Predicting Life Expectancy Across the Globe

Daniel Vasilevko

Introduction

This analysis seeks to create model that can accurately predict the life expectancy of a country based on data regarding rates of immunization, levels of malnutrition, economic prosperity, mortality rates and other variables associated with quality of life. The dataset used was taken from the World Health Organization and contains information from 179 countries across 15 years (2000-2015).

Country: The name of country from which data is coming from. 179 Countries

Region: The general region of a given country. 9 regions

Year: The year the data was collected. 2000-2015.

Infant_deaths: The number of infant deaths per 1000 live births.

Under_five_deaths: The number of deaths of children under 5 per 1000 live births.

Adult_mortality: The number of adults(age 15-59) dying before age 60 per 1000 adults.

Alcohol_consumption: The amount of liters of Alcohol consumed per capita (age 15+).

Hepatitis_B: Percentage of 1 year olds who received Hepatitis_B-B immunization.

Measles: Percentage of 1 year olds who received Measles immunization.

Polio: Percentage of 1 year olds who received Polio immunization.

Diphtheria: Percentage of 1 year old who received Diphtheria immunization.

Incidents_HIV: Incident of HIV per 1000 people (age 15-49).

BMI: Average BMI. BMI is defined as weight(k) / sqrt(height(m)).

GDP_per_capita: The GDP per capita in USD (United States Dollars).

Population_mln: The population in millions.

Thinness_ten_nineteen_years: Percentage of thinness among adolescents (age 10-19). Thinness defined as BMI < -2 standard deviations below the median.

Thinness_five_nine_years: Percentage of thinness among children aged 5 or lower. Thinness defined as BMI < -2 standard deviations below the median.

Schooling: Average years that people aged 25+ spent in formal education.

Economy_status_Developed: Whether a country is considered "developed".

Economy_status_Developing: Whether a country is considered "developing".

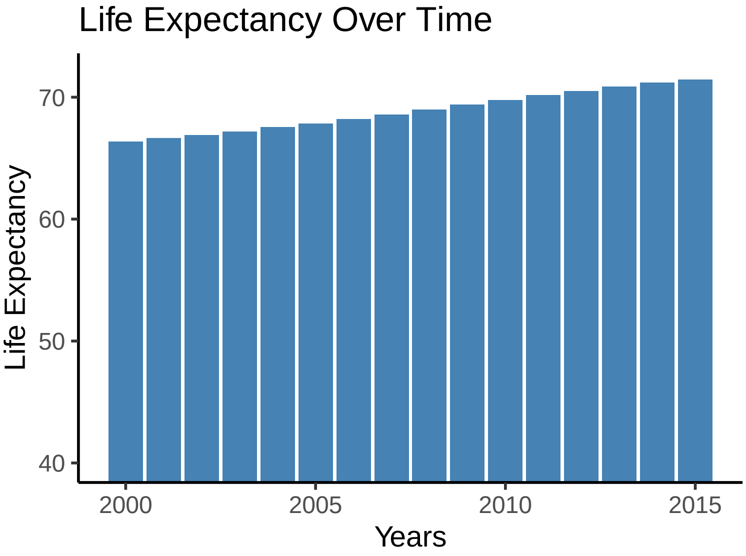
Life_expectancy: Average life expectancy.

Fitting the Model

With the dataset introduced, the proccess of selecting variables to put in the model can begin. Looking at these variables logically there are already a few that need to be pruned.

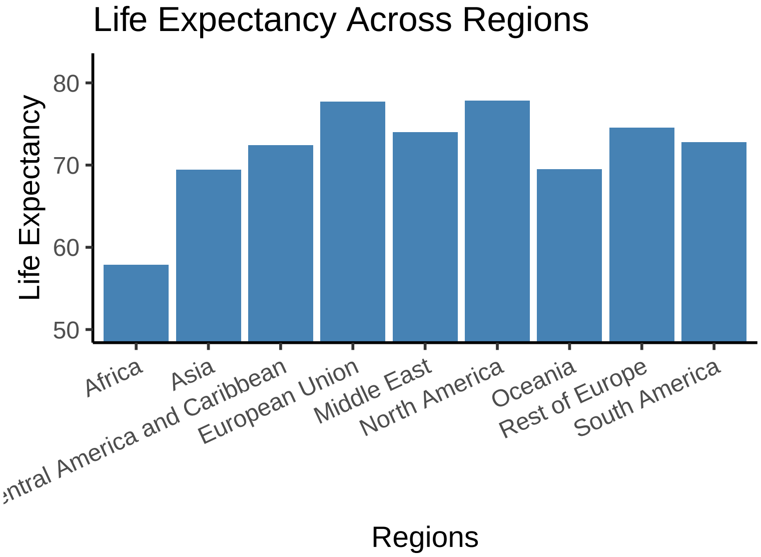
While ‘Economy_status_Developed’ and ‘Economy_status_Developing’ offer potential insight into the prosperity of a specific country, the formula used to determine if a country is “developed” or not was not provided, making this variable unreproducible by studies not sourcing data directly from the World Health Organization.

