An Analysis On Life Expectancy

Andrew Casas

Introduction

Having lived through — and currently living through — a pandemic for over a year I have realized how important our health is and the ways in which societal operations affect the way we live and survive. Studies on life expectancy can demonstrate factors that contribute to higher or lower life spans, which in turn allow for an understanding of what measures are more positive or negative to one's life. Determining the factors of life expectancy is critical for assessing the health of a population as well as being a key indicator to determine the impact of the decisions we make as a society for public health. This report attempts to demonstrate the main factors which determine life expectancy by focusing on 20 different predictors of life expectancy per country. Data used for this report was retrieved from the World Health Organization (WHO). The goal of this report aims to illustrate key factors of life expectancy in order to inform and support relief measures for issues relating to cases such as COVID-19, or other unprecedented national or global emergencies.

Methods

Before I began my exploratory data analysis I first checked if all the variables were correct and determined the relationship between each predictor to the response variable, finding the relationship to be linear. I then summarised my data to get an idea of how much of the data was explained, and examined the p-values to get an idea of what variables can be expected to be the most significant to the data (refer to figure 1). After making sure that my data was correct – by checking all the variables –I split the data in two sections at random which became the training dataset (60) and the testing dataset (40). The training dataset is the larger set and the set which I will use to perform my methods. The testing dataset will be used to validate the training dataset. I summarised both of these datasets again to ensure that both had similar properties. This was done to assure that one dataset did not contain some skewness or influential points that would make these two datasets statistically different. I then performed a confidence interval on the model to determine if the coefficients of the predictors gave us an accurate representation of the data. I then plotted the residual plots and Q-Q plot of the training dataset to see which assumptions were violated. While the normality assumption held, there did not seem to be constant variance. In order to correct this, I decided to power transform my dataset. I then explored the power transformed dataset and again summarised the data to ensure that the R^2 value did not drop significantly, and I plotted the residual plots and Q-Q plot to check whether the assumptions had been corrected. Although not all assumptions were fixed perfectly, the power transformation was able to improve the model. After power transforming my model, I found all of the leverage, outlier and influential points to determine whether there were any specific occurrences that would skew my data in any way. Luckily no significant statistics were found to do so.\ I then assessed the predictors. Firstly, I saw which predictors contain the highest multicollinearity using the vif() function. I then took note of all predictors with a multicollinearity higher than 5, as these predictors share too much information – in other words, higher than 5 – with other predictors. I also took note of all predictors with a high p-value in the summary of the dataset. With this information I performed backwards elimination where I created a series of different models, and with each of these models I removed one of the predictors with high multicollinearity. I then compared each of the AIC, BIC and adjusted R^2 value of these models, including the full model, with the aim to create a model with the lowest AIC, BIC and highest R^2. The AIC and BIC values determine which model is better while being less complicated (having more predictors) and the R^2 value is a proportion of how much of the data is explained. We consider whichever model has the lowest BIC and AIC values – and does not significantly drop the R^2 value (which cannot drop more than 0.03) – as the new reduced model. With this reduced model I ran a partial F-test with the full model and made sure that the p-value of this test was not less than 0.05. I repeated this process many times until I was able to remove as many predictors as I could while still fulfilling each one of my conditions specified. I also periodically checked the new models for the summary of their power transformations to ensure that the model was transformed appropriately and determine whether new assumptions were violated once predictors were removed. Once I constructed the least complicated model possible, I summarised the data and compared it to the full model. This ensured that the adjusted R^2 values were not different by more than a factor of 0.02. I ran a multicollinearity test and made sure that no predictors exceeded 5. If for some reason they had, which they turned out to later no, I took note of why this predictor was still important. Finally, I checked if all assumptions held and if any influential points were worth taking note of.

