ANALYSIS OF COUNTRY'S LIFE EXPECTANCY

Presented to
Dr. David Goldsman

By

Team 9
Section ISyE 6414 - MSA
Jinglin Xu, Jingyi Chen, Yunpu Zeng, Renato Maynard Etchepare, Yang Lu

College of Engineering
H. Milton Stewart School of Industrial and Systems Engineering

Georgia Institute of Technology

April 26 2024

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Life expectancy has increased by over 6 years globally since 2000. With this encouraging trend in mind, this report presents a comprehensive analysis of life expectancy for each country in the world. Life expectancy serves as a pivotal indicator of national health outcomes and is influenced by a variety of factors ranging from economic conditions to healthcare policies and social issues. This study investigates the various determinants that significantly influence national health outcomes and lifespan variations from nation to nation. Our study aims to identify and quantify these factors, providing insights into how they contribute to variations in life expectancy across different countries. By leveraging statistical techniques and regression analysis, we explore relationships between life expectancy and predictors such as Gross Domestic Product (GDP), education levels, healthcare spending, and etc. The findings of this study are intended to assist policymakers, healthcare providers, and researchers in understanding the dynamics that influence life expectancy. This, in turn, could inform strategies aimed at enhancing public health and extending life spans globally. The sections below will describe our data structure, present the data analysis, discuss the implications of our findings, and suggest areas for further research.

CHAPTER 2
DATA

2.1 Data Description

For our model, we used a dataset with 22 columns and 2,938 rows, giving a total of 64,636 data points approximately. The data was obtained from the website Kaggle, and it was collected from the World Health Organization (WHO) and United Nations website during the years 2000 and 2015. The predicting variables can be divided into four main categories: Mortality Factors, Immunization Related Factors, Social Factor, and Economical Factors. The data has 2 categorical variables and 20 numerical variables, we considered Life expectancy as our dependent Variable. Table 2.1 describes the variables considered in this study. For more detailed data, please see the dataset available at Dataset Link.

Variable Name	Type	Description	Unit
Year	Quantitative	Year in which the records were observed	Year
Adult Mortality	Quantitative	Probability of dying between 15 and 60 years per 1000 population	Per 1000 population
Infant deaths	Quantitative	Number of Infant Deaths per 1000 population	Per 1000 population
Alcohol	Quantitative	Recorded per capita consumption	Litres of pure alcohol
Percentage expenditure	Quantitative	Expenditure on health as a percentage of GDP per capita	%
Hepatitis B	Quantitative	Immunization coverage among 1-year-olds	%
Measles	Quantitative	Number of reported cases per 1000 population	Per 1000 population
BMI	Quantitative	Average Body Mass Index of entire population	Index
Under-five deaths	Quantitative	Number of under-five deaths per 1000 population	Per 1000 population
Polio	Quantitative	Immunization coverage among 1-year-olds	%
Total expenditure	Quantitative	Government expenditure on health as a percentage of total expenditure	%
Diphtheria	Quantitative	Immunization coverage among 1-year-olds	%
HIV/AIDS	Quantitative	Deaths per 1000 live births	Per 1000 live births
GDP	Quantitative	Gross Domestic Product per capita	USD
Population	Quantitative	Population of the country	Persons
Thinness 1-19 years	Quantitative	Prevalence of thinness among children and adolescents 10 to 19	%
Thinness 5-9 years	Quantitative	Prevalence of thinness among children 5 to 9	%
Income composition of resources	Quantitative	Human Development Index in terms of income composition	0 to 1 scale
Schooling	Quantitative	Number of years of Schooling	Years
Country	Qualitative	Name of the country	-
Status	Qualitative	Whether the country is developed or developing	-
Life expectancy	Quantitative	Life expectancy at birth	Years

Table 2.1: Description of Variables in the Dataset

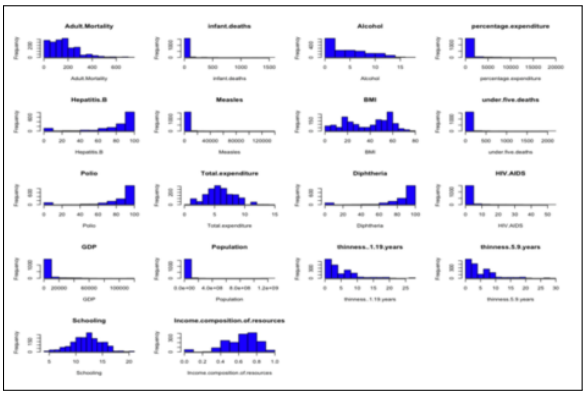
2.2 Data Pre-processing

Before using the data in our model, we addressed several pre-processing issues. First, the data had columns with missing values, in particular in Population and GDP columns. To resolve this, we deleted the rows with missing values since they did not add value to our regression model. We also considered filling the empty values using the package ”tidyverse”, which allows us to fill the columns with the data. However, certain countries such as Bolivia, had zero data in some columns, making this option not feasible. In consequence, we decided that deleting the rows without data was the best option to keep the accuracy in our prediction. Then, our dataset was reduced to 1,649 rows with 22 columns, giving a total of 36,278 data points.

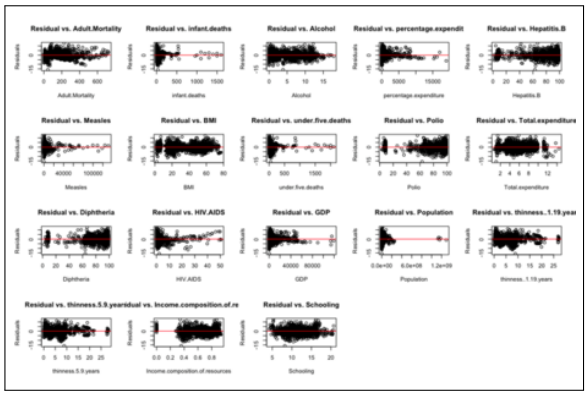
The second challenge was the multicollinearity caused by the 193 countries of our data. We believe that it was caused by countries close to each other that share similar characteristic such as economic and/or social factors, leading the columns of the dataset to be nearly linearly dependent, complicating the analysis. To address this problem we reorganized the data by continents, reducing from 193 countries to 6 continents (Asia, Africa, Europe, North America, South America, Oceania). For the transcontinental countries as Turkey, we assigned a single continent to keep the analysis simple.

