STA302 Final Report

Alexandra Lomovtseva

December 17, 2021

Introduction

The scope of this research concerns the life expectancy of nations around the world, which is a common measurement of quality of life and health in a country. Specifically, this report will consider how the social infrastructure of a country can contribute to the overall public's health, and what countries can do to increase their life expectancy. As a result, the research question will be: does the socio economic state of a country have an influence on its citizens' life expectancy?

There has been much research done on the topic of life expectancy, throughout the world and following different methods. For instance, a paper called *Social Environment Determinants of Life Expectancy in Developing Countries: A Panel Data Analysis* found that the relationship between life expectancy with education index and gross domestic product in developing countries was at 1% and 5% significance levels, respectively (Hassan et al., 2016). In particular, the paper *Socioeconomic development and life expectancy relationship: Evidence from the EU Accession Candidate countries* found that higher values of GDP per capita and lower values of infant mortality levels lead to higher life expectancy at birth (Miladinov, 2020). In addition, a study done in 1989, titled: *Life expectancy in less developed countries: socio economic development or public health?* provides a precedence for this current research, in terms of their results. It was found that mortality is primarily influenced by socioeconomic development and secondarily by public health measures. The former includes measures such as urbanization, industrialization, and education, and secondarily, while the latter includes measures such as access to safe water, physicians, and adequate nutrition (Rogers & Wofford, 1989). Overall, from the literature, it is clear to see that there are many possible predictors for life expectancy, and that no matter the location, there are many socio economic factors in common that can influence quality and longevity of life.

Since the analysis concerns the relationship between life expectancy and various socio economic predictors, an easily interpretable, yet statistically durable linear regression model will be used to answer the research question.

Methods

In order to choose the predictors to be included in the final linear regression model, we will be using the following: Exploratory Data Analysis (EDA), model validation, linear regression assumptions/conditions check, power transformations, check for outliers or influential points, hypothesis tests, multicollinearity check, Akaike's Information Criterion (AIC), Bayesian

Information Criterion (BIC) in comparison with adjusted R squared value and verification with automated selection.

Before the regression, we will need the predictors and response to satisfy linear regression requirements, firstly with an EDA. For both the (equally split) training and testing sets, the distribution of the numerical variables shown in histograms will need to follow a Normal curve. Further, for each pair of the predictors and response, the corresponding scatter plot will need to show linearity. For model diagnostics, we will formally check the predictor vs residual plots, to check if there is a discernible pattern. In this case, check the two extra conditions for linear regression, to see if there is a linear relationship in the Y and Y-hat plot, as well as in the Normal QQ plot.

If the ends of the Normal QQ plot drop off, then a power transformation needs to be done, and verified until the assumptions for linear regression hold. Next, there needs to be a check for outliers and influential points, and a note of the most significant results.

To pick the predictors for the model, so that it is accurate, yet statistically reliable, we will use hypothesis t-tests and partial-F tests to check if a predictor of lesser significance can be removed without greatly lowering the adjusted R squared.

Finally, the final model will also be validated with the Variance Inflation Factor (VIF), AIC, BIC and adjusted R squared at the end. Thus, we will need the VIF to be lower than 5, for the AIC and BIC to be as small as possible, and for the adjusted R squared to be as large as possible, and for the results to be reproducible with both the train and test sets. To further verify the final model, we will use the stepwise AIC algorithm, to ensure that the same predictors are chosen for the final model.

Results

The data to be used for analysis is called "Life Expectancy (WHO)" and comes from Kaggle (Rajarshi, 2017). The main variables of concern are: Schooling (average years of schooling), Total Expenditure (General government expenditure on health as a percentage), Status (developed or developing country) and Life Expectancy (age). The first three variables are explanatory, while the last one is the response, and these should be able to help us find if there is a linear relation between socio economic pillars of a country and its life expectancy.

Looking at the variable distributions in the train dataset in Figure 1, we see that there are significantly more developing countries than there are developed. Meanwhile the distributions for the two other predictors and the response are fairly Normally distributed, with Total Expenditure experiencing a bit of a right skew and Life Expectancy having a left skew. Overall, it is important to note that transformations will have to be done to fix these skews. It is also noteworthy that Figure 6 in the Appendix, shows a similar pattern, only for the test dataset, reassuring us that the data has been split correctly.

Figure 1: Distributions of predictors and response.

Further, we check for linearity between variables, and Figure 2 shows that there is a general linear relationship between most pairs, except for Total Expenditure and Life Expectancy. In the Appendix, Figure 7, a slight negative relationship is shown between Developing Status and Life Expectancy, which, as mentioned above, shows the skew in the categorical variable. So some additional model diagnostics need to be done in order to check that assumptions are satisfied.

