Predicting Life Expectancy for Developing Countries

Rose Ellison, Laura Moses, Daisy Nsibu, Miriam Nyamwaro

5/14/2021

Contents

Introduction

Purpose and Problem

The goal of this project aims to determine which various predicting variables really affect the life expectancy in developing countries. Using this information, a multiple linear regression model can then be established to predict life expectancy in developing countries based on these appropriate factors. Additionally, we can answer key questions, such as: Could a country with a lower life expectancy increase it's health care expenditure in order to improve its average lifespan? How do infant and adult mortality rates affect life expectancy? What impact does schooling have on life expectancy, if any? Is there a positive correlation with immunizations and life expectancy? Do densely populated countries tend to have lower life expectancy? Does Life Expectancy has positive or negative correlation with eating habits, lifestyle, exercise, smoking, drinking alcohol etc. By analyzing these various coefficient estimates, we can hopefully provide insight to these meaningful questions.

Life Expectancy Data

The Global Heath Observatory (GHO) data repository under World Health Organization (WHO) keeps track of health status as well as many other related factors for all countries. The datasets are made available to the public for the purpose of health data analysis. For this particular dataset we will be using, health factors for 193 countries from the year 2000-2015 have been collected from the WHO data repository website, as well as corresponding economic data from the United Nation website.[1] The initial dataset consisted of 2938 rows and 22 columns, although we chose to subset the data and filter `NA` values, resulting in a 1407 x 22 dataframe. The following columns and corresponding descriptions are included in the data:

- `Country`= Country, 193 unique values
- `Year` = Year, 2000-2015
- `Status` = Developed (17%) or Developing (83%) status
- `Life.expectancy` = Life expectancy in age (36.3 - 89)
- `Adult.Mortality` = Adult mortality rates of both sexes (probability of dying between 15 and 60 years per 1000 population)
- `infant.deaths` = Number of Infant deaths per 1000 population (0 - 180)
- `Alcohol` = Alcohol consumption, per capita (litres of pure alcohol)
- `percentage.expenditure` = Expenditure on health as a percentage of GDP per capita (%)
- `Hepatitis.B` = Hepatitis B immunization coverage among 1-year-olds (%)
- `Measles` = Number of reported cases of Measles per 1000 population
- `BMI` = Average Body Mass Index of entire population
- `under.five.deaths` = Number of under-five deaths per 1000 population
- `Polio` = Polio immunization coverage among 1-year-olds (%)
- `Total.expenditure`= General government expenditure on health as a percentage of total government expenditure (%)
- `Diphtheria` = Diptheria tetanus toxoid and pertussis (DTP3) immunization coverage among-1-year-olds (%)
- `HIV.AIDS` = Deaths per 1000 live births due to HIV/AIDS (0-4 years)
- `GDP` = Gross Domestic Product per capita (in USD)
- `Population` = Population of the country
- `thinness..1.19.years` = Prevalence of thinness among children and adolescents for Age 10 to 19 (%)
- `thinness.5.9.years` = Prevalence of thinness among children for Age 5 to 9 (%)
- `Income.composition.of.resources` = Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
- `Schooling` = Number of years of schooling (years)

Results and Discussion

Fit a Multiple Linear Regression Model

To begin, we fit a linear model to the entire filtered developing dataset using `lm()` and `summary()`, to determine which regressors are *not* linearly significant.

From this output, we determine that regressors `Hepatitis.B`, `Measles`, `BMI`, `Polio`, `GDP`, `Population`, `thinness..1.19 years`, and `thinness.5.9.years` all have p-values greater that 0.05, therefore they do *not* have a significant linear association with `Life.expectancy`. Because of this, we remove these regressors from our model and re-fit a new one. The table below shows our summary output for the new model, which now fits 11 regressors, instead of 19.