2.3 Data Transformation

The first model was created performing a Multiple Linear Regression Analysis (MLR) considering all the predictors. The histograms and Plots of the residuals with respect to the numerical predictor variables showed the need of a transformation.



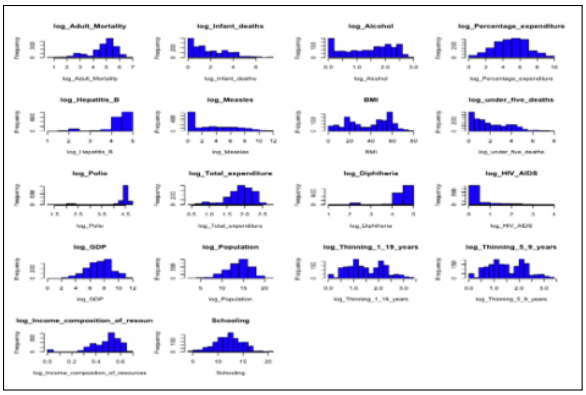
(a) Histogram of predictors before the Log Transformation.



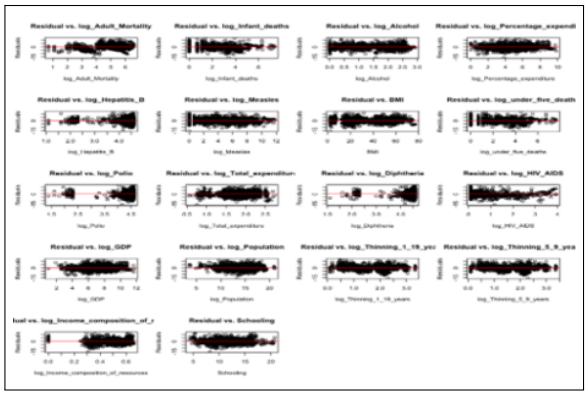
(b) Plot of residuals before the Log Transformation.

Figure 2.1: Histogram and Plot of Residuals before the Log Transformation.

The normality assumption and constant variance were severely violated for some predictors. Therefore, we applied a logarithmic transformation technique.



(a) Histogram of residuals after Log Transformation.



(b) Plot of residuals after Log Transformation.

Figure 2.2: Histogram and Plot of Residuals after the Log Transformation.

As a result, the assumption of normality and constant variance were improved across our numerical predictors. We can notice normality in almost all the histograms and a constant variance in almost all the plots.

The model after the transformation considered Continent, Status, Adult Mortality, Infants Deaths, Alcohol, Percentage Expenditure, Hepatitis B, Measles, BMI, Under 5 years old deaths, Polio, Total Expenditure, Diphtheria, HIV/AIDS, GDP, Population, Thinning between 1-19, Thinning between 5-9, Income composition of resource and schooling.

```
model_trans <- lm(log_Life_expectancy ~ Continent + Status + log_Adult_Mortality +
log_Infant_deaths + log_Alcohol + log_Percentage_expenditure + log_Hepatitis_B +
log_Measles + BMI + log_under_five_deaths + log_Polio + log_Total_expenditure +
log_Diphtheria + log_HIV_AIDS + log_GDP + log_Population + log_Thinning_1_19_years
+ log_Thinning_5_9_years + log_Income_composition_of_resources + Schooling)

Coefficients:
                            Estimate  Std. Error  t value Pr(>1t1)

(Intercept)                 4.171..00 2.109e-02   197.750  < 2e-16 ***
ContinentAsia               1.302e-02 4.465e-03   2.916   0.003593 **
ContinentEurope            -1.791e-04 6.545e-03  -0.027   0.978167
ContinentNorth America      4.614e-02 5.880e-03   7.848   7.60e-15 ***
ContinentOceania           -3.134e-02 6.925e-03  -4.526   6.44e-06 ***
ContinentSouth America      2.157e-02 6.591e-03   3.273   0.001085 **
StatusDeveloping           -1.994e-02 5.236e-03  -3.809   0.000145 ***
log_Adult_Mortality        -9.053e-03 1.279e-03  -7.080   2.14e-12 ***
log_Infant_deaths           3.206e-02 9.170e-03   3.496   0.000485 ***
log_Alcohol                -2.601e-03 2.165e-03  -1.202   0.229717
log_Percentage_expenditure 9.432e-03 1.817e-03   5.189   2.37e-07 ***
log_Hepatitis_B            -1.899e-03 2.163e-03  -0.878   0.380048
log_Measles                 1.497e-03 5.541e-04   2.701   0.006987 **
BMI                         5.430e-05 8.526e-05   0.637   0.524293
log_under_five_deaths      -3.970e-02 8.812e-03  -4.505   7.11e-06 ***
log Polio                   1.464e-03 2.431e-03   0.602   0.547062
log_Total_expenditure       2.590e-05 3.480e-03   0.007   0.994062
log_Diphtheria              4.465e-03 2.745e-03   1.627   0.103980
log_HIV_AIDS               -8.391e-02 2.310e-03  -36.323   < 2e-16 ***
log_GDP                    -4.771e-03 1.970e-03  -2.421   0.015568 *
log_Population             -1.630e-04 5.424e-04  -0.301   0.763763
log_Thinning_1_19_years     2.922e-03 4.927e-03   0.593   0.553205
log_Thinning_5_9_years     -1.840e-02 5.027e-03  -3.659   0.000261 ***
log_Inc_compos_of_resou     1.480e-01 1.527e-02   9.693    < 2e-16 ***
Schooling                   9.253e-03 8.386e-04  11.034   < 2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '  1

Residual standard error: 0.04884 on 1624 degrees of freedom
Multiple R-squared: 0.8669,      Adjusted R-squared: 0.8649
F-statistic: 440.6 on 24 and 1624 DF,   p-value: < 2.2e-16
```

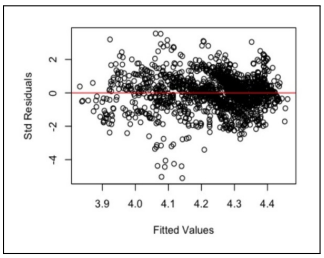
Figure 2.3: Initial Model with Log Transformation

The baseline corresponds to Africa and Developed. The model explains 86.69% of the variability in the life expectancy. We can see that Continent, Status, log_Adult_Mortality, log_Infant_deaths , log_Percentage_expenditure, measles, log_under_five_deaths,HIV/AIDS, log_Thinning_5_9_years, log_Inc_compos_of_resou, and Schooling are highly significant with P values almost 0. A goodness of fit test is performanced in the next section.

