2021

UNIVERSITY OF CALGARY

Multilinear Regression Model to Estimate the Life Expectancy based on 2015 Data

TEAM 5 – WINTER 2021 CLASS

CARLOS NASCIMENTO – ID 30135941

DANIEL GU – ID 30137628

EKPO ARCHIBONG – ID 30135831

SANG YEOP LEE – ID 10078021

Contents

1. Introduction 3
2. Data Collection and Methodology 4
   2.1 Modelling Workflow 8
3. Key Questions 10
4. Main Results of the Analysis (Exploratory Data Analysis) 11
   4.1 Full Model Analysis (Part A) 12
   4.2 Stepwise Regression Analysis (Part B) 13
   4.3 Reduced Model Analysis, With Under Five Death (Part C) 15
   4.4 Reduced Model Analysis, With Infant Death, (Part D) 16
   4.5 Reduced Model Analysis, 3 Variables, (Part E) 17
   4.6 Interaction Model Analysis (Part F) 18
   4.7 Revised Interaction Model Analysis (Part G) 20
   4.8 Higher Order Model Analysis Part 1 (Part H) 22
   4.9 Higher Order Model Analysis Part 2 (Part I) 23
   4.10 Box Cox Transformation (Part J) 24
   4.11 Revised Interaction Model With Box-Cox Transformation (Part P) 25
   4.12 Revised Interaction Model With Box-Cox Transformation (Part R) 29
5. Model Prediction 34
6. Conclusions 35
7. Discussions 36
8. Recommendations 37
9. Appendix (Files) 38
10. References 38

List of Figures

Figure 1: World life expectancy for 2015. Source https://ourworldindata.org/life-expectancy 3
Figure 2: Life Expectancy Boxplot variance by year 5
Figure 3: Life Expectancy by Country year by year 6
Figure 4: Extreme missing data identified in Alcohol and Total Expenditure 8
Figure 5: Model Development Flow Chart 11
Figure 6: Variable Plot 22
Figure 7: Leverage Plot for model P 25
Figure 8: Box-Cox plot for model P 26
Figure 9: Residual and Scale-Location plots confirm that linearity assumption is met. 27
Figure 10: Histogram of residuals and the Normal Q-Q Plot of modelP 28
Figure 11: Summary of final modelP 29
Figure 12: Leverage Plot for model R 30
Figure 13: Box-Cox plot for model R 31
Figure 14: Residual and Scale-Location plots confirm that linearity assumption is met. 32
Figure 15: Histogram of residuals and the Normal Q-Q Plot of modelR 32
Figure 16: Summary of final modelR 33
Figure 17: Model P prediction of world Life Expectancy for 2015 using mean data 34
Figure 18: Model R prediction of world Life Expectancy for 2015 using mean data 34

List of Tables

Table 1: Life Expectancy dataset used in the project 4
Table 2: Summary of comparison between all the models considered in analysis 35
Table 3: Summary of comparison of statistical parameters for modelP and modelR 36

1. Introduction

With the recent technological innovations, we have seen a strong growth in the standard of living for the global population. Technologies have allowed us to produce more food in effort to fight off starvation as well develop more effective medicines and medical techniques which have led us to conquer numerous deadly diseases. One fundamental area of life which have benefited from these technological revolutions of the last century is life expectancy. Life expectancy can be defined as the age at which a person is expected to live for the specific year of birth. Different groups of people may show different life expectancies based on number of different factors but overall, the average life expectancies have shown a strong growth. The average life expectancy of the world has seen an exponentially growth over the past century. The average life expectancy of the world in 1900 was 32 years. 100 years later, the life expectancy in 2000 has more than doubled to 66.3 years. (https://ourworldindata.org/life-expectancy) In effort to explore deeper into the life expectancy, the report will investigate on number of factors that may have a significant impact in determining the life expectancies.

