Analysis of trends in Life Expectancy

Aakash Vaithyanathan

Introduction

LIfe expectancy is an important measure to study the overall health of a country. It is one of the indicators used to assess the growth of a country and study the quality of life. As such it is no surprise why there have been numerous studies done on this topic where researchers have tried to study various trends that might be useful to understand life expectancy better.

One factor common in most of the popular studies is how they studied the *one-to-one* relationship between life expectancy to factors like alcohol, education, HIV aids etcetera. Some of the studies include *Obesity and Trends in Life Expectancy*, *Changing relationship between Alcohol and Life Expectancy*, and *Trends in Life Expectancy by Education, Norway (1961-2009)*. The studies above highlight very interesting relationships which we will analyze in our report but one thing common in each of them is how their limitation is studying life expectancy by a single factor when in reality we know it's a culmination of several determining factors. This is where our project will be different and answers the question of *why should we be interested in this project?* as we will take a closer look at variations in life expectancy as a result of *multiple factors and what kind of relationship we observe*. Our research question is

How has life expectancy changed as a result of several health and or economic factors across the globe during the periods of 2000-2015.

Methods

Our primary tool of analysis will be using *Multiple Linear Regression* model over which we will perform several computations for determining the appropriate variables, the goodness of fit for the model, checking for multicollinearity and more.

Variable Selection

In our proposed model, we will perform 3 statistical techniques to determine appropriate variables to study the variation in life expectancy namely *Stepwise AIC method, Stepwise BIC method and Lasso method with cross validation*. An aggregate result from all the 3 methods will be used to determine the best independent variables to choose and fit our model.

Model Validation

Upon selection of the appropriate variables to study the variation, we need to understand *how good is this model to study this change?* We answer this question by calculating the **coefficient of multiple determination** given by

$$R^2 = \frac{SS_{reg}}{TSS} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

A high value of R^2 (i.e close to 1) implies our model predictors play a significant role in determining the variation in our response variable and thus, is a good model fit.

Model Violations and Diagnostics

After we've selected the appropriate model parameters, we will fit a regression model and check for assumptions for **Multiple Linear Regression**. If the model violates any of the assumptions, we will perform an appropriate transformation to correct the model to satisfy these assumptions. To correct normality and linearity we can perform ***Box-Cox transformation*** and for Homoscedasticity we can perform a ***Logarithmic variance stabilizing transformation***.

Additionally, we will check for leverage points in our proposed model to check for any *outliers or influential points* that can negatively impact our interpretation. If we identify any leverage points, we can use various tests like **Cooks Distance, DFITS (*difference in fitted values*) and DFBETA (*difference in beta values*)** to find influential points and make the decision of removing them from our model and refit the new model.

Results

Descriptions of our dataset and different attributes are listed below in Table 1.

TABLE I: Dataset description

Variable	Description
life expectancy	life expectancy in ages
adult mortality	number of deaths per year
hiv-aids	deaths per 1000 live births
schooling	# years of schooling
bmi	BMI of the population
income comp	Income composition of the population

Each of the step-wise AIC, step-wise BIC and Lasso method of variable selection select ***Schooling, HIV-AIDS, Adult Mortality, Income composition of resources, BMI*** as appropriate variables. The model is then fitted using these variables as predictors and response variable being life expectancy.

We can represent our MLR equation as:

$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3 + \beta_4 \cdot x_4 + \beta_5 \cdot x_5 + \epsilon$$

where

- $y \rightarrow Life-expectancy$
- $x_1 \rightarrow HIV-AIDS,$
- $x_2 \rightarrow Schooling,$
- $x_3 \rightarrow Adult-mortality,$
- $x_4 \rightarrow Income-composition$
- $x_5 \rightarrow BMI$
- $\epsilon \rightarrow error$

TABLE II: VIF for different predictors

Predictor	VIF value
schooling	3.03
adult mortality	1.72
hiv-aids	1.43
bmi	1.60
income comp	2.90

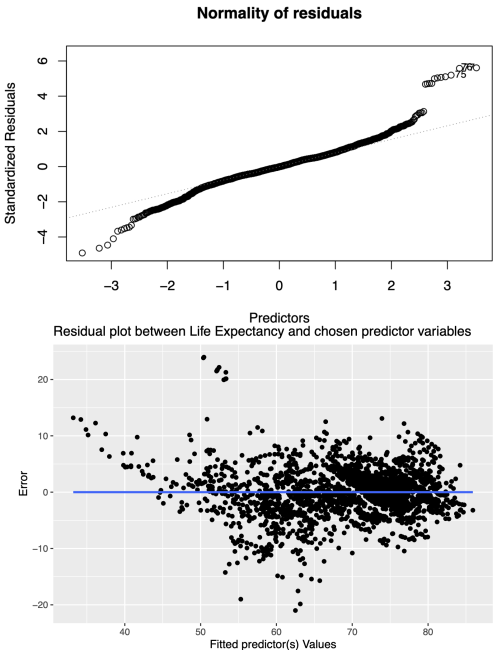

Fig. 1: Normal QQ-Plot and Residual plot to check if model satisfies assumptions of MLR

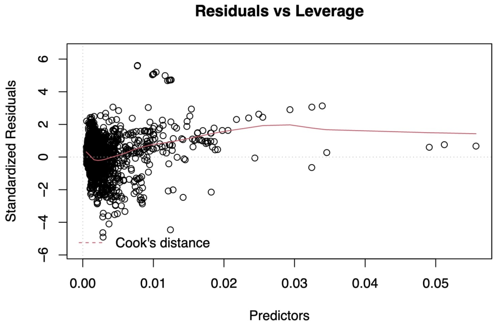

Fig. 2: Residuals vs leverage plot used to determine any influential points

The table II to the left shows the **Variance Inflation Factor (VIF)** values for each of our predictors.