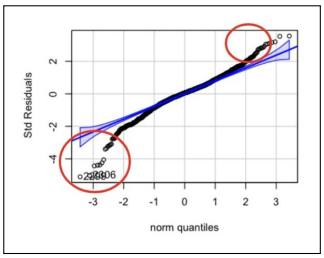
CHAPTER 3
MODEL AND ANALYSES

3.1 Goodness of Fit

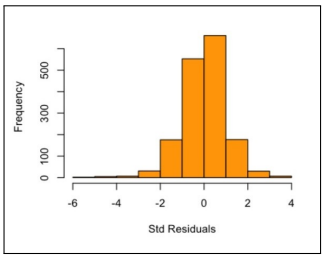
To evaluate the suitability of our initial model and verify the adherence to the four fundamental assumptions of regression—linearity, constant variance, independence, and normality—we employed three diagnostic plots: the plot of fitted values vs. residuals, the Normal Q-Q plot, and the Standard Residual Histogram. The analysis of these plots suggests that our model meets the required assumptions. Specifically, the residuals vs. fitted values plot does not show any obvious curvature or non-random distribution of residuals, and the points in the histogram are evenly distributed, supporting the assumptions of linearity, constant variance, and normality. However, the Q-Q plot indicates the presence of potential outliers in the tails, which suggests that further refinement may be necessary by addressing these outliers to enhance model accuracy. Additionally, our investigation into multicollinearity among predictors revealed a high correlation between log_under_five_deaths and log_infant_deaths. To mitigate this issue, we retained log_infant_deaths in our model and excluded log_under_five_deaths, thereby reducing the risk of multicollinearity.



(a) Residuals vs Fitted



(b) Normal Q-Q plot



(c) Std. Res. Histogram

Figure 3.1: Goodness of Fit with Initial Log Model

3.2 Variable selection

To enhance both the performance of the model and the accuracy of predictions, it's crucial to use Variable Selection. We utilize Variable Selection through Bidirectional Stepwise Selection and p-values, and compare the two methods to identify the most effective approach. Figure 3.2a illustrates the variables selected using Bidirectional Stepwise Selection while Figure 3.2b shows the Variables selected from the p-values significant Variable Selection.

```
Coefficients:
               (Intercept)       ContinentAsia
                 4.162697            0.016098
   Continent North America    ContinentOceania
                 0.046675           -0.027413
         StatusDeveloping log_Adult_Mortality
                -0.017498           -0.009712
log_Percentage_expenditure         log Measles
                 0.008848            0.001284
              log_HIV AIDS             log_GDP
                -0.086681           -0.004067
   log_Inc_compos_of_resou           Schooling
                 0.148126            0.009509
          ContinentEurope   log_Infant_deaths
                -0.001129           -0.009138
    ContinentSouth America      log_Diphtheria
                 0.020232            0.003772
    log_Thinning_5_9_years
                -0.016211
```

(a) Bidirectional stepwise selection

```
Coefficients:
                            Estimate  Std. Error  t value    Pr(>|t|)
(Intercept)                4.1782395  0.0149119  280.194   < 2e-16 ***
ContinentAsia              0.0163125  0.0043956    3.711  0.000213 ***
ContinentEurope           -0.0013162  0.0061081   -0.215  0.829414
ContinentNorth America     0.0464235  0.0055707    8.333   < 2e-16 ***
ContinentOceania          -0.0284295  0.0067495   -4.212  2.67e-05 ***
ContinentSouth America     0.0200959  0.0062154    3.233  0.001248 **
StatusDeveloping          -0.0177287  0.0051899   -3.416  0.000651 ***
log_Adult_Mortality       -0.0967654  0.0012729  -76.061  4.93e-14 ***
log_Infant_deaths         -0.0092013  0.0011945   -7.703  2.28e-14 ***
log_Percentage_expenditure 0.0089135  0.0018016    4.948  8.29e-07 ***
log_Measles                0.0012454  0.0005502    2.263  0.023746 *
log_HIV_AIDS              -0.0867943  0.0021576  -40.227   < 2e-16 ***
log_GDP                   -0.0041824  0.0019487   -2.146  0.032002 *
log_Thinning_5_9_years    -0.0163152  0.0026513   -6.154  9.50e-10 ***
log_Inc_compos_of_resou    0.1501516  0.0151874    9.887   < 2e-16 ***
Schooling                  0.0095940  0.0008014   11.971   < 2e-16 ***
```

(b) Variables selection using p-value

Figure 3.2: Variables selected using Bidirectional Stepwise Selection vs the p-values

3.3 ANOVA

By using ANOVA to compare the initial log model and reduced model from Bidirectional Stepwise Variable selection, we were able to streamline the model while maintaining predictive accuracy. The detailed Variables are shown in Figure 7. Specifically, Model 1 (Reduced Model) has fewer variables than Model 2 (Full Model), and the simplicity of the model is improved. The p-value of the ANOVA result is 0.7735, much higher than the commonly used significance level of 0.05, thus we do not reject the null hypothesis. Therefore, we can choose the simplified Model 1 with confidence, as it provides a more efficient and easy to interpret model.

```
Analysis of Variance Table

Model 1: log_Life_expectancy ~ Continent + Status + log_Adult_Mortality +
          log_Infant_deaths + log_Percentage_expenditure + log_Measles +
          log_Diphtheria + log_HIV_AIDS + log_GDP + log_Thinning_5_9_years +
          log_Income_composition_of_resources + Schooling
Model 2: log_Life_expectancy ~ Continent + Status + log_Adult_Mortality +
          log_Infant_deaths + log_Alcohol + log_Percentage_expenditure +
          log_HepatitisB + log_Measles + BMI + log_Polio + log_Total_expenditure +
          log_Diphtheria + log_HIV_AIDS + log_GDP + log_Population +
          log_Thinning_1_19_years + log_Thinning_5_9_years + log_Income_composition_of_resources +
          Schooling

          Res.Df    RSS     Df Sum of Sq      F    Pr(>F)
1         1632    3.9320
2         1625    3.9222    7   0.0097825   0.579   0.7735
```