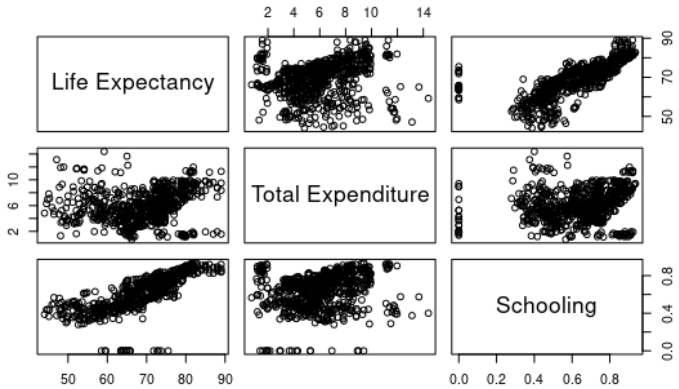

Figure 2: Scatterplots of relationships between all variables.

After checking residual plots of the variables, and finding no discernable patterns, we proceed to check the Y vs Y-hat plot to check linearity once more. In Figure 4, we see that there is a linear relationship, thus verifying the linearity assumption. Next, to verify Normality, refer to Figure 3, which shows the Normal QQ plot, that indicates skewness since the points at the ends of the line drop off. This will have to be fixed using a power transform.

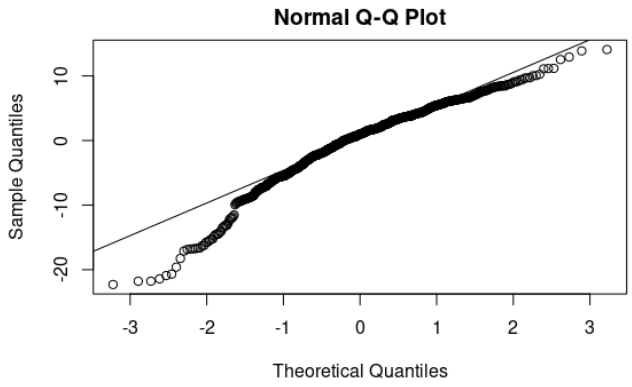

Figure 3: The ends of the Normal QQ plot drop off, which means the data is skewed.

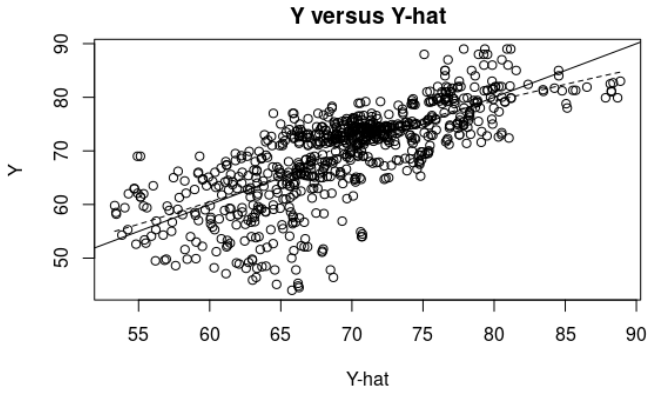

Figure 4: A linear relationship for the Y versus Y-hat plot.

After doing a power transform on the variables, and adjusting them accordingly, the new Normal QQ in Figure 5 is an improvement on skewness, since it is more linear. Afterwards, when checking for any influential points using Cook's distance, and DFFITS or checking for outliers, there were found to be many such points. In the Appendix, the scatter plot in Figure 8 shows the relationship between Life Expectancy and Schooling, with the blue points indicating the outliers, so this is noted as a limitation on the study.

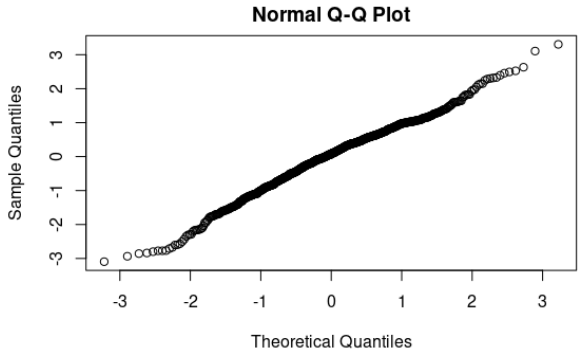

Figure 5: An improvement in the linearity of the Normal QQ plot.

Now that assumptions are satisfied with the transformed model, we move on to the hypothesis tests to pick the predictors for the model. Individual t-tests reveal that only the predictor Total Expenditure is not statistically significant, with a p-value of 0.6223 in a model with adjusted R squared of 0.605. The VIF for all predictors are around 1, significantly less than the benchmark of 5, meaning that there is little multicollinearity. By removing that predictor and doing a partial F-test, we see that the other predictors remain significant and the adjusted R squared slightly changed to 0.6054, meaning it is safe to remove Total Expenditure from the model, as it does not greatly explain the variation in the response variable. Again, the VIF remains low and the F-statistic of the original model is 402.4, increasing to 604, reinforcing the fact that removing the Total Expenditure variable was beneficial.