Results

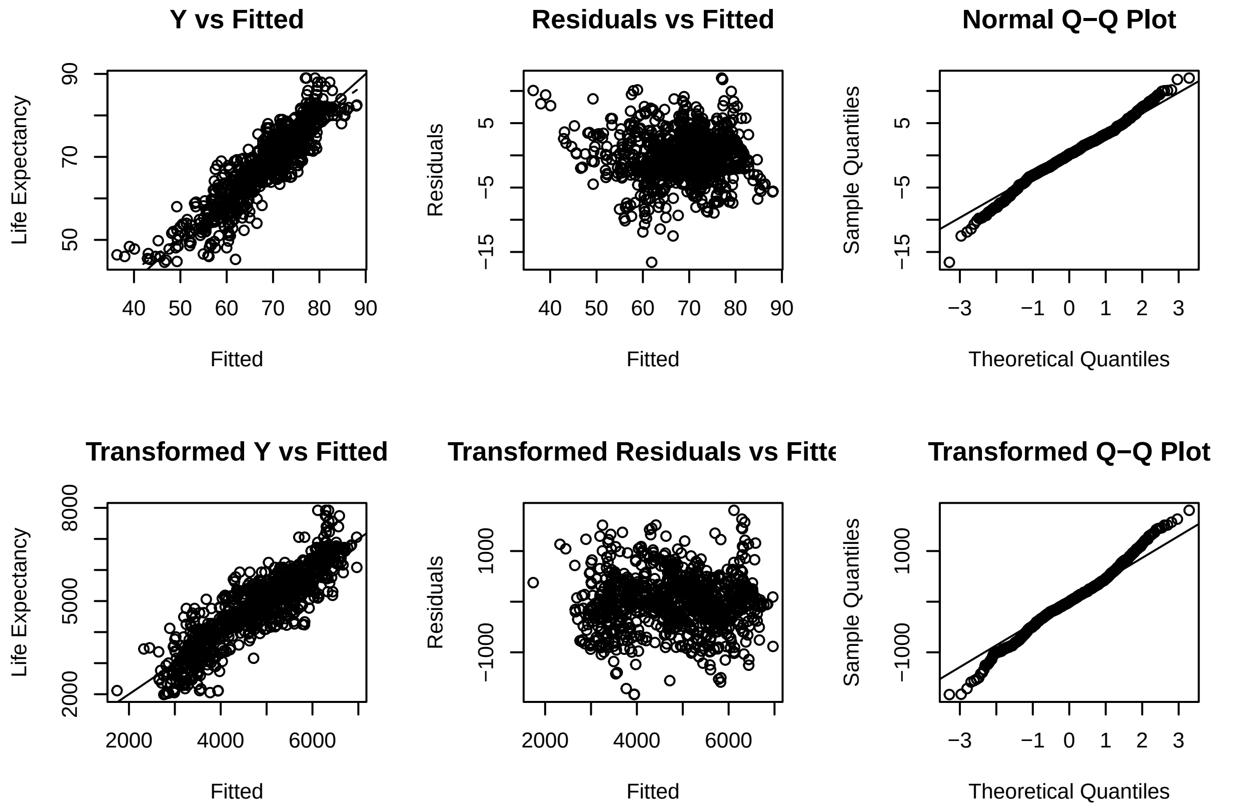
The 20 variables which predict life expectancy are: year (2001 - 2015), the status of a country (whether developed or developing), adult morality, infant deaths, alcohol, recorded per capita (15+) consumption (in litres of pure alcohol), expenditure on health as a percentage of Gross Domestic Product (GDP) per capita(%), Hepatitis B, cases of measles per 1000, BMI, deaths of infants under 5, percent of population immunized to polio, percentage of population immunized to Diphtheria tetanus toxoid and pertussis, deaths per 1 000 live births HIV/AIDS (0-4 years), population, GDP(USD), prevalence of thinness among children and adolescents for age 10 to 19 (%), number of years of Schooling. Each statistic is a country in a given year. I began this analysis with 2939 observations, but after removing all statistics containing “N/A” I was left with a sample size of 1649. After analyzing the dataset, I split the training set and test dataset. The training dataset contained 990 observations. Below we can see how the training and test dataset differed.