Figure 3.3: ANOVA Analysis

3.4 Model Evaluation

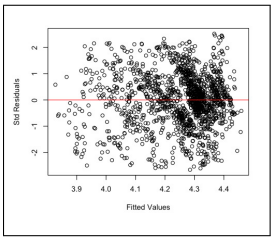
The final MLR model we generated in the previous sections is shown below

$$y = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \beta_3 d_3 + \beta_4 d_4 + \beta_5 d_5 + \beta_6 d_6 + \beta_7 x_1 + \beta_8 x_2 + \beta_9 x_3 + \beta_{10} x_4 + \beta_{11} x_5 + \beta_{12} x_6 + \beta_{13} x_7 + \beta_{14} x_8 + \beta_{15} x_9 + \beta_{16} x_{10}$$

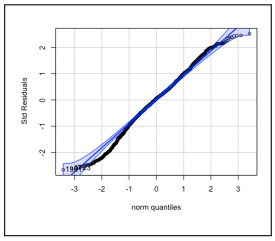
where the variable y denotes life expectancy, which has undergone a transformation using the Logit function. The coefficient β_i, for each predictor and the intercept, belongs to the set $i \in \{0, 1, 2, \ldots, 16\}$. Dummy variables d_j are used to reference continents, with Africa set as the baseline category, and j is an element of the set $\{1, 2, \ldots, 5\}$ and d_6 is used to reference status, with a baseline of developed countries. The variables x_n represent other predictors associated with the prediction of life expectancy, where n is in the set $\{1, 2, \ldots, 10\}$. Then, the final model is as follows

$$y = 4.1719 - 0.0167 d_1 - 0.06726 d_2 - 0.0426 d_3 + 0.0371 d_4 - 0.0152 d_5 - 0.0199 d_6 - 0.0009 x_1 - 0.0093 x_2 + 0.0032 x_3 + 0.0018 x_4 + 0.0036 x_5 - 0.0888 x_6 - 0.0036 x_7 - 0.014 x_8 + 0.1499 x_9 + 0.093 x_{10}$$

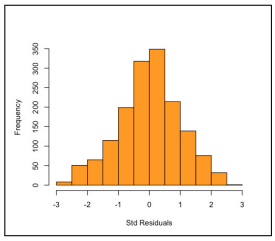
The assumptions of multiple linear regression were tested again in this section.



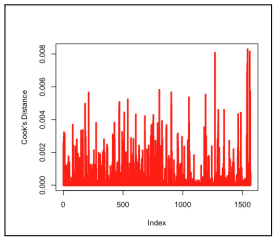
(a) Residuals vs Fitted



(b) Normal Q-Q plot



(c) Std. Res. Histogram



(d) Cook's distance

Figure 3.4: Plots for residuals and Cook's distance

The diagnostic plots suggest that the multiple linear regression model is reasonably well-fitted to the data. The Residuals vs Fitted plot shows a fairly random dispersion of residuals, indicating that the assumption of linearity holds true for most of the data. The Normal Q-Q plot's alignment along the reference line in the central quantiles supports the assumption that residuals are normally distributed, with minor deviations at the tails that are often seen in practice. The histogram of standardized residuals appears mostly symmetrical, reinforcing the notion of normality. Cook's distance values are well below the threshold, suggesting that no single observation unduly influences the model. Overall, despite slight deviations, these diagnostics collectively suggest that the model's assumptions are adequately met, making it a reliable tool for interpreting the relationship between the predictors and the response variable.

Generally, all assumptions are held. There are three observations with a Cook's Distance noticeably higher than the other observations. However, its Cook's distance is close to 0.004, suggesting that there are likely no outliers.

```
Coefficients:
                                   Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)                       4.1719746  0.0147252  283.323    < 2e-16 ***
ContinentAsia                     0.0167879  0.0036657    4.580   5.03e-06 ***
ContinentEurope                  -0.0126255  0.0050200   -2.515   0.01200  *
ContinentNorth America            0.0426576  0.0045615    9.352    < 2e-16 ***
ContinentOceania                 -0.0371682  0.0055901   -6.649   4.08e-11 ***
ContinentSouth America            0.0152563  0.0050752    3.006   0.00269  **
StatusDeveloping                 -0.0199044  0.0042489   -4.685   3.05e-06 ***
log_Adult_Mortality              -0.0089073  0.0010609   -8.396    < 2e-16 ***
log_Infant_deaths                -0.0093678  0.0009884   -9.478    < 2e-16 ***
log_Percentage_expenditure        0.0083230  0.0014817    5.617   2.29e-08 ***
log_Measles                       0.0018172  0.0004577    3.970   7.50e-05 ***
log_Diphtheria                    0.0036321  0.0019070    1.905   0.05702  .
log_HIV AIDS                     -0.0888208  0.0017911  -49.590    < 2e-16 ***
Log_GDP                          -0.0036976  0.0016101   -2.297   0.02177  *
log_Thinning_5_9_years           -0.0194446  0.0022088   -8.803    < 2e-16 ***
log_Income_composition_of_resources   0.1499833  0.0126894   11.820    < 2e-16 ***
Schooling                         0.0093658  0.0006691   13.997    < 2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03958 on 1553 degrees of freedom
Multiple R-squared: 0.906, Adjusted R-squared: 0.9051
F-statistic: 935.8 on 16 and 1553 DF, p-value: 2.2e-16
```

Figure 3.5: Summary of the Model

As for model performances, the p-values for each predictors is smaller than 0.01, showing the whole model is significant. Moreover, the R^2 is 90.6%, which indicates that 90.60% of total variability in life expectancy that can be explained by the regression. Overall, the model demonstrates a strong relationship between the predictors and log-transformed life expectancy, with a high degree of explained variability and statistically significant relationships for most included predictors.

