ABSTRACT

The research focuses on investigating the response variable life expectancy in the year 2020 and using other indicators (predictor variables) of the dataset to develop a linear model which explains the life expectancies in 2020. The dataset is a World Development Indicators (WDI), which are derived from a primary World Bank database for development data from officially-recognized international sources (Miladinov, G. 2020). It proposes a model which explains life expectancy as a whole and per continent in the world for year 2020 indicating variables with significant contribution to the model. The researcher is mindful of the objectives of the research paper that entails establishing the relationship between the predictor variables and life expectancy.

1. Introduction

Linear regression is a powerful tool used to predict the value of a dependent variable based on the value of another independent variable. This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values (Agier et al, 2016).

1.1 Preliminary Analysis

The researcher did pre-process stage of exploratory data analysis which is to familiarize with the data. This initial step enables the researcher to clean the data by detailed feature engineering of the dataset. He also identified and treated the missing values using the multiple imputation approach. The researcher also identified missing values (1981) with a blank column (EG.FEC.RNEW.ZS Renewable energy consumption (\% of total final energy consumption) which were treated and deleted. Further normalizing the data that led to the categorisation of the dataset into full model (excluding categorical variables) and reduced model (excluding columns with 80% missing value and categorical variables). The attribute's remade for easy access to columns.

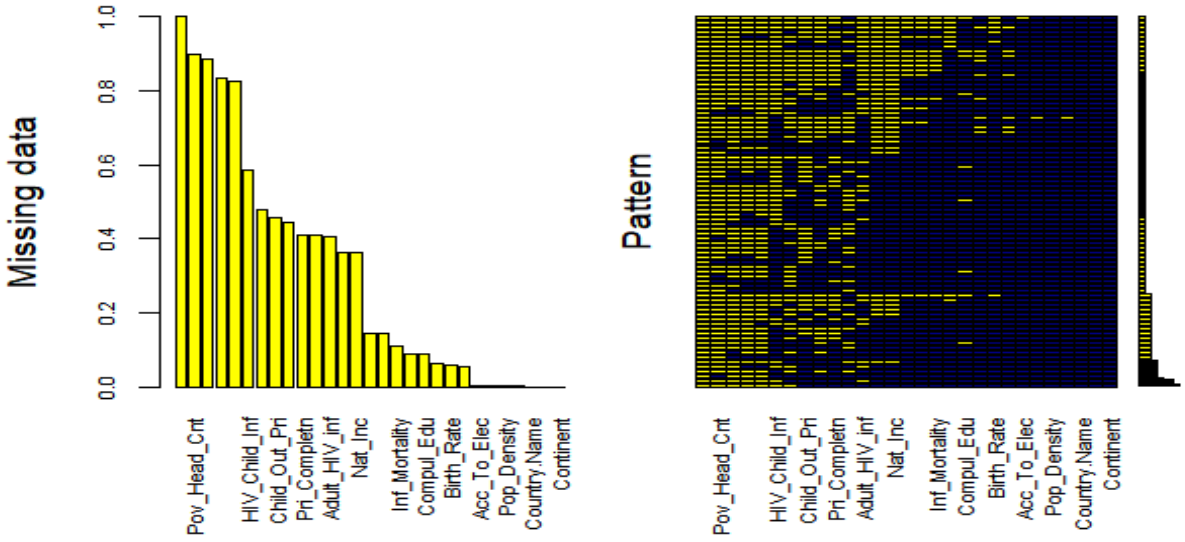

Figure 1: Missing Value Plot

After data cleaning, dataset plotted as a histogram where it gives an overall idea about the missing values.

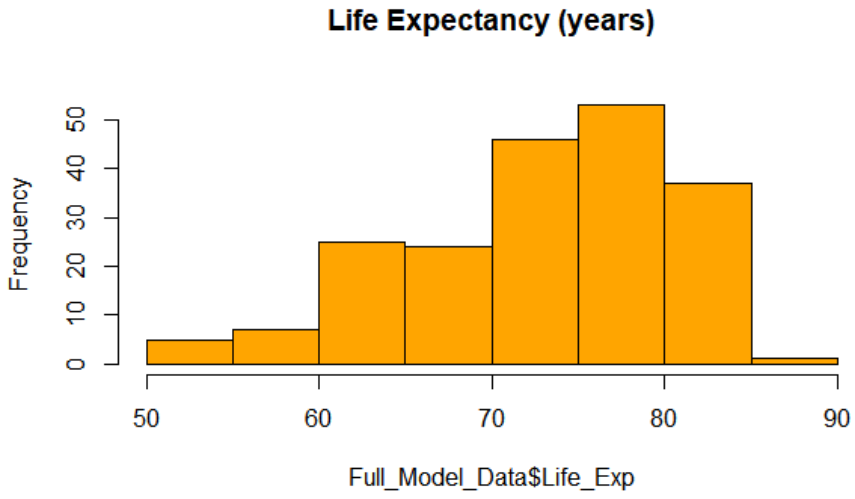

Figure 2: Histogram of Imputed Data Life Expectancy

Initial data visualisation of the relationship between the response variable and predictor variables were done as shown in the appendix. Using multiple imputation method missing values has been imputed and the complete dataset is plotted against dependent variable life expectancy. One of the plot is given below.