On paper ‘Year’ may have some level of correlation with ‘Life_expectancy’ with standards of living generally rising over time.



However the passage of time does not directly lead to people living longer. ‘Years’ merely captures a trend of life expectancy tending to rise due to other factors such as better access to healthcare or financial stability. In an MLR we want variables with a causal relationship with the response variable not a reflection of general trend. ‘Years’ will not be included in the MLR model but will still be used to better categorize the data.

The variables ‘Region’ and ‘Country’ are to ‘Year’. Both ‘Region’ and ‘Country’ each have some correlation with a life expectancy.



However it is the conditions within the ‘Country’ and ‘Region’ that causes the change in life expectancy rather than the name of a ‘Country’ or ‘Region. Additionally, this model seeks to work based on numerical data regardless of the time the analysis is conducted. Countries are not stagnant so locking the analytical ability of this model behind a preset 179 countries would be designing for obsolescence. ’Country’ will remain as a method of better categorizing data but will not be used in the MLR model and ‘Region’ will be removed entirely.

In other regards the data is clean, with 0 NA values across the 17 remaining columns and a consistent and easily readable naming scheme for each variable.

```
##                                  Column NA_Count
## Country                         Country        0
## Year                               Year        0
## Infant_deaths             Infant_deaths        0
## Under_five_deaths     Under_five_deaths        0
## Adult_mortality         Adult_mortality        0
## Alcohol_consumption Alcohol_consumption        0
```

The remaining variables are suited to be used in a MLR model.

Fitting A Multiple Linear Regression Model

The data set has been trimmed down to contain variables that could function well in creating an MLR. From there we will further clean the data by getting rid of any non-significant variables.

```
## 
## Call:
## lm(formula = Life_expectancy ~ Infant_deaths + Under_five_deaths + 
##     Adult_mortality + Alcohol_consumption + Hepatitis_B + Measles + 
##     BMI + Polio + Diphtheria + Incidents_HIV + GDP_per_capita + 
##     Population_mln + Thinness_ten_nineteen_years + Thinness_five_nine_years + 
##     Schooling, data = data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -4.7675 -0.9412 -0.0560  0.8826  7.9078 
## 
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                  8.507e+01  6.114e-01 139.141  < 2e-16 ***
## Infant_deaths               -5.152e-02  6.187e-03  -8.328  < 2e-16 ***
## Under_five_deaths           -5.295e-02  3.829e-03 -13.830  < 2e-16 ***
## Adult_mortality             -4.893e-02  6.147e-04 -79.604  < 2e-16 ***
## Alcohol_consumption          9.023e-02  8.950e-03  10.081  < 2e-16 ***
## Hepatitis_B                 -9.262e-03  2.560e-03  -3.618 0.000302 ***
## Measles                      1.228e-03  1.728e-03   0.711 0.477264    
## BMI                         -1.674e-01  1.891e-02  -8.854  < 2e-16 ***
## Polio                        8.691e-04  5.876e-03   0.148 0.882421    
## Diphtheria                   3.257e-03  5.926e-03   0.550 0.582629    
## Incidents_HIV                1.005e-01  1.827e-02   5.502 4.08e-08 ***
## GDP_per_capita               3.043e-05  2.146e-06  14.181  < 2e-16 ***
## Population_mln              -1.814e-04  2.011e-04  -0.902 0.366939    
## Thinness_ten_nineteen_years -3.313e-02  1.725e-02  -1.921 0.054867 .  
## Thinness_five_nine_years    -2.449e-03  1.690e-02  -0.145 0.884797    
## Schooling                    1.119e-01  1.668e-02   6.709 2.35e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.369 on 2848 degrees of freedom
## Multiple R-squared:  0.9789, Adjusted R-squared:  0.9788
## F-statistic:  8817 on 15 and 2848 DF,  p-value: < 2.2e-16
```

From the P-values shown above we find 6 variables that are not needed in the model. Measles, Polio, Diphtheria, Population_mln, Thinness_ten_nineteen_years, and Thinness_five_nine_years all have a P-value above .05 meaning they do not accurately explain the behavior of 'Life_expectancy'.