Table 1: Fitted Multiple Linear Regression Model with 11 Regressors

term	estimate	std.error	statistic	p.value
(Intercept)	443.8225	49.6260	8.9433	0.0000
Year	-0.1952	0.0248	-7.8794	0.0000
Adult.Mortality	-0.0161	0.0010	-16.7308	0.0000
infant.deaths	0.0830	0.0099	8.3471	0.0000
Alcohol	-0.1667	0.0360	-4.6343	0.0000
percentage.expenditure	0.0010	0.0001	8.5032	0.0000
under.five.deaths	-0.0637	0.0074	-8.5901	0.0000
Total.expenditure	0.1154	0.0452	2.5529	0.0108
Diphtheria	0.0126	0.0045	2.7751	0.0056
HIV.AIDS	-0.4647	0.0179	-25.9499	0.0000
Income.composition.of.resources	10.0501	0.8379	11.9937	0.0000
Schooling	1.0820	0.0623	17.3687	0.0000

Checking for Multicollinearity

From the correlation plot below, we can see that there are a few strongly correlated relationships between regressors. When we check for variance inflation factors, all values are less than 5 except for `under.five.deaths` and `infant.deaths`, which have values over 180. Running an `mctest()` on the fit also detects multicollinearity in these two regressors. Plotting them against each other, we can see there is a clear linear relationship between these two regressors, so we will remove `under.five.deaths`, which has the highest VIF value, from the model. Doing this fixes the multicollinearity issue.

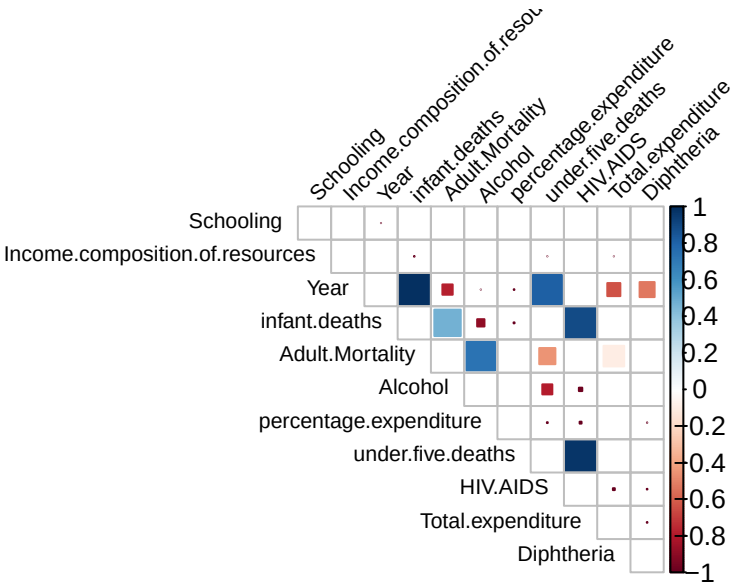
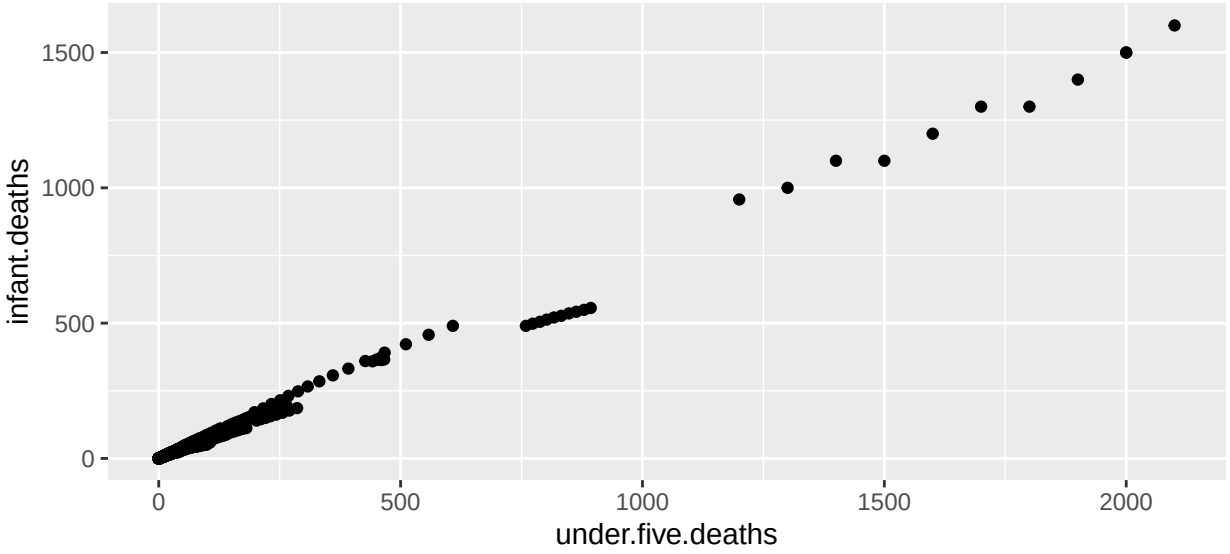