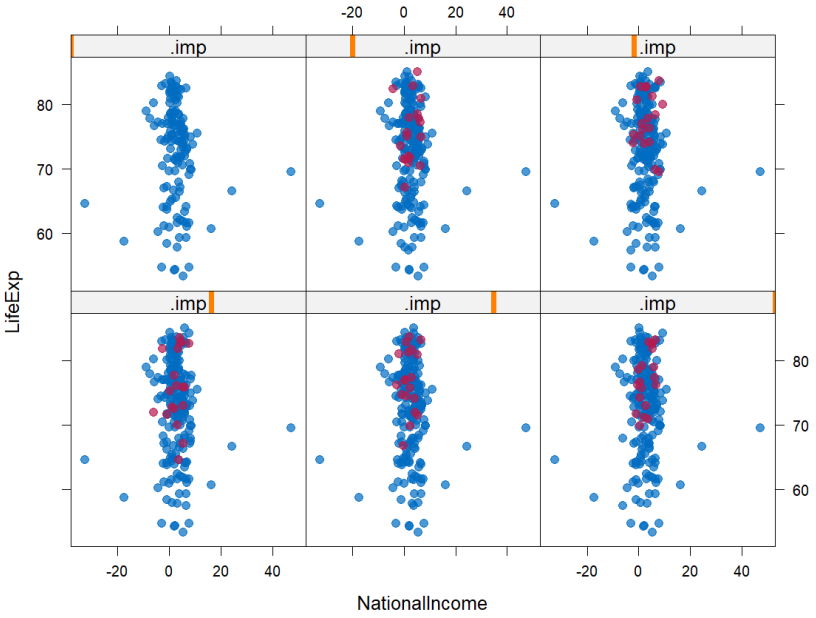

Figure 3: Dataset plot after imputation

1.2 Analysis

In the analysis stage, the researcher conducted a normalcy test using the Shapiro-wilk test. Findings shows that the response variable (life expectancy) significantly deviated from normality (W = 0.95574, p-value = 7.773e-06). This outcome further justified the use of mean in our imputation stage. The researcher performed a linear regression analysis on the full dataset and the model revealed that out of the 24 predictor variables used to predict life expectancy, only seven were found significant, with Adjusted R-squared of 0.8733. The model is given by,

LifeExpectancy~ 6.578e-02*accessElectricity+ 1.326e+00*InterestRate+ 1.595e-01HealthExpenditure-7.055e-02Unemployment-4.381e-01BirthRate-3.207e-05HIVAdult+ 4.497e-02DrinkingWater (Considered significant variables only)

The p values of the imputed model also give us a clarity that the model should be reduced to decrease its p value to make the model fit. The model revealed that out of the 4 predictor variables used to predict life expectancy are actually significant, with improved Adjusted R-squared of 0.8148 at 100% significance when compared to the previous model of 0.812 at 100% significance. With improved Residual standard error: 3.071 on 212 degrees of freedom which are pictorially shown in appendix,

revealed a near perfect model looking at the graph of Residual vs Fitted, Normal Q-Q, Scale-Location and Residual vs leverage.

2. Multi-Collinearity

One of the main assumptions of multiple regression is that there is no perfect multi-collinearity between the predictor variables. This is rarely the case; but it is quite possible for the predictor variables to be highly correlated. This makes it difficult to infer the effect of one predictor variable while holding all constant. While pairs() is good for a general look at the relation between two variables, it doesn't exactly give a more specific number to look at. Also, black and white is not a particular colour scheme to look at. That is why corrplot was used. Below is our outcome:

