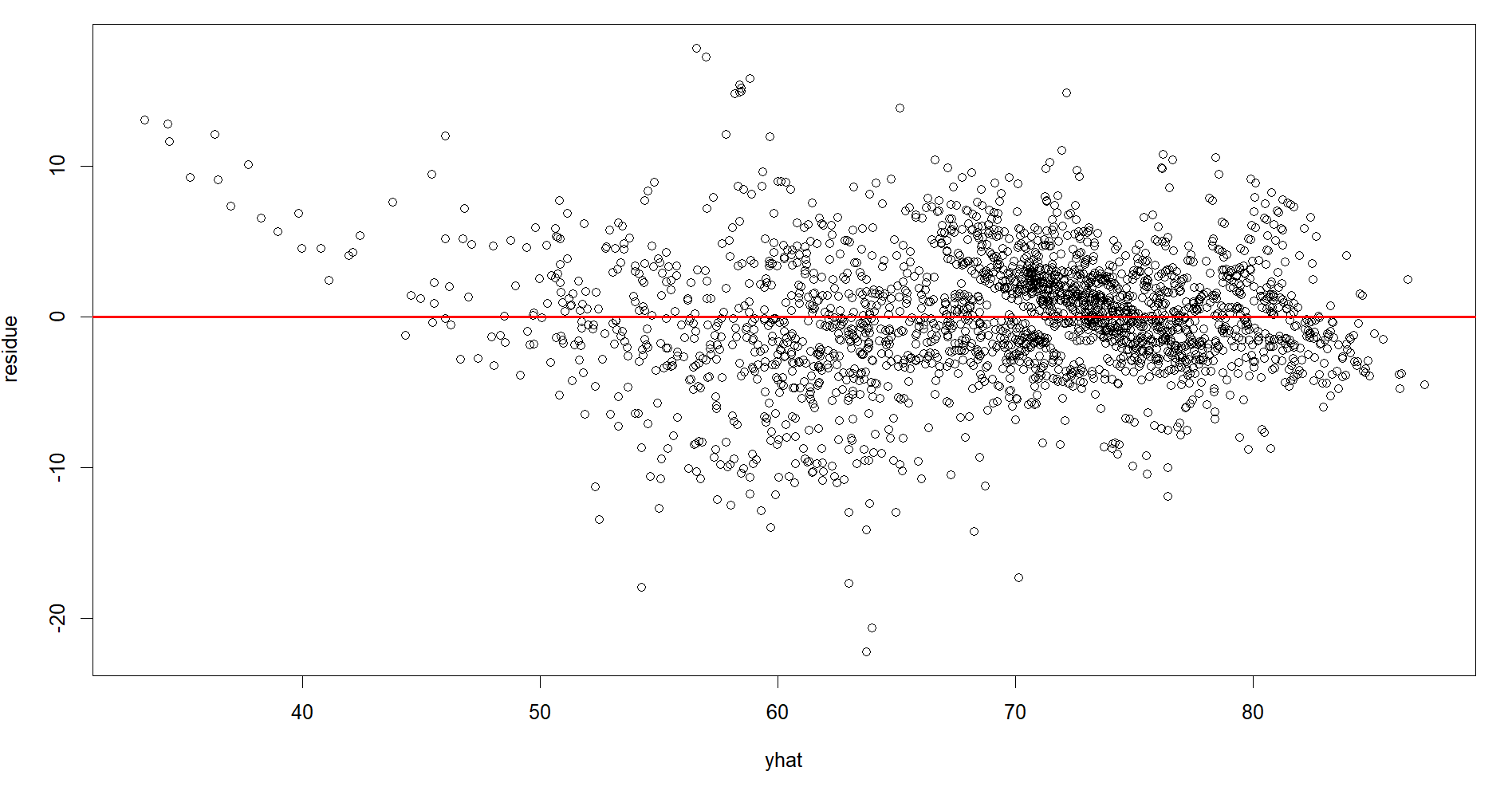


Fig: Plot of Residuals versus Predicted values

We also wanted to check if errors follow a normal distribution; we used a Q-Q plot to compare the residual to observations from a normal distribution. The Q-Q plot showed that the tails do not follow the normal line. So, we tried to improve the model using square root and log transformations. As we can see from the plots below, transformations did not improve the model, so it was decided to stay with the original model.

