Table of Contents

[Introduction 2](#_Toc46434929)

[Data: 2](#_Toc46434930)

[Methodology 5](#_Toc46434931)

[Data Compilation 5](#_Toc46434932)

[Imputation - Filling Missing Values 5](#_Toc46434933)

[Choice of Analysis Methods 5](#_Toc46434934)

[Assumptions 5](#_Toc46434935)

[Descriptive Analysis of Variables 6](#_Toc46434936)

[Response Variable: Life Expectancy in Years 6](#_Toc46434937)

[Input Variables 6](#_Toc46434938)

[Categorical independent variables: Region, Income Group 7](#_Toc46434939)

[Results 8](#_Toc46434940)

[ANOVA Model 8](#_Toc46434941)

[ANOVA model using Region as the Classification Variable: 8](#_Toc46434942)

[Initially, Levene’s test for homogeneity for the ANOVA model was used, however, this model failed to satisfy the assumption of equal variances. Thus, we opted to apply Welch’s ANOVA. 8](#_Toc46434943)

[ANOVA model using Income Groups as the Classification Variable 9](#_Toc46434944)

[Regression Analysis 11](#_Toc46434945)

[1) Backward Selection method for Linear Regression. 11](#_Toc46434946)

[2) Evaluating the fit of the Model. 12](#_Toc46434947)

[3) Checking for Correlation, Collinearity and Outliers. 15](#_Toc46434948)

[4) Scoring and Validating the model 17](#_Toc46434949)

[Conclusion 18](#_Toc46434950)

[Additional Information and Future Considerations 18](#_Toc46434951)

[Appendices: 19](#_Toc46434952)

[APPENDIX A: Example - For missing values in Tuberculosis incidents 19](#_Toc46434953)

[APPENDIX B: Summary Statistics and Plots of Life Expectancy 19](#_Toc46434954)

[APPENDIX C: Summary Statistics of Input Variables 20](#_Toc46434955)

[APPENDIX D: Initial ANOVA with Region as the Classification Variable 20](#_Toc46434956)

[APPENDIX E: Summary Statistics of ANOVA with Income Group as Classification Variable 21](#_Toc46434957)

[APPENDIX F: Summary of Variables Removed from Each Iteration 22](#_Toc46434958)

[APPENDIX G: Results of Backwards Elimination – All Indicators 23](#_Toc46434959)

[APPENDIX H: Metrics of Finalized Model 24](#_Toc46434960)

# Introduction

Over the past century, it has been seen that life expectancy around the world has greatly improved. The average life expectancy in 1950 was only 46 years of age. In 2015, the global average life expectancy is 71 years old [(Roser et al., 2019)](https://ourworldindata.org/life-expectancy#how-did-life-expectancy-change-over-time).

This research explores the effects of indicators across health, economic, environmental and educational factors on life expectancy across the globe. Life expectancy at birth is defined as the "number of years a newborn infant could expect to live if prevailing patterns of age-specific mortality rates at the time of birth stay the same throughout the infant's life” [(United Nations Development Programme[UNDP], 2015)](http://hdr.undp.org/sites/default/files/hdr_2015_statistical_annex.pdf). It is commonly viewed as a key measure of populations' health [(Australian Government Department of Health, 2012)](https://www1.health.gov.au/internet/publications/publishing.nsf/Content/oatsih-hpf-2012-toc~tier1~life-exp-wellb~119). Using significant indicators, our research goal was to explore the life expectancy of a new-born child.

The findings of this report aim to be useful for policy-makers and aide organizations that address the health of new-born children. The inclusion of diverse factors, noted above, will provide a comprehensive approach to better understand the role that certain indicators have on the health and longevity of a new-born child. This information will provide a better understanding for addressing key issues to increase global life expectancy.

# Data:

Data for 217 countries in the world were obtained from the World Bank's agencies. The response variable is the total life expectancy at birth in years.