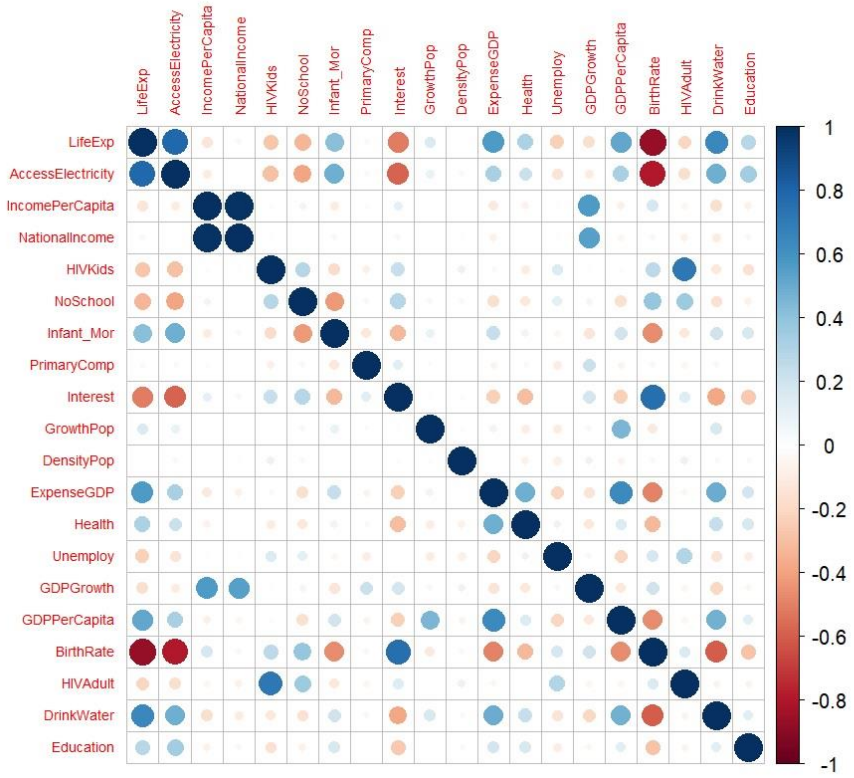

Figure 4: Corplot Matrix

By this point we can see a correlation that is abs(var) =~ 1 and it isn't between our response variable and itself, you started questioning what is happening in your data. Then we applied multi-collinearity check using VIF (Variance influence factor). VIF histogram is the application of VIF on a

cut-off point of 5 which is ideal and generally accepted. Therefore, we can infer that multi-collinearity exits. Attributes with high VIF usually greater than 5 are being considered as must be removed from the list to more fitted one.

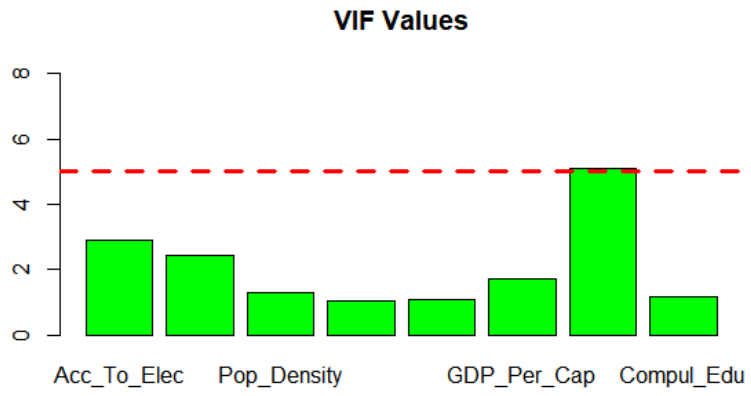

Figure 5: VIF Plot Before Column Drop

2.1 Solving Multi-Collinearity

This ultimately is dependent on what you ultimate objective is. Multi-colinearly is a problem because it makes it difficult to infer the effect of one variable where everything else. So this is not a problem if:

- We just care about the predictions
- the factors with high collinearity don't matter in our analysis

But because we need to infer the effect of the variable (Birth_Rate) that has severe multi-collinearity here are some of the simpler solutions:

- Get more data. Actually, this solves a lot of problems. Like making your prediction more accurate in the first place. But we don't have more data and not an option.
- Transform your predictors, whether by standardization, dummy variables or any other measure.
- Drop the one of the offending variables, though this is generally a last resort unless you want to redesign the model from scratch

For simplicity's sake in this case, we'd actually drop Birth_Rate partially because it's correlated with many of the variables here, but mainly because of the structure of Birth_Rate at the moment. And appendix 6 shows the final output:

3. Finding the Best Model

We can use different approaches to find the best model to predict life expectancy. This project particularly aims to follow forward and backward stepwise regression analysis which involves 3 different stages such as build forward stepwise regression model, build backward stepwise regression model and interpret the results to select the best model.

1. Variable Selection.

In each first and second step the systemic identification of predictive variables is done with backward and forward stepwise regression. As the first step build a complete model and observed that very few variables like GDP growth rate, Adult HIV counts, drinking water quality etc hold quality p value to proceed further with a standard residual error 2.999 with an adjusted R^2 of 0.8335.
In the first forward stepwise regression model is built by beginning with AIC value of 436.68 until no changes noticed with an AIC value of 424.94 with 8 predictive variables. Likewise with backward approach also confined with 8 variables with and AIC 1053.221 same as in forward.

2. Model Selection

When expanding the model with the selected variables, the statistics shows that out of 6, Health expenses, electricity access, birth rate contribute quality values to predict life expectancy and people who managed to drink clean water are too good to predict the life expectancy. With very low p value of < 2.2e-16 predicts the model better with the data and high R^2 is also suggests a range of variation which best fits for the model and scenario.