Table 2: Variance Inflation Factors for Regressors

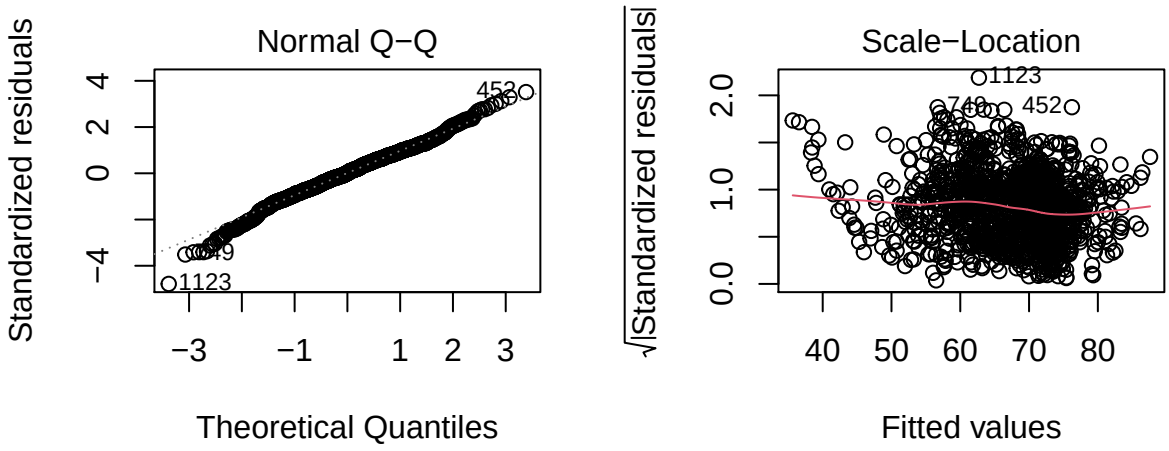
	vif(fit)
Year	1.1106
Adult.Mortality	1.6901
infant.deaths	185.1859
Alcohol	1.5899
percentage.expenditure	1.2644
under.five.deaths	186.7404
Total.expenditure	1.0774
Diphtheria	1.1722
HIV.AIDS	1.4876
Income.composition.of.resources	2.3595
Schooling	2.6711