The above 2 figures (figure 1 & 2) show that our model doesn't violate any assumptions of MLR and doesn't have any influential points that negatively impact the interpretation of our model result.

We perform a Hypothesis test using **partial F-Test** on our proposed model and a reduced model (excluding BMI and HIV-AIDS) with the null hypothesis being: $\beta_1 = \beta_5 = 0$ and the alternate hypothesis is either $\beta_1, \beta_5 \neq 0$. From the test, we get a p-value = $2.2e^{-16}$ indicating our result is significant at the 5% significance level and that the full model is a better representation for our problem.

We perform another Hypothesis test on our proposed model to determine if the coefficients of our predictors are significant or not. Our null hypothesis being: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ and alternate hypothesis being either of $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5 \neq 0$. From the test, we get a p-value = $2.2e^{-16}$ indicating our result is significant at the 5% significance level and that each of the predictors are relevant to the problem statement.

The table below presents the 95% confidence interval for each of the predictor variables.

TABLE III: 95% CI for predictors

Predictor	2.5%	97.5%
schooling	0.93	1.13
adult mortality	-0.02	-0.02
hiv-aids	-0.53	-0.46
bmi	0.04	0.06
income comp	8.0	10.75

Using the population mean for each of our predictors, we **predict** the life expectancy of the population = 63.96924 years *(incl. errors)*

Lastly, the coefficient of multiple determination for our model = 0.81 indicating that the predictors play a significant role in determining life expectancy and thus our model is a good fit.

Discusion

The results from our model indicate that there are 5 primary factors that play a crucial role in determining the life expectancy of a community or population. One of the important relationships we must check for when answering our question is

1. *If there is any collinearity among the various predictors?*
2. *Are two or more predictors required to answer this question or is one enough?*
3. *What relationship if any exist among the different predictors?*

These were the questions that were answered by when we checked for **multicollinearity** in our model using the Variance Inflation Factor. A VIF value of ≤ 5 is considered acceptable and conclusive of the fact that there exists no multicollinearity among the model and that all the variables with VIF ≤ 5 are essential to the model. As seen from *Table 2*, this was the case for our predictor variables and thus our model predictors are essential to answering our research question.

Further validation is provided on why our parameters are essential for answering the question from the result of the hypothesis tests and partial f-test that was done. Both the tests arrived at the conclusion that our result is significant indicating that *each of the predictors is essential and that no predictor can be omitted.* Additionally, the 95% confidence interval constructed provided useful information about each of the predictors. For instance, the 95% confidence interval for the schooling parameter is $[.93, 1.13]$ indicating that for a unit rise in life expectancy of the population, we are 95% confident that our true population parameter estimate for schooling will be as low as 0.93 years to as high as 1.13 years. A similar conclusion is made for each of the other population parameters listed in *Table 3*.

We have performed several statistical analyses for our model and gotten deterministic results for each. However, one question still remains and that is *How do these results tie to past studies done by others?*

The study from Walls HL, Backholer K, et.al[1] concluded that life expectancy was observed to be lower in individuals with high

BMI but only up to a certain limit. Beyond this point, they identified that BMI could not be used as the only factor for determination. Another study from Danilova I, et.al[2] and A, Baal Pvan, et.al[3] identified that alcohol was shown to lower life expectancy and that education showed a *positive and strong* correlation to increased life expectancy. These results were no coincidence in our report as we saw a similar trend in behaviour as seen from our several cited results above.

In conclusion, this project helped us better understand the relationship of life expectancy to several factors, what the nature of the relationship is and if they're important. The results of the report were in line with relevant studies of the past.

Limitations

Despite the conclusions drawn from the project, there are certain limitations in our model that are worth mentioning. One of them is the LASSO method of variable selection. If in our model we had a group of predictors that individually do not give relevant information but as a group do, then the LASSO model would only select one arbitrary variable from this group.

Another limitation is on multicollinearity. For multicollinearity, we continue to re-specify the model until we obtain the desired model with VIF value $\leq$ 5. An alternative would be to calculate the correlation values between the predictors to logically determine which variables could be removed and which are not significant. Both options pose problems due to a limited understanding of how the dataset was collected and if a certain predictor played a crucial role in the determination of our question.

Finally, there is also the issue of dealing with influential observations. As mentioned in the methods section, we have *Cook's distance, DFITS, DFBETAs* to determine influential points and make the decision of keeping or deleting an observation. However, if the 3 tests give conflicting results on whether to keep the point or not, we may get an inaccurate model.

References

1. Walls HL, Backholer K, Proietto J, McNeil JJ. Obesity and trends in life expectancy. Journal of Obesity. https://www.hindawi.com/journals/jobe/2012 /1079
89/. Published May 13, 2012. Accessed October 21, 2022

2. Danilova I. Changing relationship between alcohol and life expectancy. Online Library Wiley. https://onlinelibrary.wiley.com/doi/full/10.1111/dar .13034. Accessed October 21, 2022

3. A, Baal Pvan. Trends in Life Expectancy by Education, Norway (1961-2009). SpringerLink. https://link.springer.com/article/10.1007/s10654-012-9663-0. Accessed October 21, 2022.