The same can be achieved by combining full model and reduced model, the detailed code is attached in the appendix. The model selection is carried out in both the ways to analyse the quality of selection process and in both the cases common 6 predictive variables called electricity access, interest rate, health expense, GDP per capita, Birth rate and drinking water made considerable contributions.

While selecting the model, its necessary to identify quality attribute to predict the variable effectively. To make it more accurate, analysing VIF value is a crucial step to finalize. As, the way in which VIF plotted as a histogram, the number of columns dropped may changes from the approach we followed to find the best model. BirthRate is the only attribute which removed through VIF cut-off.

After analysing the new reduced model, the result showed that very less attributes have significant values and eliminated those values with less significant to get the model more fit. The attributes which were kept after keen examination of significant levels were, Access to electricity,

Health expenses, Unemployment, GDP growth, GDP per capita and drinking water with an intercept of 6.723e+01, which is considerably good to proceed with the model. Finally the best model is derived by,

LifeExpectancy= 6.723e+01+ 7.332e-02* AccessElectricity+ Interest *1.347e+00+ Health*2.197e-01+ GDPPerCapita*4.179e-05+ BirthRate*-4.057e-01 DrinkWater*5.532e-02

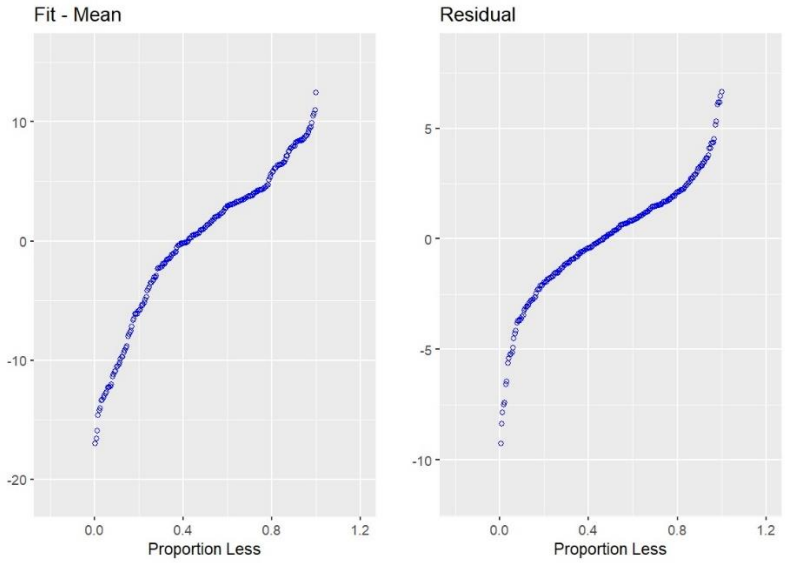

Figure 6: Model Fit and the Residual

4. Life Expectancy across Continent

From the model, it can be inferred from the figure below that life expectancy is least in Africa owing to the fact that all the major indicators for improved life expectancy are very low compared to other continents. It also laid credence to the fact that Europe is the biggest economic powerhouse with the largest Gross Domestic Product with is one of the indicators for life expectancy. In general we can conclude that life expectancy is dependent majorly on Access to electricity, Population Growth, Population density and Gross Domestic Product per Capital. These are the ingredients of industrialized nations.

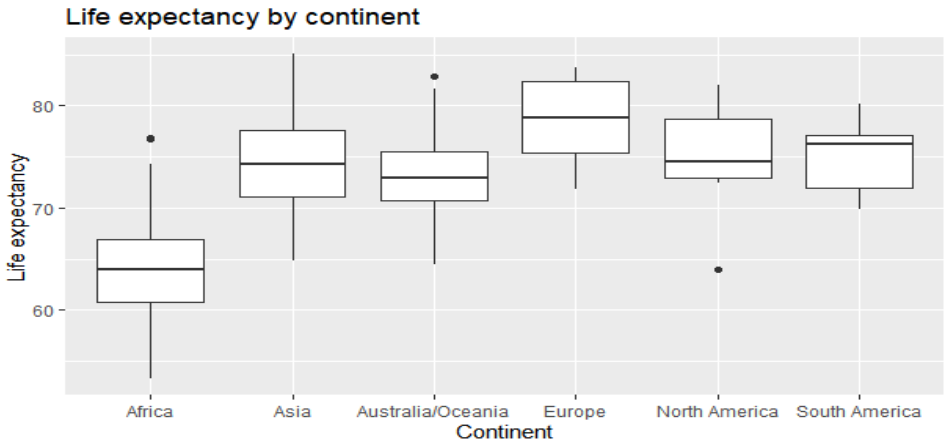

Figure 7: Life Expectancy across Continent