Checking Assumptions

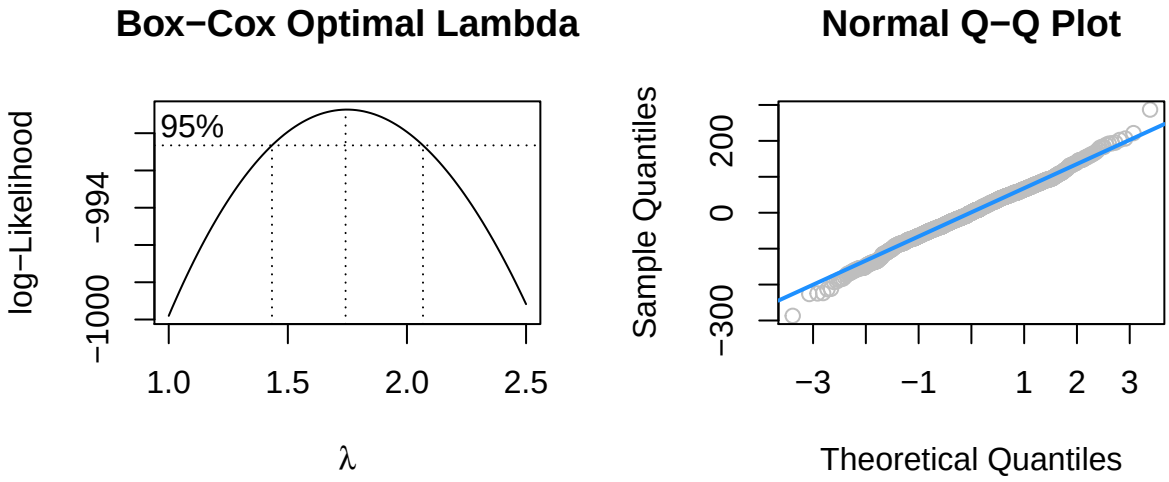
In order for our model to be accurate, the residuals must follow a normal distribution and have a constant variance. While the variance assumption does not appear to be violated, there is a slight problem with the normality. Performing a Shapiro-Wilk normality test results in a p-value of $1.8 \times 10^{-7} < 0.05$, implying that the distribution of the data *are not* normal. Similarly, an Anderson-Darling normality test resulted in a $p - value = 8.3 \times 10^{-6} < 0.05$, meaning we must reject the null hypothesis that our data follow a normal distribution.

There are two methods in which we can address this, so we will investigate these. The first option is to perform a Box-Cox transformation on y, life expectancy in order to correct normality. If that does not work, then we can look more closely at influential observations using Cook's distance.



Box-Cox Transformation on `Life.expectancy`

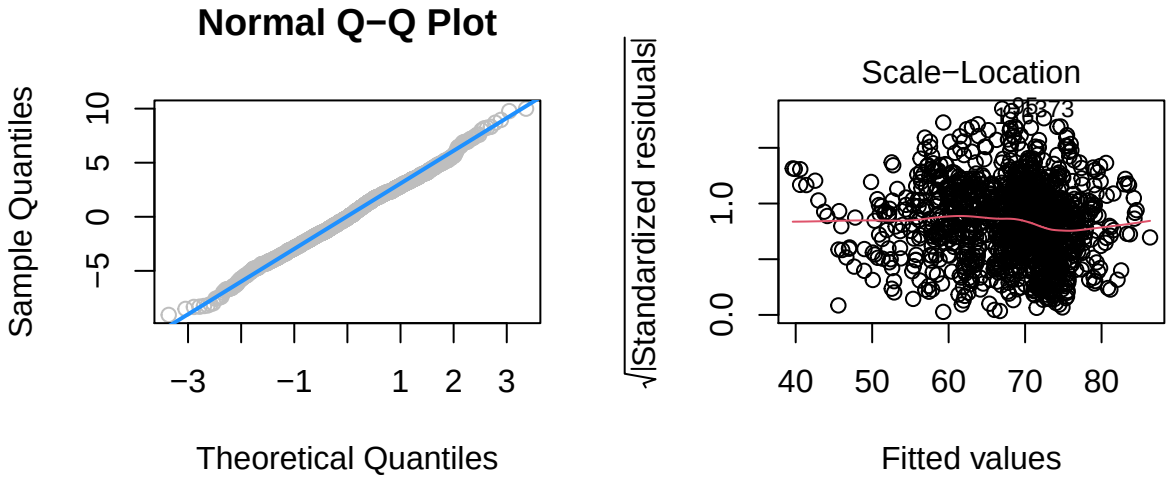
The first plot below shows the optimal λ value as well as confidence interval to perform a Box-Cox transformation on the response y variable, `Life.expectancy`. In the second plot, we can see that there is still an issue with normality at the tail ends, even after performing our transformation.



Performing a Shapiro-Wilks and Anderson-Darling normality test on the transformed model fit produces p-values of 0.00056 and 0.0039, respectively, so although they have gotten better, they still do not pass these normality tests.

Cook's Distance for Influential Observations

We can clearly see some issues in the normality plot where a number of observations are affecting the residual normality and fit of the data. Using Cook's distance, we find 108 such observations which are influence both in the x and y direction. Since transforming `Life.expectancy` did not correct the normality issue, we will try removing the observations that are the most influential.



