Tiffany Tran
Professor Cristiana Drake
STA 108
8 December 2023

Term Project

Executive Summary

The goal of this statistical project is to determine the best predictors of life expectancy and use those variables to develop a multiple regression model that accurately predicts it. To achieve this, I built a model by using forwards selection and adjusted R^2 criterion testing, assessing residual assumptions and transforming the predictor. From these methods, I concluded that the percentage of government expenditures towards health care, percentage of population with internet access, births per 1000 people, and percentage of population over 65 years old are significant country-level predictors in determining life expectancy.

Introduction

When determining the well-being of a country, it is often valuable to examine the longevity of their people as it can be a good indicator of their general health, mortality, and standard of living. These factors can be used to address issues faced by countries with lower average life expectancies and in deciding if countries may require additional support from other governments. However, many variables are related to longevity, such as distribution of the population in rural areas or the amount of CO2 emissions produced per capita. In this report, I will identify the best country-level predictors of longevity using statistical methods and develop a parsimonious model that predicts longevity with those variables.

Methods

Before determining which variables are good predictors, it is important to do some initial observations on what the variables actually represent. This starts with some basic descriptives like the mean and standard deviation for each predictor to get a general idea of the data that is being worked with. Notably two of the variables from our data set, land area of countries in square kilometers and GDP, seemed to have extremely large values in comparison to the others, which are mostly percentages from 0 to 100. However this does not say anything about the relationship of these variables with life expectancy and only gives us a basic understanding of each variable being worked with, so I performed simple linear regressions with the outcome variable against each potential predictor.

The process of simple linear regression started with scatterplots, which revealed that all variables besides land area and population seemed to have some relationship, linear or not, with longevity. Then I moved onto residual plot analysis to assess assumptions about the variables to more formally test my observations. This involved looking at residual vs predicted plots to assess linearity and constant error variance. Once again, land area and population seemed to not have any relationship at all with longevity, while the other variables seemed to have either linearity or

constant error variance around nonlinearity (which typically indicates that an x transformation of the predictor is possible to achieve a more linear relationship with life expectancy).

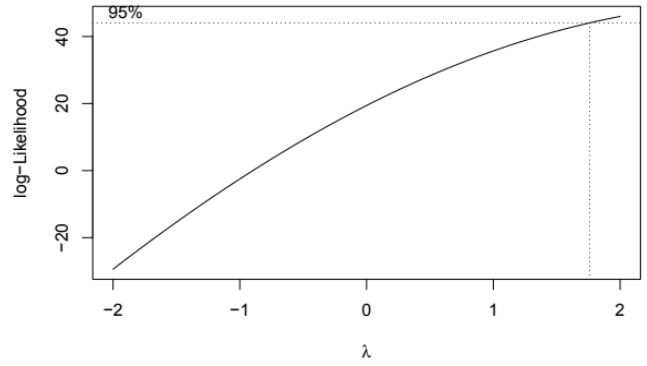
I also obtained the q-q residual plots for each predictor to assess the normality of errors, which is used to see if the normality assumption for possible hypothesis testing is met. It appeared that most of these variables had non-normality, barring internet access percentage, birth rate, elderly population, and cell phone subscriptions, so initial hypothesis testing of these variables would be invalid. But as these tests would not be used for any formal conclusions, I performed t-tests for the slope of the predictors (H_0: $\beta_k = 0$ vs H_a: $\beta_k \neq 0$ where k is the predictor variable). Similar to the observations from before, all the variables besides land area and population seemed to have a significant p-value (less than $\alpha = 0.05$), meaning that these variables have a significant relationship with life expectancy. From all these initial findings, it appears that land area and population are not predictors of longevity. However, these variables should not be entirely excluded from the rest of the model selection process just yet.

After performing these basic statistical observations, model selection can properly begin. This starts with fitting a full model of life expectancy onto all of the predictors and seeing which predictors are significant in this model through partial F-testing. In the case of multiple regression, the partial F-test for each predictor evaluates if the beta coefficient (slope) is equal to zero, but also accounts for all other predictors in the model. As such, we get different conclusions from these tests than from the previous t-tests: healthcare expenditures, internet access, birth rate, and elderly population are significant predictors (p-value less than 0.05), with cell phone subscriptions being almost significant (p-value only a bit larger than 0.05). These five variables seem to be the best predictors of longevity from this simple test, but we need to look more formally at the specifics.

We can further utilize the idea of partial F-testing in a method called forward selection, which aims to develop a model with only relevant predictors. Forward selection starts with the most significant predictor (smallest p-value) and determines whether or not to add the next most significant variable based on its p-value ($\alpha = 0.01$). Using the leaps package in R, I obtained the following suggested order of predictors from forward selection: birth rate, healthcare, internet access, elderly population, cell phone subscriptions, CO^2 emissions, GDP, land area, rural population, and total population. I then checked these results by starting with a model with only birth rate as a predictor, then slowly added other predictors and checked if their partial F-test concluded them as significant. If no remaining variables that could be added were significant, then this process of forward selection would arrive at a “final” model. For this data, this left me with birth rate, healthcare, internet access, elderly population, and cell phone subscriptions as significant predictors of longevity, which is similar to the conclusions made from the full model F-tests. I checked this conclusion with the leaps package’s backward selection, which is a method that removes the least significant variables until only the significant predictors remain ($\alpha = 0.05$), and ended with the same order as forward selection. Thus, I found it reasonable to conclude that this model was a reasonable “final” model from this method.