The indicators included in the research span across economic, educational, environmental and health factors.  These factors in addition to the economic factor, were chosen as they were included in the Human Development Report Office's Human Development Report for the UNDP in 2015. It explored educational, environmental and health factors through the human development approach - which focused on multiple factors that improve life rather than rely on the assumption that solely economic growth leads to greater wellbeing [(HDR UNDP)](http://hdr.undp.org/en/humandev).

**Economic:**

***GDP per capita*** measures the country's economic output by its midyear population. This indicator was selected as GDP per capita shown prior positive association with life expectancy [(Messias, 2003)](https://www-ncbi-nlm-nih-gov.ezproxy.library.yorku.ca/pmc/articles/PMC1447959/). ***Waged and Salaried Workers*** (as a percentage of total employment) are workers who hold jobs that are paid employment jobs with contracts that give basic remuneration that is not commission-based. This indicator was chosen because a high proportion of wage and salaried employees in a nation can signify advanced economic development [(World Bank, 2020)](https://data.worldbank.org/indicator/SL.EMP.WORK.ZS).

**Environment:**

***Access to electricity*** (as a percentage of the population) was chosen as an indicator because the maintenance of secure electrical services is necessary for creating conditions that are conducive for economic growth and stability [(World Bank)](https://data.worldbank.org/indicator/EG.ELC.ACCS.ZS). Additionally, close associations have been found between household electricity and life expectancy at birth [(Wang, S. et al., 2019)](https://www-sciencedirect-com.ezproxy.library.yorku.ca/science/article/pii/S0360544219307327?via%3Dihub). The ***PM2.5 air pollution*** measures the mean annual exposure (of micrograms per cubic meter) is the average level of exposure that a country's population. This indicator was chosen due to prolonged exposure to the concentrations of these fine particles of air pollution are capable of causing severe health damage through penetrating the respiratory tract that can lead to diseases [(Brauer, M et al., 2018)](https://pubs-acs-org.ezproxy.library.yorku.ca/doi/10.1021/acs.est.8b02864).The ***percentage of the population that uses at least basic sanitation services*** was chosen as an indicator because sanitation is fundamental to human development. With poor sanitation linked to the transmission of diseases [(World Health Organization, 2019)](https://www.who.int/news-room/fact-sheets/detail/sanitation), hygiene facilities for basic sanitation have been used to fight against poverty, disease and death.

**Health:**

The ***prevalence of anemia among children*** (the percentage of children under 5 years old) was chosen because anaemia in infancy are at risk of cognitive, socio-emotional and adaptive development difficulties [(Abu-Ouf and Jan, 2015)](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4375689/). ***Immunization of diptheria-tetanus-pertussis (DPT)*** measures the percentage of children from 12-23 months old who were vaccinated before 12 months old. This indicator was chosen because this vaccine protects infants against infectious diseases which can cause serious, or even fatal illnesses and disabilities [(World Health Organization, 2020)](https://www.who.int/en/news-room/fact-sheets/detail/immunization-coverage).  ***Immunization of measles*** measures the percentage of children 12-23 months old who received the measles vaccine before being 12 months old. This indicator was chosen as measles is a highly contagious viral disease that can lead to high fever, blindness, or even death [(World Health Organization, 2020)](https://www.who.int/en/news-room/fact-sheets/detail/immunization-coverage). ***Mortality Caused by Road Traffic Injury (per 100,000 people)*** was chosen because road traffic injuries are the leading cause of death for children and youth that are 5-29 years old [(World Health Organization, 2020)](https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries). The ***cause of death by non-communicable diseases*** is a percentage of the total deaths for all ages. This includes cancer, diabetes, mellitus, cardiovascular diseases, digestive diseases, skin diseases, musculoskeletal diseases, and congenital anomalies (World Health Organization). This indicator was chosen because non-communicable diseases are leading causes of premature disability and death globally [(Licher et al., 2019)](https://pubmed.ncbi.nlm.nih.gov/30716101/)***.*** The ***Incidence of tuberculosis (per 100,000 people)*** indicator was chosen because tuberculosis is an air-borne bacterial disease that can be fatal. It is one of the top ten causes of death in the world. [(World Health Organization, 2020).](https://www.who.int/news-room/fact-sheets/detail/tuberculosis) The ***Current Health Expenditure (% of GDP)*** was chosen because countries where health spending grew more rapidly also had rapid increases in life expectancy [(OECD, 2011)](https://www-oecd-ilibrary-org.ezproxy.library.yorku.ca/docserver/soc_glance-2011-en.pdf?expires=1595387660&id=id&accname=ocid194763&checksum=3E9186C64764F39A27BD325D412AE8A4). The ***total alcohol consumption per capita*** is the total litres of pure alcohol consumed by each person that is over 15 years old. This indicator was chosen because alcohol consumption was identified as a causal factor in over 200 disease and injury conditions by the World Health Organization [(World Bank)](https://data.worldbank.org/indicator/SH.ALC.PCAP.LI).