Variable	mean (s.d.) in training	mean (s.d.) in test
Year	2007.717 (4.074)	2008.026 (4.104)
Status	0.134 (0.34)	0.168 (0.373)
Life.expectancy	69.125 (8.637)	69.568 (9.031)
Adult.Mortality	168.635 (120.949)	167.584 (131.683)
infant.deaths	30.935 (106.317)	34.985 (139.92)
Alcohol	4.453 (4.014)	4.654 (4.051)
percentage.expenditure	672.594 (1737.491)	738.603 (1791.989)
Hepatitis.B	80.259 (24.582)	77.654 (27.013)
Measles	1981.982 (8929.541)	2588.815 (1.1606182×10^4)
BMI	37.386 (19.893)	39.244 (19.506)
under.five.deaths	42.22 (144.727)	47.225 (186.983)
Polio	83.71 (22.291)	83.346 (22.703)
Total.expenditure	5.877 (2.302)	6.074 (2.292)
Diphtheria	83.821 (22.377)	84.657 (20.329)
HIV.AIDS	1.859 (5.721)	2.171 (6.473)
GDP	5325.793 (1.1290021×10^4)	5926.938 (1.1749024×10^4)
Population	13445099.042 (6.1698486×10^7)	16469166.98 (8.1897368×10^7)
thinness..1.19.years	4.859 (4.604)	4.838 (4.595)
Measles	4.936 (4.674)	4.866 (4.626)
Income.composition.of.resources	0.625 (0.186)	0.642 (0.177)
Schooling	12.041 (2.779)	12.239 (2.818)

Transforming the model Although there seems to be some slight differences, both datasets produce the same data. Because of this, we will now work with the training set and validate our findings later on. In order to transform the model, I changed the status of the value from ‘Developed’ to 0.001 and the value ‘Developing’ into 1. This was because positive numerical values are necessary to transform the model.

To determine whether the findings had resulted in a better model, a power transformation was performed to the dataset by formally checking the assumptions. This was done by using a y vs $\hat{y}$ to determine linearity, applying a residual plot to determine constant variance and using a Q-Q plot to determine normality.

```
## Warning in estimateTransform.default(X, Y, weights, family, ...): Convergence
## failure: return code = 1
```



Above we can see six plots. The top three plots showing the original model, with the lower three representing the transformed model.
Notice that there is no longer a cluster leaning to the right in the residual plot, which means that the assumption of uncorrelated errors as well as constant variance hold because there is no fanning patterns in the residual plot. We can also see that there is a linear relation among both of the y vs *haty* plots which means that both the linearity assumption and the normality assumption also hold.

Reducing the model Now that we have transformed our model, we can examine how the data has changed by comparing figure 1 and figure 2 in the appendix. In order to simplify the data, predictors will be removed to determine which predictors have the highest multicollinearity. The percentage expenditure was found to have a collinearity of 6.51, as expected 'thinness 1-19' and 'thinness 5-9' have a high collinearity because the data is very similar, having the values of 9.58 and 9.78. As well as GDP having a collinearity of 6.76. As described in my methods, I will perform the backwards elimination and compare four models. Model 1 will be missing 'percentage expenditure', model 2 will be missing 'thinness 1-19', model 3 will be missing 'thinness 5-9' and model 4 will be missing 'GDP'.