To further cut any unnecessary variables from the model it is a good idea to test for multi-collinearity. Although 'Infant_deaths' and 'Under_five_deaths' are both significant, logically both are very likely to have multi-collinearity as they measure mortality range over essentially the same age range. To see the variables with multi-collinearity we will find each variables Variance Inflation Factor (VIF) which essentially shows how much of the variable's relevance is casused by collinearity in the model.

```
##       Infant_deaths   Under_five_deaths     Adult_mortality Alcohol_consumption
##           43.501866           41.931188            7.468546            1.856158
##         Hepatitis_B                 BMI       Incidents_HIV      GDP_per_capita
##            1.420358            2.009466            2.790705            1.985467
##           Schooling
##            4.110131
```

The normal cut off for large multicollinearity is a VIF larger than 10, so suffice to say 'Infant_deaths' and 'Under_five_deaths' are definitely above the cutoff. Since 'Infant_deaths' has a higher VIF than 'Under_five_deaths' and has a slightly higher Standard-Error it will be the variable to be removed.

```
##   Under_five_deaths     Adult_mortality Alcohol_consumption         Hepatitis_B
##            6.427603            7.399285            1.758321            1.396022
##                 BMI       Incidents_HIV      GDP_per_capita           Schooling
##            2.008984            2.763603            1.912481            4.106971
```

The remaining 2 variables with moderate MC are 'Under_five_deaths' and 'Adult_mortality'. Since the 2 variables only have moderate multicollinearity and the age ranges of the 2 mortality rates do not intersect for now neither variable will be removed.

Outliers

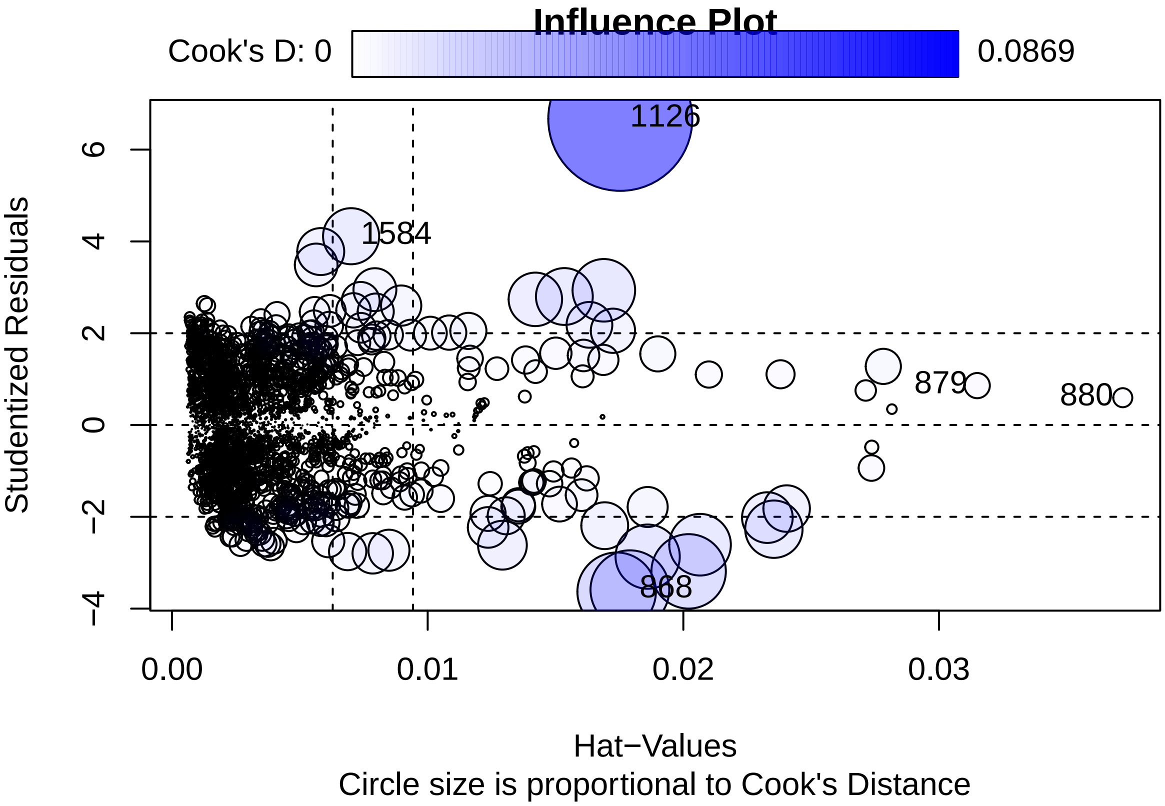
The presence of outliers can greatly skew data and can have disproportionate effects of the inferences made by a model.