3.5 Model Prediction

For the model prediction, we firstly spitted data using K-fold method. In k-fold cross-validation, the data set is randomly divided into k equally sized segments, known as "folds". Out of these, one fold is set aside as the validation set for evaluating the model's performance, while the combined remaining k - 1 folds are used to train the model. This process is systematically repeated so that each fold serves as the validation set exactly once, ensuring that every data point has had the opportunity to be part of the validation set and thus contribute to the overall model assessment. In this study, we set $k = 10$ which is commonly used in k-fold method.

Model accuracy is commonly evaluated using several statistical metrics that quantify the difference between observed values and predictions made by the model. Three widely used accuracy

indices are the Mean Squared Prediction Error (MSPE), the Mean Absolute Error (MAE), and the Mean Absolute Percentage Error (MAPE). In the model's prediction, the MSPE, MAE and MAPE indicate a high level of accuracy

$$\text{MSPE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\%$$

Figure 3.6: Accuracy indices formulas

Model Accuracy	Value
MSPE	0.0015
MAE	0.0308
MAPE	0.7305

Table 3.1: Summary of model accuracy

To compare the ability of various methods to predict life expectancy, we introduced two machine learning methods at the same time. They are gradient boosting method and random forest method. Gradient boosting is a machine learning technique based on boosting in a functional space, where the target is pseudo-residuals rather than the typical residuals used in traditional boosting. It gives a prediction model in the form of an ensemble of weak prediction models, i.e., models that make very few assumptions about the data, which are typically simple decision trees. [3, 1]. When a decision tree is the weak learner, the resulting algorithm is called gradient-boosted trees; it usually outperforms random forest. [4]. Bagging or bootstrap aggregation is a technique for reducing the variance of an estimated prediction function. Bagging seems to work especially well for high-variance, low-bias procedures, such as trees. For regression, we simply fit the same regression tree many times to bootstrap-sampled versions of the training data, and average the result. Random forests is a substantial modification of bagging that builds a large collection of de-correlated trees, and then averages them. On many problems the performance of random forests is very similar to boosting, and they are simpler to train and tune. As a consequence, random forests are popular, and are implemented in a variety of packages. [2]. The model performances of these two machine learning methods are shown below

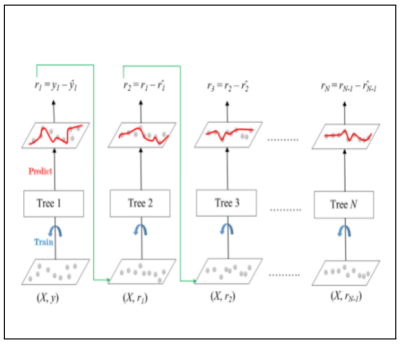
```
Stochastic Gradient Boosting

1570 samples
  12 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 1413, 1414, 1411, 1414, 1414 ...
Resampling resuls across tuning parameters:

 interaction.depth n.trees RMSE       R.squared  MAE
 1                  50     0.03996146 0.910763   0.03075851
 1                 100     0.03521004 0.9245958  0.02686290
 1                 150     0.03369966 0.9307513  0.02553831
 2                  50     0.03405066 0.9303239  0.02604782
 2                 100     0.03013934 0.9445881  0.02286643
 2                 150     0.02835450 0.9509791  0.02125959
 3                  50     0.03110630 0.9414049  0.02332171
 3                 100     0.02774580 0.9529884  0.02058620
 3                 150     0.02612523 0.9582796  0.01913945

Tuning parameter 'shrinkage' was held constant at a value of 0.1
Tuning parameter 'n.minobsinnode' was held constant
 at a value of 10
RMSE was used to select the optimal model using the smallest
value.
The final values used for the model were n.trees = 150,
interaction.depth = 3, shrinkage = 0.1 and n.minobsinnode = 10.
```

(a) Results of gradient boosting

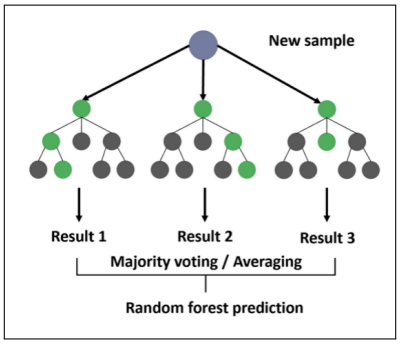
```
Random Forest

1570 samples
  12 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 1413, 1414, 1411, 1414, 1414, ...
Resampling results across tuning parameters:

  mtry RMSE       Rsquared  MAE

    2  0.02932698 0.9546289 0.01964892
    9  0.02054003 0.9747191 0.01331485
    16 0.02079078 0.9738984 0.01343330

RMSE was used to select the optimal model using the
smallest value.
The final value used for the model was mtry = 9.
```

(b) Results of random forest

Figure 3.7: Results of gradient boosting and Random Forest.



(a) Gradient boosting.



(b) Random Forest method.

Figure 3.8: Gradient boosting and random forest method

The k-fold validation with $k = 10$ was employed for these two methods, and the prediction accuracy for three models are show below:

	MLR	Gradient Boosting	Random Forest
MSPE	0.0015	0.000 487	7.2618×10^{-5}
MAE	0.0308	0.016 46	0.005 38
MAPE	0.7305	0.389 06	0.127 63

Table 3.2: Model Prediction Accuracy

Generally, the prediction performances of random forest method is the best among the three models. However, the ML methods may encounter over-fitting issue and are less interpretable.