Model	AIC	BIC	Adjusted R^2
Full model	1.5241736×10^4	1.5349485×10^4	0.7907679
Model 1	1.5259975×10^4	1.5362827×10^4	0.7866666
Model 2	1.5242952×10^4	1.5345804×10^4	0.7903035
Model 3	1.5250431×10^4	1.5353283×10^4	0.7887133

Model	AIC	BIC	Adjusted R^2
Model 4	1.5243989×10^4	1.534684×10^4	0.7900838

From the figure above we can see that the best model is Model 2. Before making this into our new reduced model we will run a partial f-test and determine if this model is acceptable. Running this test gives us a value of 0.076 which means this model is better than the full. This process is repeated recursively until each of the predictors are statistically significant to the model.

validating

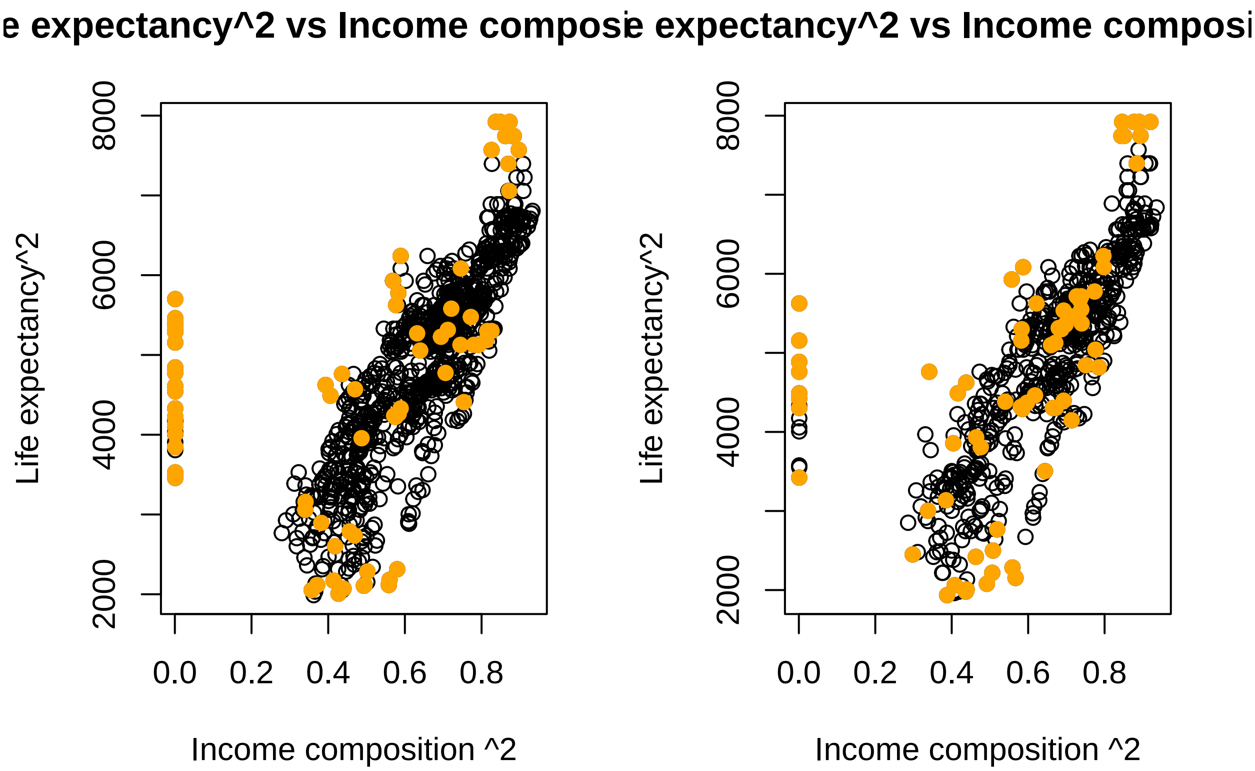
Variable	Estimate	Standard Error
Life.expectancy	2758.5925482 (151.0442129)	2608.7504581 (192.6279083)
Status	-0.1529593 (0.0591136)	-0.1127614 (0.0673019)
Adult.Mortality	-46.7848034 (4.0382762)	-41.9806444 (4.8890661)
percentage.expenditure	59.1192581 (10.4056498)	61.3295925 (11.9989805)
Measles	-10.5629982 (3.0046545)	-7.4259356 (3.5933393)
Diphtheria	2.2926664 (0.805994)	2.6408184 (1.1126206)
HIV.AIDS	363.5945512 (19.6762474)	372.1688653 (24.6219219)
thinness.5.9.years	-93.0667562 (19.561625)	-107.4880055 (25.2876807)
Income.composition.of.resources	1312.2957723 (143.7828699)	1926.1973282 (210.7393854)
Schooling	64.4087265 (10.755389)	34.6148723 (14.5019751)