Using the built in outlier finder in R we can find all outlier in the data set.

```
##      rstudent unadjusted p-value Bonferroni p
## 1126 6.670251         3.0551e-11   8.7498e-08
```

With an Bonferroni P-value of 8.7498e-08, under the assumtions of our model, the chances of this data point to exist due to random chance is extremely low, thus making it an outlier.

This built in function only finds the most egregious outliers so it is a good idea to manually look for any outliers in the data. A good way to manually find outliers is by looking at 3 different values. Leverage measures how far an observations predictor values are from the mean Standardized Residuals measure how far an observations response values are from the mean. Cooks Distance measures how much influence a given point has on a variable.

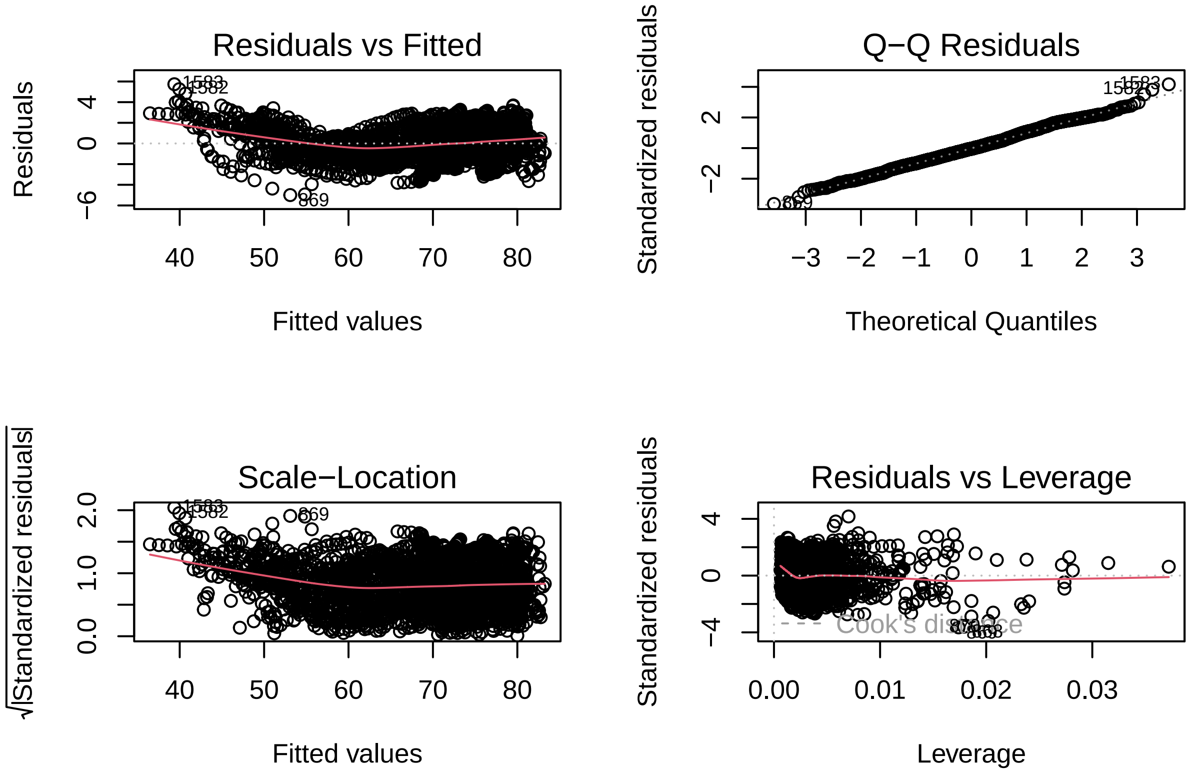


```
##          StudRes         Hat       CookD
## 868  -3.5877956 0.017909032 0.025973534
## 879   0.8585213 0.031495986 0.002663504
## 880   0.5950417 0.037189710 0.001519962
## 1126  6.6702513 0.017534435 0.086906108
## 1584  4.1131168 0.007004009 0.013185122
```

With the influence plot we can clearly see that that obsetvation 1126 exudes a high level of influence on the data while having a high residual value meaning the point should be removed.