**Education**:

The ***percentage of primary school enrollment*** is the gross enrollment ratio of primary school which indicates the capacity level of the education system in a nation. This indicator was chosen because primary education teaches basic literacy and mathematical skills, to which there is a strong relationship between educational attainment and life expectancy [(Steingrímsdóttir, Ó et al., 2012)](https://link-springer-com.ezproxy.library.yorku.ca/article/10.1007/s10654-012-9663-0).

In addition to literature review, these factors were chosen based on the completeness of the dataset. All indicators were screened to ensure that at least 75% of the data points were present.

# Methodology

## Data Compilation

To prepare the dataset, the data tables of all aforementioned indicators were extracted from the World Bank's DataBank. In addition, the Regional and Income Group classification was also extracted from the DataBank as it classifies countries by region and income levels.

The data tables were merged into a single file for the year 2016 for the training data set and year 2010 for the test data set.

## Imputation - Filling Missing Values

The data sets from the World Bank were not fully complete. Due to the absence of several missing values, imputation was conducted for each of the indicators. For each of the numeric indicators, average values of the regions were calculated. After which, the regional average was used to impute any missing value based on the region the country belonged to (refer to Appendix A for example). This method was applied for both the training data and the test data sets.

## Choice of Analysis Methods

With the data set containing two categorical variables and fourteen numeric variables, the ANOVA and linear regression models were used. The ANOVA was used to analyze if there are any statistically significant differences in the mean life expectancy values of the different region and income group classes. After which, an explanatory linear regression model was used to estimate the life expectancy value of a given country. Through the regression model, the indicators and their respective associations that are most impactful on life expectancy were identified. Backward model selection was used as it gives the best result (Anton, MBAN 5110 - Lecture 9, 2020). The regression model was then tested over a test dataset to measure its explain-ability.

## Assumptions

In order to impute the missing values of an indicator with respect to a given country, the respective regional average value was used. The assumption made is that the regional average for a particular indicator is close to the actual value of the indicator.

# Descriptive Analysis of Variables

## Response Variable: Life Expectancy in Years

In the distribution of life expectancy, among 217 observations, the median life expectancy in 2016 has reached 74.17. However, the mean is lower due to the presence of outliers below the first quartile (Central African Republic and Lesotho with life expectancies of 51.593 and 52.059 respectively). The overall distribution is left skewed, bimodal and the s-shaped normal probability line suggests that the life expectancy does not follow normal distribution (refer to Appendix B for diagrams).

## Input Variables

To get a prior idea about the distribution of each independent variable, the box and whisker plot standardized using mean is presented below.

ox Plots Of Standardized Interval Variables 
1 , 0 
0 0 
0 
0 
0 


The presence of outliers is present in ten of the fourteen variables. Upon visual observation, the variables of Alcohol Consumption per Capita, Gross Enrollment in Primary school, and Mortality Caused by Road Traffic have normal distributions. (Refer to Appendix C for summary statistics).