In addition to using this selection method, I also performed criterion testing using adjusted R^2 to gather further information on the best predictors of longevity. For this analysis, I used another function from the leaps package in R which lists out the best adjusted R^2 values for each possible subset size. In the case of adjusted R^2, the best values are the biggest ones (closest to 1) as those subsets will explain the greatest proportion of the variance in the outcome variable, life expectancy. When performing this test, I also removed the variables of land area, rural population, and total population as they were the "worst" predictors of life expectancy according to our previous forward selection method. From the list of adjusted R^2 values generated by the function, I determined that the subset with predictors birth rate, healthcare, internet access, elderly population, and cell phone subscriptions would produce an appropriately large adjusted R^2 value. This five variable subset was chosen as the adjusted R^2 values barely changed between the five and six variable subset models, indicating that adding a sixth variable would not explain much more information about the regression. This is important because we are searching for a parsimonious model so it is better to stick with a model that balances accuracy and simplicity.

Notably, these five variables from our adjusted R^2 test are the same as the ones used in the model found from forward selection. So using this "final" model, I tested its regression assumptions to see if any of them were violated as this would indicate a possible change in the model. Like the simple linear regressions done earlier, this was done using residual vs predicted and q-q plots. From these residual plots, it appeared that this model had nonlinearity and possibly some nonconstant error, but had normality of errors barring some outliers. And upon observing plots of the residuals against each possible predictor, I found that none of them indicated major issues with normality. Thus I found it reasonable to determine if a transformation of y was necessary using a box-cox plot, which is shown to the left. From this plot, it appears that the ideal transformation for this data would be greater than $\lambda = 2$, which would unfortunately not be ideal for interpretation. So I proceeded with a squared transformation of y, which results in a new model that must be reevaluated.



In the reevaluation process, I started with looking at the residual plots of the new model to see if it fixed the violated assumptions of linearity and constant variance. I observed that the residual vs predicted plot indicated more linearity and better constant variance than the previous model, but normality became worse in the q-q plot. To assess this, I performed partial F-tests on this new model and noticed that cell phone subscriptions were no longer a significant predictor in the model. I double-checked this with the adjusted R^2 criterion test, which now preferred a 4 predictor subset that excluded cell phone subscriptions from the original subset. This felt like adequate evidence to remove cell phone subscriptions as a predictor in this transformed model.

Upon doing residual plots for this updated model, it appeared that the removal of this variable fixed the previously observed change in normality while keeping linearity and constant error variance. Since all assumptions are now met with this new, transformed model, I found it sufficient to conclude it as my true final model for predicting longevity.

Finally, I checked for multicollinearity in my model as it could be the cause for some issues with it. I first used calculations of the variance inflation factor for each of my significant predictors, which did not result in any value being greater than 4. This means that according to the VIF, there is not much collinearity within the model.

Results

The final model is $\text{LifeExpectancy}^2 = 6243.653 + 43.826(\text{Health}) + 15.343(\text{Internet}) - 90.051(\text{BirthRate}) - 46.820(\text{ElderlyPop})$. As the outcome variable life expectancy is transformed through a square, then the interpretation of this model can be described through an increase of squared life expectancy such as follows: There is no proper interpretation of the intercept of 6243.653; Per percentage increase of government expenditures spent on healthcare, the squared average life expectancy increases by 43.826 years; Per percentage increase of the population with internet access, the squared average life expectancy increases by 15.343 years; Per unit increase of birth rate (births per 1000 people), the squared average life expectancy decreases by 90.051 years; Per percentage increase of the population being at least 65 years old, the squared average life expectancy decreases by 46.820 years.

Conclusion

From this project, it appears that the best predictors of longevity are measures of either overall healthiness or wealthiness of a country, such as healthcare expenditures, birth rate, elderly population, and internet access. Government healthcare spending makes immediate sense as a predictor, as better healthcare means that more people have the appropriate resources to get better whenever they are unwell. Birth rate also falls under this more obvious predictor category, as typically poorer countries have greater birth rates to accommodate for other limited resources. Internet access is an interesting predictor as it does not seem directly related to health, and instead serves as an indicator of country wealth. This attribute is greatly predicative of longevity since wealthier countries tend to have healthier people. Unfortunately I do not have a good reason why elderly population negatively affects life expectancy, but I can guess that a greater population of elders means that there are less healthy children which then decreases average longevity.

However, there are some possible issues that I noticed with my project that might have affected the result of model building. First, the boxcox plot suggested a much larger transformation, but I opted for a smaller one for a cleaner interpretation. Next, my t-tests for the simple linear regressions of the predictors were done in the absence of correct testing assumptions of normality, although they were ultimately not very useful tests. Finally, there may be multicollinearity with my selected predictors that I did not observe closely as I used the VIF

method instead of something like a correlation matrix. Besides some possible deeper analysis of these features, I feel like my model is very predicative of longevity based on my process of model selection.