Assumptions of Multiple Linear Regression

Using a set of plots we can confirm assumptions of Linearity, Normality of Residuals, Homoscedasticity and the model was checked for outliers.



Residuals vs Fitted Plot: The line is approximately straight and the errors tend to cluster around 0 indicating that the model satisfies the assumption of linearity.

Normal Q-Q Plot: The points follow a straight line indicating that residuals are normally distributed meaning the model satisfies the assumption of normality of errors.

Scale-Location Plot: The line is approximately straight and has no clear pattern indicating that model satisfies the assumption of homoscedasticity.

Residuals vs Leverage: No points with high leverage have a high standardized residual meaning there it is confirmed there are no outliers in the data set.

Model Selection

With all assumptions of Multiple Linear Assumption confirmed the best model can be selected. To find the best model Backwards Stepwise Regression can be used which takes away variables from the model to try to create the most optimized model. The test used to compute the quality of the model was Akaike Information Criterion which rewards good fit while penalizing high complexity.

```
## Start:  AIC=1861.35
## Life_expectancy ~ Under_five_deaths + Adult_mortality + Alcohol_consumption +
##     Hepatitis_B + BMI + Incidents_HIV + GDP_per_capita + Schooling
##
##                       Df Sum of Sq    RSS    AIC
## <none>                              5450.6 1861.4
## - Hepatitis_B          1      11.9  5462.5 1865.6
## - Incidents_HIV        1      60.5  5511.1 1891.0
## - Schooling            1     101.8  5552.4 1912.3
## - BMI                  1     120.7  5571.3 1922.1
## - Alcohol_consumption  1     348.0  5798.6 2036.6
## - GDP_per_capita       1     535.2  5985.8 2127.5
## - Under_five_deaths    1    6051.6 11502.2 3997.5
## - Adult_mortality      1   12142.5 17593.1 5214.2
```

```
##
## Call:
## lm(formula = Life_expectancy ~ Under_five_deaths + Adult_mortality +
##     Alcohol_consumption + Hepatitis_B + BMI + Incidents_HIV +
##     GDP_per_capita + Schooling, data = data)
##
## Coefficients:
##         (Intercept)    Under_five_deaths      Adult_mortality
##           8.358e+01           -8.338e-02           -4.890e-02
## Alcohol_consumption          Hepatitis_B                  BMI
##           1.161e-01           -4.764e-03           -1.327e-01
##       Incidents_HIV       GDP_per_capita            Schooling
##           1.016e-01            3.532e-05            1.206e-01
```

As shown above, every variable taken from the model resulted in worse performance, meaning the base model was the best. The influence of each variable is then observed with ANOVA.

```
## Anova Table (Type II tests)
##
## Response: Life_expectancy
##                      Sum Sq   Df   F value    Pr(>F)
## Under_five_deaths    6051.6    1 3168.6903 < 2.2e-16 ***
## Adult_mortality     12142.5    1 6357.9555 < 2.2e-16 ***
## Alcohol_consumption   348.0    1  182.2234 < 2.2e-16 ***
## Hepatitis_B            11.9    1    6.2329    0.0126 *
## BMI                   120.7    1   63.1947 2.672e-15 ***
## Incidents_HIV          60.5    1   31.7032 1.971e-08 ***
## GDP_per_capita        535.2    1  280.2225 < 2.2e-16 ***
## Schooling             101.8    1   53.3028 3.689e-13 ***
## Residuals            5450.6 2854
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA table shows that each predictor has a significant effect on the outcome, as indicated by their small p-values (all less than 0.05). The variable Adult_mortality stands out as the most influential predictor, with highest F-value (6357.96). The other significant measurement was GDP_per_capita. The least influential variable was Hepatitis_B.