## Categorical independent variables: Region, Income Group

egion 
Region 
East Asia & Pacific 
Europe & Central Asia 
Latin America & Caribbe ncome Group 
Income_group 
High income 
Low income 
Lower middle income 


It was found that the groups divided by region are not balanced, with only 3 countries in North America and 8 in South Asia, both of them only accounting for less than 5% of the total countries.

The variable income group also divides the whole dataset into unbalanced groups, with more than 60% belonging to either high income or upper middle income group and less than 40% in the remaining two lower income groups.

These class variables were not used in linear regression since their values were used to impute the missing values of the numeric variables.

# Results

## ANOVA Model

ANOVA models were generated to observe if there are any significant differences in the mean life expectancy amongst varying income levels and regions in the world.

### ANOVA model using Region as the Classification Variable:

### Initially, Levene’s test for homogeneity for the ANOVA model was used, however, this model failed to satisfy the assumption of equal variances. Thus, we opted to apply Welch’s ANOVA.

The summary statistics using region as an classification variable suggests that there appears to be a difference among the life expectancies of different regions.

ifeExpectancy 
Level of 
Region 
East Asia & Pacific 
Europe & Central Asia  istribution of LifeExpectancy 
61.30 
Prob > F <.OOOV 
80 
70 
50 

As mentioned previously, the equal variance assumption does not hold among these region groups in the initial Levene's Homogeneity test, thus Welch’s ANOVA test is a more appropriate option as it does not assume equal variances. The result shows a significant P-value under Welch’s ANOVA test, meaning that there is enough evidence to reject null hypothesis, thus there is at least one group mean that is statistically significantly different from the other group means.

elch's ANOVA for LifeExpectancy 
Source 
Region 
Error 
DF 
6.0000 
24.1113 
F Value F 

The Tukey-Kramer adjustment was then applied in the ANOVA model. Based on the diffogram, there appears to be a significant difference between Sub-Saharan Africa and all other regions, South Asia and Europe & Central Asia, South Asia and North America, East Asia & Pacific.

0 
70 
60 
LifeExpectancy Comparisons for Region 
Arneri

### ANOVA model using Income Groups as the Classification Variable

The summary statistics and box plots with income group as a categorical variable suggests that there appears to be a difference among the life expectancy of varied income groups. Countries with high and upper-middle income seem to enjoy a longer life, not contradictory to our group’s intuition. (Refer to appendix D summary statistics).

istribution of life_expectancy_clean 
80 
70 
60 
50 
High income 
L

The F-value and P-value suggests that there is a significant difference in mean life expectancy among the income groups. The diffogram further illustrates that differences between *all levels* are significant. This may mean that economic factors may have a large contribution to life expectancy.

ource 
Model 
Error 
Corrected Total 
DF 
3 
213 
216 
Sum of Squares 
7942.64322 


ife_expectancy Comparisons for Income_group 
80 
75 
70 
65 
60 
60 
Lower 'Icon* 
incon" 
Upper 'Icon* 
Low

## Regression Analysis

The goal of the multiple regression model is to estimate the life expectancy value for a given country and explain the effects of the different input variables.

The regression model was built using the backward selection technique. Afterwards, linear relation assumption and model fit evaluation was conducted. Next, outliers and multicollinearity were checked. Finally, the model was scored and validated.

The model parameter threshold of α = 0.05 was used.

### Backward Selection method for Linear Regression.

In the first iteration of the model in which all the variables were present, the R squared value is 0.8701 but there are several variables in the model which were not significant (refer to Appendix F for variables removed in each iteration). The model ran for eight more iterations. In each iteration, the variable with the highest p-value above 0.05 was eliminated.

In the final iteration, only six variables remain. Tuberculosis incidence (per 100k), DPT Vaccination (in children aged 1-2 years), GDP per capita, access to electricity, non-communicable disease deaths (% of total deaths), Anemia (in children aged under five).

Contrary to our group's initial intuition, basic sanitation, health care expenditure, and pollution were not statistically significant to explaining life expectancy.