Using the AIC and BIC to verify results, we see that after removing the Total Expenditure variable, the AIC and BIC slightly reduced, 2 and 5 points respectively to 19932.52 and 19951.19, indicating that the goodness of fit improved. Then, comparing to the test set, we see similar low VIF values, and similar AIC and BIC to the training dataset based model. Finally, to verify the model further, a stepwise AIC algorithm on the test set also concluded that removing Total Expenditure was beneficial, concluding that Life Expectancy is accurately predicted by Developing Status and Schooling.

Discussion and Limitations

Some limitations to note for this study are that: the normality of Total Expenditure and Life Expectancy was skewed, so it required a power transformation to correct. Additionally, the relationship between Total Expenditure and Life Expectancy was not linearly, but in the end Total Expenditure was not included in the model. Moreover, the skew in Status comes from the fact that there are more developing countries than developed countries, so this may have slightly impacted the result and significance of the study. Finally, there were found to be a large number of influential points and outliers, which could not be removed due to ethical considerations, but are important to note since they could have a big impact on the results.

All in all, this study set out to answer the research question: does the socio economic state of a country have an influence on its citizens' life expectancy? Despite the limitations, it concluded that there is indeed a strong relationship between a countrys’ socio economic state (specifically whether its status is developing and its average years of schooling) and its life expectancy. Therefore, governments should recognize that an increase in average schooling years, and an improvement to a status of ‘developed’ should also result in an increase to life expectancy for its citizens. The linear model created in this study for predicting life expectancy is both statistically durable and easy to understand using only two predictors: status and schooling. It also, rather surprisingly, found that general government expenditure on health as a percentage did not play a role in determining a country’s life expectancy, and thus health of the population.

Appendix

Figure 6: Distributions of predictors and response.

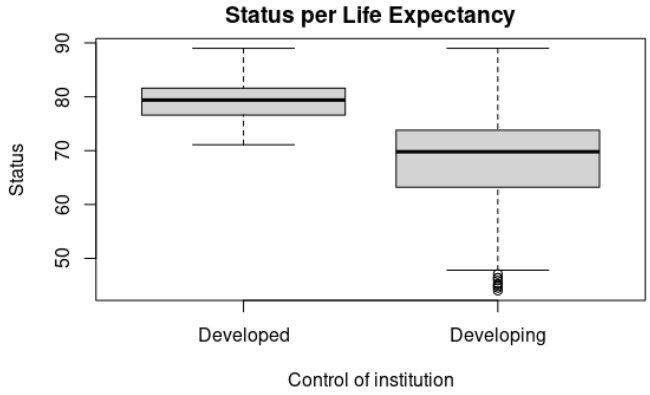

Figure 7: Boxplot showing relationship between Status and Life Expectancy.

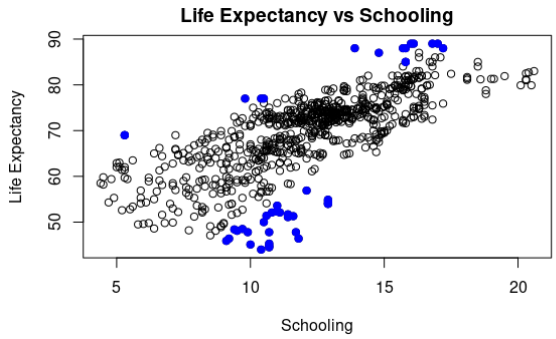

Figure 8: Scatter plot showing outliers in relationship between Life Expectancy and Schooling.

References

Hassan, F. A., Minato, N., Ishida, S., & Mohamed Nor, N. (2016). Social Environment Determinants of life expectancy in developing countries: A panel data analysis. Global Journal of Health Science, 9(5), 105. https://doi.org/10.5539/gjhs.v9n5p105

Miladinov, G. (2020). Socioeconomic development and life expectancy relationship: Evidence from the EU Accession Candidate countries. Genus, 76(1). https://doi.org/10.1186/s41118-019-0071-0

Rajarshi, K. (2017). Life Expectancy (WHO) (Version 1).[Data file]. Retrieved from https://www.kaggle.com/kumarajarshi/life-expectancy-who

Rogers, R. G., &; Wofford, S. (1989). Life expectancy in less developed countries: Socioeconomic development or public health? Journal of Biosocial Science, 21(2), 245–252. https://doi.org/10.1017/s0021932000017934