Example of Prediction Life Expectancy in North America:

Assumptions: we assume Continent="North America"; Status="Developed"; the other values of predictors we select is the median of the predictors. The detailed formula is described as follows:

$$y = \beta_0 + \beta_3 d_3 + \beta_7 x_1 + \beta_8 x_2 + \beta_9 x_3 + \beta_{10} x_4 + \beta_{11} x_5 + \beta_{12} x_6 + \beta_{13} x_7 + \beta_{14} x_8 + \beta_{15} x_9 + \beta_{16} x_{10}$$

where x_i is the median value of each predictors and $d_3 = 1$ is North America. The predictions of life expectancy in North America for each method are shown below

Results	Model	Life Expectancy
	MLR	77.4
	Gradient Boosting	74.3
	Random Forest	73.7

Table 3.3: Life Expectancy Results by Model

From the results shown in the table, the two machine learning methods appear to have more similar results compared with the multiple linear regression model. The results of predicted life expectancy among the three models are: MLR 77.4 $>$ Gredient Boosting 74.3 $>$ Random Forest 73.7. The average value among the three models is calculated as 75.1.

CHAPTER 4
CONCLUSION AND RECOMMENDATIONS

This study has successfully explored the various determinants impacting life expectancy across different nations and continents, utilizing a robust regression analysis framework with all necessary steps.

We have identified several key factors, economic indicators such as income, health infrastructure such as adult mortality rate and HIV/AIDS vaccination rate, and societal factors such as schooling, that significantly influence life expectancy. The analysis provided clear evidence that improvements in these areas are closely linked with enhancements in national longevity. This study can give a clear picture to policy makers as well as stake holders in best allocating resources and funding in support of citizens' health outcomes.

Future studies could incorporate additional predictors, such as environmental factors, genetic predispositions, and more detailed socioeconomic data, to capture the complexity of health outcomes more comprehensively. In addition, an analytical focus on the impact of specific health policies on life expectancy across different regimes expanding from current study could guide effective policy formulation and public health improvements.