The goal of knowing the average life of expectancy of population helps countries to allocate economic resources to improve the quality life of people, like heathy system, as well as to define plans to guarantee the budget in a level that can pay for the retirement of old people. In addition, the financial institutions and insurance companies use this parameter to estimate the expected cost to pass to their clients for products (e.g. insurance, loans, retirement plans) that they normally offer to them. Besides that, organizations, like World Health Organization (WHO), compare this number of different countries to verify the improvement of life expectancy of their population year by year and identify priority of countries that need more resources. For example, it is presented in **Figure 1:** World life expectancy for 2015. Source https://ourworldindata.org/life-expectancythe actual life expectancy map for the year 2015.

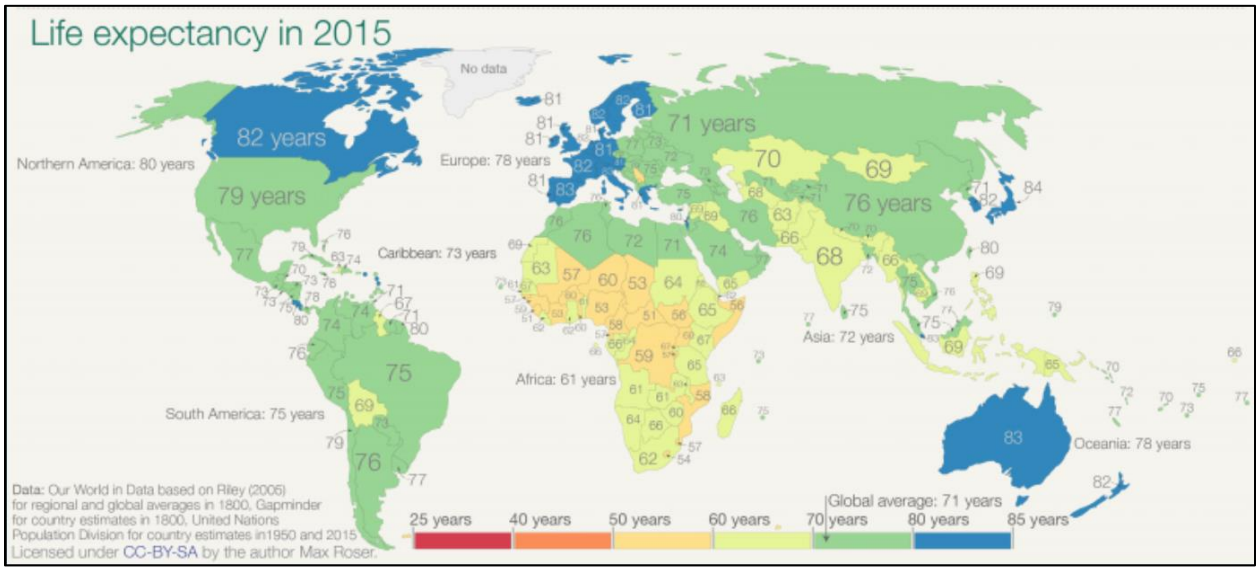


Figure 1: World life expectancy for 2015. Source https://ourworldindata.org/life-expectancy

Number of factors that will be utilized for the statistical analysis can be divided into different areas. First is the economic status of the country. This will include factors such as the economic status of each countries where they are defined as either developed or developing. It also includes economic measures such as Gross Domestic Product (GDP) and income. The other area is on the healthcare system of each countries. Factors that are part of the healthcare system are healthcare expenditure, immunization for different diseases and measure of reported diseases. The next area we are exploring are on the personal habits and lifestyles that may impact the overall health of an individual. This includes the percentage of population who are drinking as well as average Body Mass Index (BMI) to reflect the health condition of each countries. On top of all the areas discussed, we are also exploring how education, population and mortality can impact the life expectancy.

The goals of this project are identify based on available data collected by WHO the main variables that statistically influences the life expectancy for a specific year (2015) and build some linear regression model to help to predict the life expectancy.