### 

### 2) Evaluating the fit of the Model.

In the beginning, it was assumed that the predictor variables have a linear relation with target variable of life expectancy. To validate this, we assessed the following:

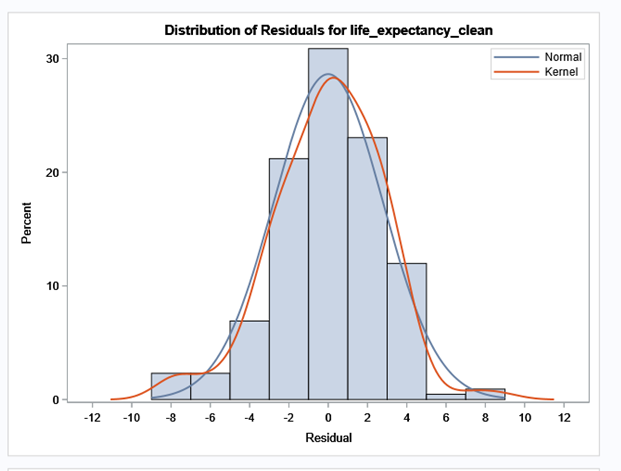
1. ***Equal Variance, Sum of residuals = 0 and no diverging or converging pattern formation.***

esidualbyPredictedforlife_expectancy_clean 
0 。 00 0 
Predicted Value 

From the residuals plot, we can see that they are randomly scattered across the value of zero. There is no converging or diverging patterns the variance is fairly equal across the zero-reference line.  This implies that the assumption that the sum of residuals is equal to 0, is satisfied.

1. ***The residuals are normally distributed across zero.***

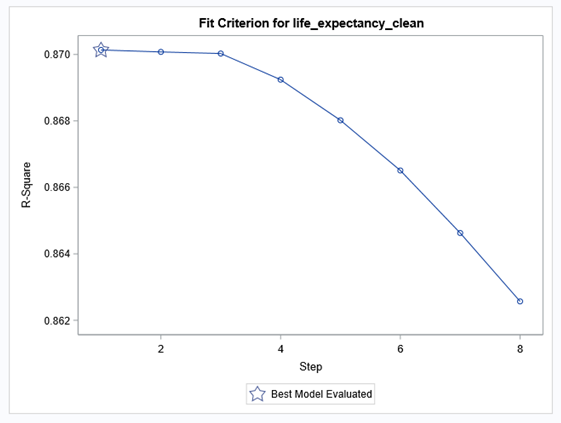
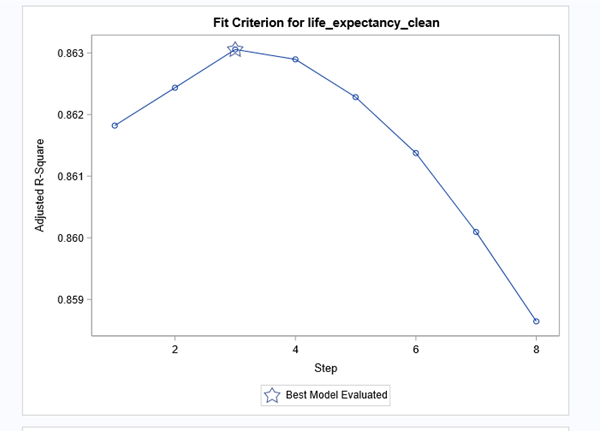
The QQ plot and the residuals distribution histogram plot were generated to analyze distribution.

 -Q Plot Of Residuals for 
000 
0 00 
Quantile 

The QQ plot closely follows the diagonal straight line for the residuals. Also, the distribution of residuals in the histogram resembles a normal distribution very closely.  Hence the assumption of normality in residual distribution is also satisfied.