APPENDICES

```
# Project ISYE 6414
# Library
library(lmtest)
library(car)
library(MASS)
library(ggplot2)
library(leaps)

# Set the working directory to where your CSV file is located
setwd('C:\\Users\\...')#import data
file_name <- "Life_Expectancy_Data.csv"
df <- read.csv(file_name)

# There are two approach here, delete the rows without info, or inpute missing values.

# Delete Rows
df <- na.omit(df)
cat("Data after removing rows with any missing values:\n")
print(dim(df))

#See the head of the dataframe
head(df)

# Access with the name of the columns
Country = df$Country
Year = df$Year
Status = df$Status
Life_expectancy =df$Life.expectancy
Adult_Mortality =df$Adult.Mortality
Infant_deaths=df$infant.deaths
Alcohol=df$Alcohol
Percentage_expenditure=df$percentage.expenditure
Hepatitis_B=df$Hepatitis.B
Measles = df$Measles
BMI = df$BMI
under_five_deaths = df$under.five.deaths
Polio = df$Polio
Total_expenditure = df$Total.expenditure
Diphtheria = df$Diphtheria
HIV_AIDS = df$HIV.AIDS
GDP = df$GDP
Population = df$Population
Thinning_1_19_years = df$thinness..1.19.years
Thinning_5_9_years = df$thinness.5.9.years
Income_composition_of_resources = df$Income.composition.of.resources
Schooling = df$Schooling

#Regroup the countries in Continents to improve the assumptions and reduce collinearity.
continent_mapping <- c(
  "Afghanistan" = "Asia",
        .
        .
        .
  "Zimbabwe" = "Africa"
)

#Create new column Continent
df$Continent <- continent_mapping[df$Country]
sum(is.na(df$Continent))
Continent = df$Continent

#Deleting Country
df$Country <- NULL

# Multiple Linear Regression
```

```r
#MLR
model <- lm(Life_expectancy ~ Continent + Year + Status + Adult_Mortality +
              Infant_deaths + Alcohol + Percentage_expenditure + Hepatitis_B +
              Measles + BMI + under_five_deaths + Polio + Total_expenditure +
              Diphtheria + HIV_AIDS + GDP + Population + Thinning_1_19_years +
              Thinning_5_9_years + Income_composition_of_resources + Schooling,
            data = df)
model_summary <- summary(model)
model_summary

# Data exploration
# residuals here
par(mfrow=c(4,5))
predictor_vars <- c("Adult.Mortality", "infant.deaths", "Alcohol",
                    "percentage.expenditure", "Hepatitis.B",
                    "Measles", "BMI", "under.five.deaths", "Polio",
                    "Total.expenditure", "Diphtheria", "HIV.AIDS", "GDP",
                    "Population", "thinness..1.19.years", "thinness.5.9.years",
                    "Income.composition.of.resources", "Schooling"
                    )

for(var in predictor_vars){
  if(is.numeric(df[[var]]) | is.integer(df[[var]])) {
    plot(df[[var]], model$residuals, xlab=var,ylab="Residuals")
    abline(h=0, col='red')
    title(paste("Residual vs.", var))
  }}

plot(df$Adult_Mortality, model$residuals, xlab=Adult_Mortality,ylab="Residuals")

print(dim(df$Adult_Mortality))

# histogram here
par(mfrow=c(5,4))
hist_vars <- c("Adult.Mortality", "infant.deaths", "Alcohol",
                    "percentage.expenditure", "Hepatitis.B",
                    "Measles", "BMI", "under.five.deaths", "Polio",
                    "Total.expenditure", "Diphtheria", "HIV.AIDS", "GDP",
                    "Population", "thinness..1.19.years", "thinness.5.9.years","Schooling"
                    ,"Income.composition.of.resources"
)

for(var in hist_vars){
  if(is.numeric(df[[var]]) || is.integer(df[[var]])) {
    hist(df[[var]], main=paste(var), xlab=var, col="blue")
  }
}
# goodness of fit
resids_inital = stdres(model)
fits_intial = model$fitted
par(mfrow =c(2,2))
plot(fits_intial, resids_inital, xlab="Fitted Values",ylab="Std Residuals")
abline(0,0,col="red")

qqPlot(resids_inital, ylab=" Std Residuals", main = "")
hist(resids_inital, xlab="Std Residuals", main = "",nclass=10,col="orange")

# Do the log transformation
df$log_Life_expectancy <- log(df$Life.expectancy + 1)
df$log_Infant_deaths <- log(df$infant.deaths + 1)
df$log_Adult_Mortality <- log(df$Adult.Mortality + 1)
df$log_Alcohol <- log(df$Alcohol + 1)
df$log_Percentage_expenditure <- log(df$percentage.expenditure + 1)
df$log_Hepatitis_B <- log(df$Hepatitis.B + 1)
df$log_Measles <- log(df$Measles + 1)
df$log_under_five_deaths <- log(df$under.five.deaths + 1)
df$log_Polio <- log(df$Polio + 1)
df$log_Total_expenditure <- log(df$Total.expenditure + 1)
df$log_Diphtheria <- log(df$Diphtheria + 1)
df$log_HIV_AIDS <- log(df$HIV.AIDS + 1)
df$log_GDP <- log(df$GDP + 1)
```

```r
df$log_Population <- log(df$Population + 1)
df$log_Thinning_1_19_years <- log(df$thinness..1.19.years + 1)
df$log_Thinning_5_9_years <- log(df$thinness.5.9.years + 1)
df$log_Income_composition_of_resources <- log(df$Income.composition.of.resources + 1)

# check the residuals and histogram again after log transformation
# residuals and histogram
par(mfrow=c(5,4))
predictor_vars <- c("log_Adult_Mortality", "log_Infant_deaths",
                    "log_Alcohol", "log_Percentage_expenditure", "log_Hepatitis_B",
                    "log_Measles", "BMI", "log_under_five_deaths", "log_Polio",
                    "log_Total_expenditure", "log_Diphtheria", "log_HIV_AIDS", "log_GDP",
                    "log_Population", "log_Thinning_1_19_years", "log_Thinning_5_9_years",
                    "log_Income_composition_of_resources", "Schooling"
)

for(var in predictor_vars){
  if(is.numeric(df[[var]]) | is.integer(df[[var]])) {
    plot(df[[var]], model$residuals, xlab=var,ylab="Residuals")
    abline(h=0, col='red')
    title(paste("Residual vs.", var))
  }}
# histogram
par(mfrow=c(5,4))
hist_vars <- c("log_Adult_Mortality", "log_Infant_deaths",
               "log_Alcohol", "log_Percentage_expenditure", "log_Hepatitis_B",
               "log_Measles", "BMI", "log_under_five_deaths", "log_Polio",
               "log_Total_expenditure", "log_Diphtheria", "log_HIV_AIDS", "log_GDP",
               "log_Population", "log_Thinning_1_19_years", "log_Thinning_5_9_years",
               "log_Income_composition_of_resources", "Schooling"
)

for(var in hist_vars){
  if(is.numeric(df[[var]]) || is.integer(df[[var]])) {
    hist(df[[var]], main=paste(var), xlab=var, col="blue")
  }
}
# Get the log transformation model
model_trans <- lm(log_Life_expectancy ~ Continent + Status + log_Adult_Mortality +
                    log_Infant_deaths + log_Alcohol + log_Percentage_expenditure + log_Hepatitis_B +
                  log_Measles + BMI + log_under_five_deaths + log_Polio + log_Total_expenditure +
                  log_Diphtheria + log_HIV_AIDS + log_GDP + log_Population + log_Thinning_1_19_years +
                  log_Thinning_5_9_years + log_Income_composition_of_resources + Schooling,
            data = df)
model_trans_summary <- summary(model_trans)
model_trans_summary

# qqplot goodness of fit
resids_inital = stdres(model_trans)
fits_intial = model_trans$fitted
par(mfrow =c(2,2))
plot(fits_intial, resids_inital, xlab="Fitted Values",ylab="Std Residuals")
abline(0,0,col="red")

qqPlot(resids_inital, ylab=" Std Residuals", main = "")
hist(resids_inital, xlab="Std Residuals", main = "",nclass=10,col="orange")

# Multicollinearity VIF check, we delete the log_under_five_deaths
vif.results <- vif(model_trans)
print(vif.results)

####### Variable Stepwise Bidirectional Stepwise Selection
model_trans <- lm(log_Life_expectancy ~ Continent + Status + log_Adult_Mortality + log_Infant_deaths +
                    log_Alcohol + log_Percentage_expenditure + log_Hepatitis_B +
                    log_Measles + BMI + log_Polio + log_Total_expenditure +