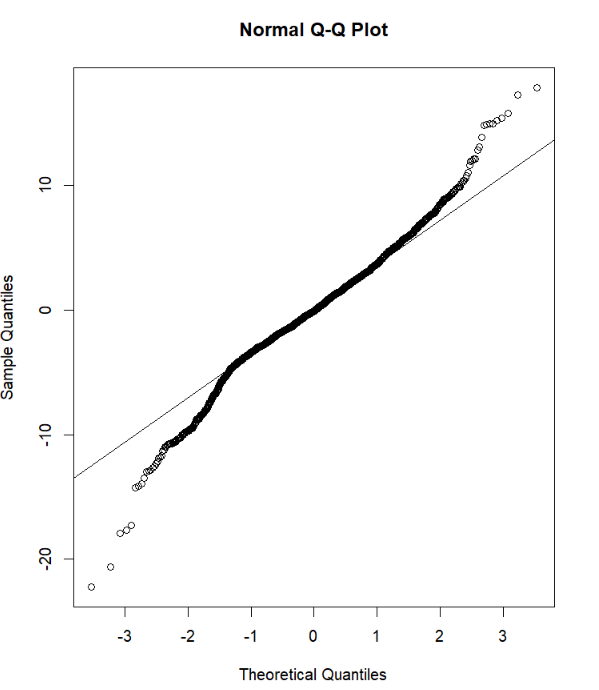


Fig: Q-Q plot of original model

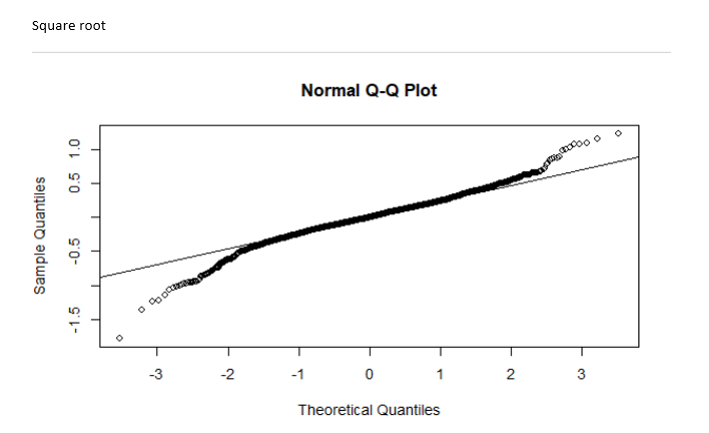
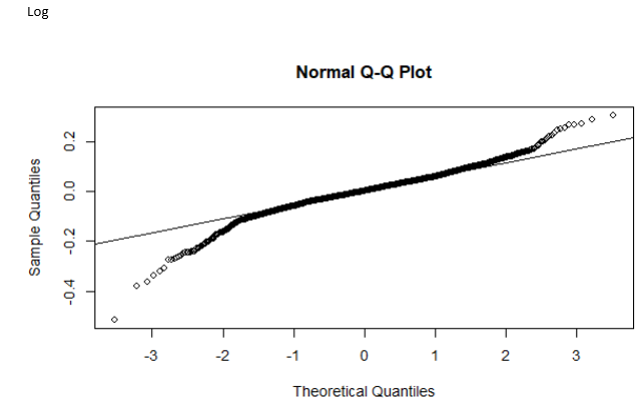
 

Fig: Q-Q plot of square root transformed model Fig: Q-Q plot of log transformed model

We performed a t-test to find if there is a difference between mean life expectancy in developed and developing countries. We found that there is a significant difference.

Graphical user interface, text

Description automatically generated

Fig: T-test between life expectancy in developed and developing countries

1. **Conclusion**

In this project, our objectives were to find statistically significant predictors by building a multiple linear regression. After building a multiple linear regression model, we found that 15 predictors were statistically significant: infant deaths, measles, polio, diphtheria, GDP, thinness 1-19 years, schooling, adult mortality, hepatitis B, BMI, total expenditure, HIV/AIDS, population, income composition of resources, status.

We also wanted to check the difference in life expectancy between developing and developed countries. After running a t-test, we found that the difference is significant. So, we can answer the last question of our project, which predictors are significant for developed and developing countries. There are 8 significant predictors for predicting life expectancy in developed countries, such as adult mortality, alcohol, BMI, total expenditure, diphtheria, GDP, thinness 1-19 years, and income composition of resources whereas there are 13 significant predictors for life expectancy in developing countries, namely adult mortality, infant deaths, alcohol, hepatitis B, measles, BMI, polio, total expenditure, diphtheria, HIV/AIDS, GDP, income composition of resources, and schooling. For developing countries, immunization-related predictors have a significant impact on life expectancy, and for developed countries, it is social and economic predictors.

1. **References**
2. The global health observatory. Explore a world of health data. Life expectancy at birth. || <https://www.who.int/data/gho/indicator-metadata-registry/imr-details/65>
3. [Irene Lebrusán Murillo](https://cenie.eu/en/users/irene-lebrusan-murillo). The life expectancy: what is it and why does it matter. || <https://cenie.eu/en/blogs/age-society/life-expectancy-what-it-and-why-does-it-matter>
4. Max Roser, Esteban Ortiz-Ospina and Hannah Ritchie (2013) - "Life Expectancy". Published online at OurWorldInData.org. || <https://ourworldindata.org/life-expectancy#:~:text=Life%20expectancy%20is%20the%20key,of%20death%20in%20a%20population>.