Visually, the residual normality plot appears to be normal and we can see an improvement over the previous plots. Moreover, the Shapiro-Wilk normality test on the Cook's distance modified model resulted in a p-value of 0.05074, implying we have sufficient evidence to say that our data *does* follow a normal distribution. Thus, we will use this model fit going forward.

```
# Check normality of cooks d fit
shapiro.test(fit_cd$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fit_cd$residuals
## W = 0.9976, p-value = 0.05074
```

Variable Selection

In order to choose the best model, we will perform variable selection using an all-possible-regressions approach. There 5 models determined as the best options:

- **Model 1:** Life.expectancy ~ Year + Adult.Mortality + infant.deaths + Alcohol + percentage.expenditure + Total.expenditure + Diphtheria + HIV.AIDS + Income.composition.of.resources + Schooling *(All regressors)*
- **Model 2:** Life.expectancy ~ Year + Adult.Mortality + Alcohol + percentage.expenditure + Total.expenditure + Diphtheria + HIV.AIDS + Income.composition.of.resources + Schooling
- **Model 3:** Life.expectancy ~ Year + Adult.Mortality + infant.deaths + Alcohol + percentage.expenditure + Diphtheria + HIV.AIDS + Income.composition.of.resources + Schooling
- **Model 4:** (Life.expectancy ~ Year + Adult.Mortality + percentage.expenditure + Total.expenditure + Diphtheria + HIV.AIDS + Income.composition.of.resources + Schooling
- **Model 5:** Life.expectancy ~ Year + Adult.Mortality + Alcohol + percentage.expenditure + Diphtheria + HIV.AIDS + Income.composition.of.resources + Schooling

```
##                                                                      Best Subsets Regression
## ------------------------------------------------------------------------------------------------------------
## Model Index    Predictors
## ------------------------------------------------------------------------------------------------------------
##      1         Adult.Mortality
##      2         Adult.Mortality Income.composition.of.resources
##      3         Adult.Mortality HIV.AIDS Income.composition.of.resources
##      4         Adult.Mortality HIV.AIDS Income.composition.of.resources Schooling
##      5         Adult.Mortality Total.expenditure HIV.AIDS Income.composition.of.resources Schooling
##      6         Year Adult.Mortality percentage.expenditure HIV.AIDS Income.composition.of.resources S
##      7         Year Adult.Mortality percentage.expenditure Total.expenditure HIV.AIDS Income.composit
##      8         Year Adult.Mortality percentage.expenditure Total.expenditure Diphtheria HIV.AIDS Inco
##      9         Year Adult.Mortality Alcohol percentage.expenditure Total.expenditure Diphtheria HIV.A
##     10         Year Adult.Mortality infant.deaths Alcohol percentage.expenditure Total.expenditure Di
## ------------------------------------------------------------------------------------------------------------
##
##                                                         Subsets Regression Summary
## ------------------------------------------------------------------------------------------------------------
##                        Adj.        Pred
## Model    R-Square    R-Square    R-Square      C(p)          AIC         SBIC          SBC          M
## ------------------------------------------------------------------------------------------------------------
##   1        0.5569      0.5565      0.5554    2696.4958    7935.2941    4245.2835    7950.8022    341
##   2        0.7540      0.7536      0.7514     922.8498    7172.7936    3483.5662    7193.4710    189
##   3        0.8095      0.8091      0.8071     424.6544    6842.4556    3153.9925    6868.3024    146
##   4        0.8361      0.8356      0.8342     187.6101    6649.5464    2961.9123    6680.5625    126
##   5        0.8417      0.8411      0.8395     139.1226    6606.3546    2918.8667    6642.5400    122
##   6        0.8474      0.8467      0.8454      89.6742    6560.6260    2873.4470    6601.9808    117
##   7        0.8516      0.8508      0.8493      54.0254    6526.5535    2839.6945    6573.0776    114
##   8        0.8540      0.8531      0.8515      34.3897    6507.3611    2820.7342    6559.0546    112
##   9        0.8561      0.8551      0.8536      17.0822    6490.1529    2803.7956    6547.0158    111
##  10        0.8570      0.8559      0.8543      11.0000    6484.0272    2797.8126    6546.0594    110
## ------------------------------------------------------------------------------------------------------------
## AIC: Akaike Information Criteria
##  SBIC: Sawa's Bayesian Information Criteria
##  SBC: Schwarz Bayesian Criteria
##  MSEP: Estimated error of prediction, assuming multivariate normality
##  FPE: Final Prediction Error
##  HSP: Hocking's Sp
##  APC: Amemiya Prediction Criteria
```