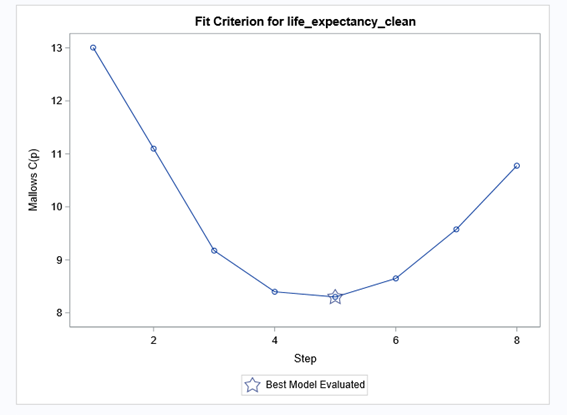
1. ***Model Fit***

R2 and Adjusted R2

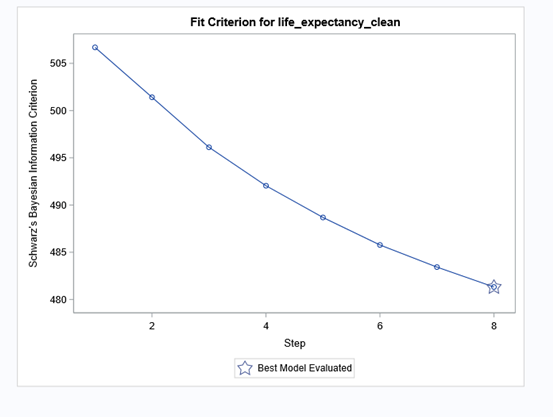
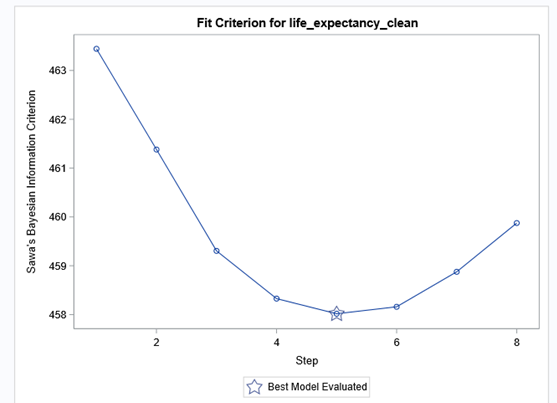
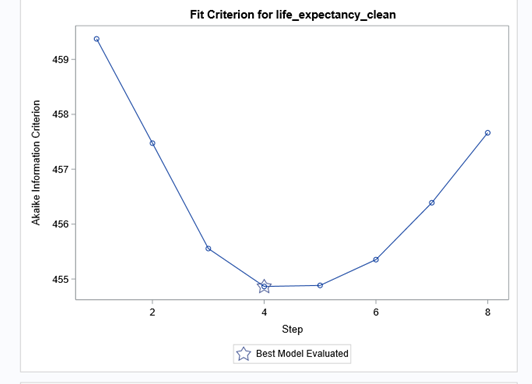
For the final iteration, the R2 and adjusted R2 are 0.85 and 0.86 respectively. When adjusted for the number of input variables, the model accounts for 86% of variability in the data. The very high adjusted R2 with all statistically significant values suggests that the final model accounts for 86% of variability in the data, meaning the regression model fits the data well.

Mallow’s Cp



However, it should be noted that the lowest value of Mallows Cp was for the fifth iteration where it reached between 8-9. This criterion was not perfectly met.

Information Criteria



AIC BIC SBC

Amongst the information criteria, the smaller the values, the better. The Akaike Information Criterion (AIC) is a fined technique based on in-sample fit to estimate the likelihood of a model to predict/estimate the future values. A good model is the one that has minimum AIC among all the other models. The values for AIC in our model are fairly low. However, the Schwarz Bayesian Information Criterion (SBC) was referenced in to choosing the finalized model in conjunction with observing the variables in which the p-values were statistically significant.