                    log_Diphtheria + log_HIV_AIDS + log_GDP + log_Population + log_Thinning_1_19_years +
                    log_Thinning_5_9_years + log_Income_composition_of_resources + Schooling,
                  data = df)

selected_model <- stepAIC(model_trans, direction = "both", trace = FALSE)
summary(selected_model)
```

```r
selected_model2 <- stepAIC(model_trans, direction = "forward", trace = FALSE)
summary(selected_model2)

########select significant variables using p-values
model_trans2 <- lm(log_Life_expectancy ~ Continent + Status + log_Adult_Mortality +
                     log_Infant_deaths  + log_Percentage_expenditure +
                     log_Measles + log_HIV_AIDS + log_GDP +
                     log_Thinning_5_9_years + log_Income_composition_of_resources + Schooling,
                   data = df)
summary(model_trans2)
resids = stdres(model_trans2)
fits = model_trans2$fitted
cook = cooks.distance(model_trans2)
par(mfrow =c(2,2))
plot(fits, resids, xlab="Fitted Values",ylab="Std Residuals")
abline(0,0,col="red")

qqPlot(resids, ylab="Std Residuals", main = "")
hist(resids, xlab="Std Residuals", main = "",nclass=10,col="orange")
plot(cook,type="h",lwd=3,col="red", ylab = "Cook's Distance")

#AIC CHECK
aic_value <- AIC(selected_model)
aic_value
aic_value <- AIC(model_trans2)
aic_value

#ANOVA
anova(selected_model, model_trans)


########get rid of outliers
resids = stdres(selected_model)
outliers <- which(resids > 2 | resids < -2)
print(length(outliers))
df_clean <- df[-outliers, ]
print(names(df_clean))
selected_model_clean <- lm(log_Life_expectancy ~ Continent + Status + log_Adult_Mortality +
                             log_Infant_deaths + log_Percentage_expenditure +
                             log_Measles + log_Diphtheria + log_HIV_AIDS +
                             log_GDP + log_Thinning_5_9_years + log_Income_composition_of_resources +
                             Schooling,data = df_clean)
summary(selected_model_clean)
resids = stdres(selected_model_clean)
fits = selected_model_clean$fitted
cook = cooks.distance(selected_model_clean)
par(mfrow =c(2,2))
plot(fits, resids, xlab="Fitted Values",ylab="Std Residuals")
abline(0,0,col="red")

qqPlot(resids, ylab=" Std Residuals", main = "")
hist(resids, xlab="Std Residuals", main = "",nclass=10,col="orange")
plot(cook,type="h",lwd=3,col="red", ylab = "Cook's Distance")
vif.results <- vif(selected_model_clean)
print(vif.results)

######### Prediction
library(caret)
set.seed(123)
ctrl <- trainControl(method = "cv", number = 10)
model_mlr <- train(log_Life_expectancy ~ Continent + Status + log_Adult_Mortality +
                     log_Infant_deaths + log_Percentage_expenditure +
                     log_Measles + log_Diphtheria + log_HIV_AIDS +
                     log_GDP + log_Thinning_5_9_years + log_Income_composition_of_resources +
                     Schooling,
                   data = df_clean,
                   method = "lm",
                   trControl = ctrl)
print(model_mlr)
predictions <- predict(model_mlr, newdata = df_clean)
actual <- df_clean$log_Life_expectancy
```

```
results <- postResample(predictions, actual)
MSPE <- results["RMSE"]^2
MAE <- results["MAE"]
MAPE <- mean(abs((actual - predictions) / actual), na.rm = TRUE) * 100
print(paste("MSPE:", MSPE))
print(paste("MAE:", MAE))
print(paste("MAPE:", MAPE))

########Gredient Boosting Machine
library(caret)
library(gbm)
set.seed(123)
ctrl <- trainControl(method = "cv", number = 10, verboseIter = FALSE)
model_gbm <- train(log_Life_expectancy ~ Continent + Status + log_Adult_Mortality +
                     log_Infant_deaths + log_Percentage_expenditure +
                     log_Measles + log_Diphtheria + log_HIV_AIDS +
                     log_GDP + log_Thinning_5_9_years + log_Income_composition_of_resources +
                     Schooling,
                   data = df_clean,
                   method = "gbm",
                   trControl = ctrl,
                   verbose = FALSE)
print(model_gbm)

predictions <- predict(model_gbm, newdata = df_clean)

actual <- df_clean$log_Life_expectancy

actual <- df_clean$log_Life_expectancy
results <- postResample(predictions, actual)
MSPE <- results["RMSE"]^2
MAE <- results["MAE"]
MAPE <- mean(abs((actual - predictions) / actual), na.rm = TRUE) * 100
print(paste("MSPE:", MSPE))
print(paste("MAE:", MAE))
print(paste("MAPE:", MAPE))

df_north_america <- subset(df_clean, Continent == "Europe",Status=="Developed")
print(df_north_america)

############Random Forest
install.packages('randomForest')
library(caret)
library(randomForest)
set.seed(123)
ctrl <- trainControl(method = "cv", number = 10, verboseIter = FALSE)

model_rf <- train(log_Life_expectancy ~ Continent + Status + log_Adult_Mortality +
                    log_Infant_deaths + log_Percentage_expenditure +
                    log_Measles + log_Diphtheria + log_HIV_AIDS +
                    log_GDP + log_Thinning_5_9_years + log_Income_composition_of_resources +
                    Schooling,
                  data = df_clean,
                  method = "rf",
                  trControl = ctrl,
                  verbose = FALSE)

print(model_rf)

predictions <- predict(model_rf, newdata = df_clean)
actual <- df_clean$log_Life_expectancy

results <- postResample(predictions, actual)
MSPE <- results["RMSE"]^2
MAE <- results["MAE"]
MAPE <- mean(abs((actual - predictions) / actual), na.rm = TRUE) * 100
print(paste("MSPE:", MSPE))
print(paste("MAE:", MAE))
print(paste("MAPE:", MAPE))
```

```r
###############Prediciton in North America
print(df_clean$Continent)
library(caret)
library(randomForest)
library(gbm)

new_data <- data.frame(
  Continent = "North America",
  Status = "Developed",
  log_Adult_Mortality = median(df_clean$log_Adult_Mortality, na.rm = TRUE),
  log_Percentage_expenditure = median(df_clean$log_Percentage_expenditure, na.rm = TRUE),
  log_Measles = median(df_clean$log_Measles, na.rm = TRUE),
  log_HIV_AIDS = median(df_clean$log_HIV_AIDS, na.rm = TRUE),
  log_GDP = median(df_clean$log_GDP, na.rm = TRUE),
  log_Infant_deaths = median(df_clean$log_Infant_deaths, na.rm = TRUE),
  log_Diphtheria = median(df_clean$log_Diphtheria, na.rm = TRUE),
  log_Thinning_5_9_years = median(df_clean$log_Thinning_5_9_years, na.rm = TRUE),
  log_Income_composition_of_resources = median(df_clean$log_Income_composition_of_resources,
  na.rm = TRUE),
  Schooling = median(df_clean$Schooling, na.rm = TRUE)
)

print(new_data)


predictions_mlr <- predict(model_mlr, new_data)
predictions_gbm <- predict(model_gbm, new_data)
predictions_rf <- predict(model_rf, new_data)

print(paste("Predicted log_Life Expectancy in North America using MLR:", exp(predictions_mlr)))
print(paste("Predicted log_Life Expectancy in North America using GBM:", exp(predictions_gbm)))
print(paste("Predicted log_Life Expectancy in North America using RF:", exp(predictions_rf)))
```

REFERENCES

[1] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning*, 2nd. New York: Springer, 2009, pp. 337–384.

[2] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer New York, 2009, ch. Random Forests, pp. 587–604.

[3] S. M. Piryonesi and T. E. El-Diraby, "Data analytics in asset management: Cost-effective prediction of the pavement condition index," *Journal of Infrastructure Systems*, vol. 26, no. 1, p. 04 019 036, 2020.

[4] S. M. Piryonesi and T. E. El-Diraby, "Using machine learning to examine impact of type of performance indicator on flexible pavement deterioration modeling," *Journal of Infrastructure Systems*, vol. 27, no. 1, p. 04 020 035, Feb. 2021.