Model Validation

Taking into consideration error comparison, multicollinearity, residuals normality and variance, as well as Radj^2, all models returned very similar results. Model 1, however, had a slightly smaller error and a slightly larger Radj^2 than the others. However, when running a cross-validation analysis, Model 4 was deemed the best model. Again, the cross-validation values, as seen below, were also approximately the same. Thus, we chose to use Model 1 as our best model fit, since it had the lower error and better Radj^2.

Table 3: Cross-validation on 5 Models.

Fit	CV
LS1	1.820
LS2	1.817
LS3	1.831
LS4	1.807
LS5	1.828

Table 4: Best Fit Multiple Linear Regression Model Beta Estimates

term	estimate	std.error	statistic	p.value
(Intercept)	377.8776	41.8185	9.0361	0.0000
Year	-0.1620	0.0209	-7.7601	0.0000
Adult.Mortality	-0.0208	0.0009	-22.7961	0.0000
infant.deaths	-0.0021	0.0007	-2.8429	0.0045
Alcohol	-0.1274	0.0308	-4.1319	0.0000
percentage.expenditure	0.0009	0.0001	7.1660	0.0000
Total.expenditure	0.2411	0.0406	5.9414	0.0000
Diphtheria	0.0181	0.0040	4.5403	0.0000
HIV.AIDS	-0.4512	0.0185	-24.3462	0.0000
Income.composition.of.resources	14.3302	0.9362	15.3075	0.0000
Schooling	0.7549	0.0594	12.7084	0.0000

Best Fitted Model

The table below shows the β coefficient estimates for the best multiple linear regression fitted model.

The best model is:

$$\begin{aligned} \hat{life\ expectancy} = 377.88 - 0.16 \cdot\ & year - 0.02 \cdot\ adult\ mortality - 0.0021 \cdot infant\ deaths \\ & - 0.13 \cdot alcohol + 0.0009 \cdot\ percentage\ expenditure + 0.24 \cdot\ total\ expenditure + 0.018 \cdot\ diptheria \\ & - 0.45 \cdot\ HIV/AIDS + 14.33 \cdot\ income\ composition\ of\ resources + 0.75 \cdot\ schooling \end{aligned}$$

With an $R_{adj} = 0.8559$, we know that 85.6% of the variation in life expectancy can be explained by the regressors.

Conclusion

Summary

The relevant predicting variables for life expectancy we determined are year, adult mortality rates, number of infant deaths, alcohol consumption, expenditure on health as a percentage of GDP per capita, government expenditure on health as a percentage of total government expenditure, DPT3 immunization coverage among 1-year-olds, deaths due to HIV/AIDS, income composition of resources index, and number of years of schooling. Based on our model, an increase in the total expenditure and/or percentage expenditure would improve the average lifespan. Moreover, having more schooling and DPT3 immunizations among children can also improve life expectancy. In contrast, infant and adult mortality rates negatively affect life expectancy, as

well as HIV/AIDS and drinking alcohol. Apart from drinking habits, other lifestyle habits that affect BMI were not found to have a linearly significant impact on life expectancy. Country population was also deemed insignificant, so is unclear whether life expectancy is lower in densely populated countries. However, we are confident that this model can be used to reasonably predict life expectancy's for developing countries using the health and economic information available from the World Health Organization and United Nations databases.

Future Research

In the future, it would be interesting to do a comparative analysis between the developing countries and developed countries. Do the predictor variables for the two classes of country status differ? Additionally, factoring various economic, social, and health by groups may also prove interesting.

References

1. Rajarshi, K. (2018, February 10). Life Expectancy (WHO). Kaggle.