### 3) Checking for Correlation, Collinearity and Outliers.

#### Correlation

earson Correlation Coefficients, N = 217 
LifeExpectancy 
Life Expectancy 
Prob 

From the Pearson Correlation Coefficients table, all predictor variables except Gross Primary School Enrollment appear to have a strong relationship with the Life Expectancy variable with the absolute value of correlation value larger than 0.3 and P-value less than 0.0001.

Specifically, four variables seem to have a negative relationship with life expectancy, namely anemia incidences, mortality due to road traffic, Tuberculosis incidences per 100,000, and pollution (PM2.5) which fit our intuition.  As the prevalence of anemia among children, mortality due to road traffic injury, incidence of Tuberculosis and pollution (PM2.5) increases, the overall life expectancy suffer a decrease.

However, the strong positive relationship between Non-Communicable Disease and life expectancy is quite mysterious at this stage, a explanation is that non-communicable disease, such as cancer, cardiovascular diseases, chronic obstructive pulmonary diseases, are more likely to occur the longer people age.

Positive relationships between life expectancy and basic sanitation, access to electricity, wage and salaried employees, immunization for measles, and GDP per capita are also aligned with intuition. However, possible collinearity needs to be addressed.

#### Collinearity

Collinearity was analyzed in the model using the **Variance Inflation Factor (VIF)**. VIF is a measure of the amount of multicollinearity in a set of multiple regression variables. Mathematically, the VIF for a regression model variable is equal to the ratio of the overall model variance to the variance of a model that includes only that single independent variable. A VIF that is greater than 10 indicates collinearity, presence of collinearity can be ignored if the VIF is less than 10.

arameter E tin 
DF 
tub 0 i $ - p -100k - ~ n 
gdp—per_capita_clean 
~ cce 一 ~ 0

Since all VIF values are well below 10, it can be concluded that there is no presence of multicollinearity in our model.

#### Outliers

The following criteria were used to assess outliers in the model:

* **|Studentized residuals| > 3**

since the model has a relatively large number of observations, the studentized residuals cutoff of three is most suitable (SAS Institute, 2019).

* **|DFFITS|  > -**  the cut off point for the measure of impact of the ith observation being influential
* Cook's D  > : this is the cut off point for Cook's D, which measures the simultaneous change in the parameter estimates when the ith observation is deleted from the analysis.

For each of these criterion, a flag was created in the dataset and found the following values had one or more flags being raised. In the table of outliers, most of the countries are very small nations.

1 Co 4- vo 一 「 
9 Le 0 0 
8 Y—-., Rep 
5 C30 SAR,

Monaco was deleted from the dataset since it flagged all three criteria for being outliers.  For now, we chose to let the other values in the dataset remain there because they did not flag all three criteria.

### 4) Scoring and Validating the model

To score and validate the model, a test data set was prepared based on data for the indicators during year 2010. The same data preparation steps to impute the values using regional averages was used in preparation of the test set.  Using this data, the same linear regression model on the training set was applied to the test (validation) dataset. After which, the model was scored and validated through external validation.

ouanbes て a と 3 
bu を 朝 」 1 ー ー 0 ー ー 
品 あ を も s 
期 リ pambs a6eaenv 

The plot above shows the mean squared error (MSE) for the test and train datasets. The test dataset has a higher MSE, however it did not drastically increase with more iterations. The value for MSE for validation dataset is very similar to the trend of the training set, meaning that we did not overfitting our dataset by eliminating variables that are statistically significant using a backward selection method.

# Conclusion

The final explanatory regression model:

**Life Expectancycountry** =

There are significant differences in life expectancy between regions and income groups. The six most important factors affecting life expectancy are health, environmental, and economic (refer to Appendix H for full summary statistics of the finalized model). The indicators in relation to the health factor are Tuberculosis Incidences, Anemia Incidences, DPT vaccination, and Non-Communicable Diseases. The environmental indicator that was significant is Access to Electricity. The significant economic factor is GDP per capita.

Tuberculosis and Anemia incidences have a strong negative effect on the value of life expectancy.