Here we can also see the final model, with the predictors being Status, Adult Mortality, percentage expenditure, measles, diphtheria, HIV/Aids, thinness 5-9, income composition and schooling, where we can refer to figure 3 in the appendix to see which transformations were applied to the model.

Discussion

By refining the number of predictors in the initial model it became clear that the predictors that significantly impact life expectancy are mostly related to accessibility to resources, such as access to medicine and schooling. This proves to be useful because this may inform the public of what actions need to be taken in order to support countries which are struggling with lower life expectancy. When analysing this in the context of the COVID-19 pandemic, we can see that a crucial aspect of life quality is accessibility to vaccines and other forms of medical immunization.

limitations Comparing the training and test models we determined that one predictor was significantly different in both models. We will compare the two datasets by highlighting the influential points in simple linear regression models to determine if this was the reason.



The plots above show the influential points in the simple regression model life expectancy vs income composition. The left is the training set and the left is the testing, we can see that there are significantly more influential points on the left of the training dataset plot. This is causing the difference when comparing the datasets. However we will still accept this dataset to be validated due to all if its similarities.

We must also notice that the final model had an R^2 value of 0.80 which indicates that we were not able to truly determine all factors which predict life expectancy.

References

World Health Organization. (n.d.). Home. World Health Organization. Retrieved December 18, 2021, from https://www.who.int/ Pardoe, I. (2021). Applied regression modeling. Wiley.

Appendix

Figure 1.

```
## 
## Call:
## lm(formula = (Life.expectancy)^2 ~ (Year) + I((Status)^(-1)) + 
##     sqrt(Adult.Mortality) + log(infant.deaths) + sqrt(Alcohol) + 
##     log(percentage.expenditure) + (Hepatitis.B)^3 + log(Measles) + 
##     BMI + log(under.five.deaths) + (Polio)^4 + Total.expenditure + 
##     (Diphtheria)^4 + I((HIV.AIDS)^-0.5) + log(GDP) + log(Population) + 
##     log(thinness..1.19.years) + log(thinness.5.9.years) + (Income.composition.of.resources)^2 + 
##     Schooling, data = train)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1834.08  -288.55    -4.16   293.24  1804.42 
## 
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                     -6.743e+03  9.297e+03  -0.725 0.468455    
## Year                             4.775e+00  4.646e+00   1.028 0.304308    
## I((Status)^(-1))                -1.841e-01  6.405e-02  -2.875 0.004135 ** 
## sqrt(Adult.Mortality)           -4.541e+01  4.083e+00 -11.121  < 2e-16 ***
## log(infant.deaths)               5.945e+00  1.580e+01   0.376 0.706829    
## sqrt(Alcohol)                    3.979e-01  2.245e+01   0.018 0.985861    
## log(percentage.expenditure)      9.356e+01  2.091e+01   4.474 8.60e-06 ***
## Hepatitis.B                     -1.998e+00  8.968e-01  -2.228 0.026134 *  
## log(Measles)                    -1.175e+01  3.499e+00  -3.358 0.000814 ***
## BMI                              1.502e+00  1.140e+00   1.317 0.187995    
## log(under.five.deaths)          -2.898e+00  1.599e+01  -0.181 0.856208    
## Polio                            1.540e+00  9.883e-01   1.558 0.119518    
## Total.expenditure                7.225e-01  7.884e+00   0.092 0.927005    
## Diphtheria                       2.670e+00  1.094e+00   2.441 0.014839 *  
## I((HIV.AIDS)^-0.5)               3.738e+02  2.176e+01  17.177  < 2e-16 ***
## log(GDP)                        -5.117e+01  2.505e+01  -2.042 0.041378 *  
## log(Population)                  2.159e+00  7.534e+00   0.287 0.774462    
## log(thinness..1.19.years)        8.653e+01  4.873e+01   1.776 0.076089 .  
## log(thinness.5.9.years)         -1.567e+02  4.830e+01  -3.244 0.001218 ** 
## Income.composition.of.resources  1.279e+03  1.494e+02   8.560  < 2e-16 ***
## Schooling                        6.566e+01  1.156e+01   5.681 1.77e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 527.1 on 969 degrees of freedom
## Multiple R-squared:  0.795,  Adjusted R-squared:  0.7908 
## F-statistic: 187.9 on 20 and 969 DF,  p-value: < 2.2e-16
```