GDP per capita, Access to electricity, portion of non-communicable disease deaths, DPT vaccines have a strong positive effect on the value of life expectancy.

# Additional Information and Future Considerations

As life expectancy consists of a multitude of factors, going forward, health care policy makers and aid organizations may desire to include other pertinent factors specific to their cases into the model. The data used is all from the World Bank and some of the other indicators that may contribute to correlation have too many missing values. More meaningful datasets from other sources could be consulted by the audience.

# Appendices:

## APPENDIX A: Example - For missing values in Tuberculosis incidents

Missing Tuberculosis Incidents Data

|  |  |
| --- | --- |
| Region | Regional Average |
| East Asia & Pacific | 32.57777778 |
| Europe & Central Asia | 21.51836735 |
| Latin America & Caribbean | 30.88787879 |
| Middle East & North Africa | 31.44285714 |
| North America | 8.95 |
| South Asia | 44.2 |
| Sub-Saharan Africa | 59.02083333 |

## APPENDIX B: Summary Statistics and Plots of Life Expectancy

ummary Statistics for Dependent Variables 
Analysis Variable : LifeExpectancy Life 

istribution of Life Expectancy 
80 
60 
52.059 8 .593 
50  istribution of life_expectancy 
30 
25 
20 
15 
10 
5 
Summary Stati robability Plot for life_expectancy 
90 
80 
70 
60 
0 00 
50 
5 
10 
Normal Lin

## APPENDIX C: Summary Statistics of Input Variables

ummary Statistics for Interval Variables 
Variable 
AccessToElectricity 
AlcoholPer

## 

## APPENDIX D: Initial ANOVA with Region as the Classification Variable

The summary statistic using region as an input variable suggests that there seems to be a difference along the life expectancy of different regions, and the box plot shows the same result in a more straightforward way.

ife_expectancy_clean 
Level of 
Region 
East Asia & Pacific 
Europe & Central Asia 
Latin America & Caribbea  
Sub-Saharan A' 
South Asia 
North Amenca 
Middle East & Nom Africa 
Labn Amenca & Caribbean 
Europe & Central

The F-value and P-value show that there is a significant difference in mean life expectancy among 6 regions.  More specifically, based on the diffogram shown, there seems to be a significant difference between Sub-Saharan Africa and all other regions, South Asia and Europe& Central Asia, South Asia and North America, East Asia & Pacific and Europe and Central Asia when Tukey-Kramer adjustment applied in the ANOVA model.

ource 
Model 
Error 
Corrected Total 
DF 
6 
210 
216 
Sum of Squares 
7765.22970 


0 
70 
60 
life_expectancy_clean Comparisons for Region 

Even though the normality assumption is held in the assumption test, the result of both Levene’s test and Welch’s test show that our model fails the variance equality assumption, which means the result of our ANOVA test is not reliable enough

evene's Test for Homogeneity of life_expectancy_clean Variance 
ANOVA of Squ

## 

## APPENDIX E: Summary Statistics of ANOVA with Income Group as Classification Variable

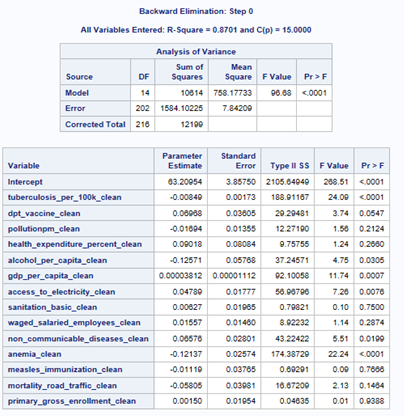
evel of 
Income_group 
High income 
Low income 
Lower middle income 
Upper m

## 

## APPENDIX F: Summary of Variables Removed from Each Iteration

-2106 , 112.74 
Α" model M O .05M 
Backwud 
p .rtial 


## APPENDIX G: Results of Backwards Elimination – All Indicators



## APPENDIX H: Metrics of Finalized Model