Figure 2.

```
##
```

```
## Call:
## lm(formula = (Life.expectancy)^2 ~ I((Status)^(-1)) + sqrt(Adult.Mortality) +
##     log(percentage.expenditure) + log(Measles) + (Diphtheria)^4 +
##     I((HIV.AIDS)^-0.5) + log(thinness.5.9.years) + (Income.composition.of.resources)^2 +
##     Schooling, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1877.09  -283.73     5.23   320.44  1944.26
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     2758.59255  151.04421  18.263  < 2e-16 ***
## I((Status)^(-1))                  -0.15296    0.05911  -2.588 0.009809 **
## sqrt(Adult.Mortality)            -46.78480    4.03828 -11.585  < 2e-16 ***
## log(percentage.expenditure)       59.11926   10.40565   5.681 1.76e-08 ***
## log(Measles)                     -10.56300    3.00465  -3.516 0.000459 ***
## Diphtheria                         2.29267    0.80599   2.845 0.004540 **
## I((HIV.AIDS)^-0.5)               363.59455   19.67625  18.479  < 2e-16 ***
## log(thinness.5.9.years)          -93.06676   19.56162  -4.758 2.25e-06 ***
## Income.composition.of.resources 1312.29577  143.78287   9.127  < 2e-16 ***
## Schooling                         64.40873   10.75539   5.989 2.97e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 528.4 on 980 degrees of freedom
## Multiple R-squared:  0.7917, Adjusted R-squared:  0.7898
## F-statistic: 413.8 on 9 and 980 DF,  p-value: < 2.2e-16
```

Figure 3.

```
##
## Call:
## lm(formula = (Life.expectancy)^2 ~ I((Status)^(-1)) + sqrt(Adult.Mortality) +
##     log(percentage.expenditure) + log(Measles) + (Diphtheria)^4 +
##     I((HIV.AIDS)^-0.5) + log(thinness.5.9.years) + (Income.composition.of.resources)^2 +
##     Schooling, data = final_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1854.25  -289.96    -1.85   306.69  1911.92
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     2713.68439  118.44179  22.912  < 2e-16 ***
## I((Status)^(-1))                  -0.13728    0.04428  -3.100 0.001966 **
## sqrt(Adult.Mortality)            -45.35420    3.09955 -14.633  < 2e-16 ***
## log(percentage.expenditure)       60.02392    7.83655   7.659 3.17e-14 ***
## log(Measles)                      -9.16199    2.30157  -3.981 7.17e-05 ***
## Diphtheria                         2.36809    0.65026   3.642 0.000279 ***
## I((HIV.AIDS)^-0.5)               366.79878   15.33730  23.915  < 2e-16 ***
## log(thinness.5.9.years)          -99.41923   15.41860  -6.448 1.49e-10 ***
## Income.composition.of.resources 1501.86478  118.23350  12.703  < 2e-16 ***
## Schooling                         55.27209    8.56525   6.453 1.44e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 523.9 on 1639 degrees of freedom
## Multiple R-squared:  0.8023, Adjusted R-squared:  0.8012
## F-statistic: 738.9 on 9 and 1639 DF,  p-value: < 2.2e-16
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.