2. Data Collection and Methodology

The data collection was carried out by the Global Health Observatory (GHO) under the World Health Organization (WHO). The dataset is related to life-expectancy and other health factors for 193 countries. The corresponding economic data was collected from the United Nations (UN) website and combined with the WHO data into a single dataset that is open to public view and can be found in the website below:

https://www.kaggle.com/kumarajarshi/life-expectancy-who

Country	Year	Status	Lifeexpectancy	Adult Mort	infant deaths	Alcohol	percentage	Hepatitis B	Measles	BMI	under-five	Polio	Total expend	Diphtheria	HIV/AIDS	GDP	Population	thinness
Afghanistan	2015	Developing	65	263	62	0.01	71.27962	65	1154	19.1	83	6	8.16	65	0.1	584.2592	33736494	17.2
Afghanistan	2014	Developing	59.9	271	64	0.01	73.52358	62	492	18.6	86	58	8.18	62	0.1	612.6965	327582	17.5
Afghanistan	2013	Developing	59.9	268	66	0.01	73.21924	64	430	18.1	89	62	8.13	64	0.1	631.745	31731688	17.7
Afghanistan	2012	Developing	59.5	272	69	0.01	78.18422	67	2787	17.6	93	67	8.52	67	0.1	669.959	3696958	17.9
Afghanistan	2011	Developing	59.2	275	71	0.01	7.097109	68	3013	17.2	97	68	7.87	68	0.1	63.53723	2978599	18.2
Afghanistan	2010	Developing	58.8	279	74	0.01	79.67937	66	1989	16.7	102	66	9.2	66	0.1	553.3289	2883167	18.4
Afghanistan	2009	Developing	58.6	281	77	0.01	56.76222	63	2861	16.2	106	63	9.42	63	0.1	445.8933	284331	18.6
Afghanistan	2008	Developing	58.1	287	80	0.03	25.87393	64	1599	15.7	110	64	8.33	64	0.1	373.3611	2729431	18.8
Afghanistan	2007	Developing	57.5	295	82	0.02	10.91016	63	1141	15.2	113	63	6.73	63	0.1	369.8358	26616792	19
Afghanistan	2006	Developing	57.3	295	84	0.03	17.17152	64	1990	14.7	116	58	7.43	58	0.1	272.5638	2589345	19.2
Afghanistan	2005	Developing	57.3	291	85	0.02	1.388648	66	1296	14.2	118	58	8.7	58	0.1	25.29413	257798	19.3
Afghanistan	2004	Developing	57	293	87	0.02	15.29607	67	466	13.8	120	5	8.79	5	0.1	219.1414	24118979	19.5
Afghanistan	2003	Developing	56.7	295	87	0.01	11.08905	65	798	13.4	122	41	8.82	41	0.1	198.7285	2364851	19.7
Afghanistan	2002	Developing	56.2	3	88	0.01	16.88735	64	2486	13	122	36	7.76	36	0.1	187.846	21979923	19.9
Afghanistan	2001	Developing	55.3	316	88	0.01	10.57473	63	8762	12.6	122	35	7.8	33	0.1	117.497	2966463	2.1
Afghanistan	2000	Developing	54.8	321	88	0.01	10.42496	62	6532	12.2	122	24	8.2	24	0.1	114.56	293756	2.3
Albania	2015	Developing	77.8	74	0	4.6	364.9752	99	0	58	0	99	6	99	0.1	3954.228	28873	1.2
Albania	2014	Developing	77.5	8	0	4.51	428.7491	98	0	57.2	1	98	5.88	98	0.1	4575.764	288914	1.2
Albania	2013	Developing	77.2	84	0	4.76	430.877	99	0	56.5	1	99	5.66	99	0.1	4414.723	289592	1.3
Albania	2012	Developing	76.9	86	0	5.14	412.4434	99	9	55.8	1	99	5.59	99	0.1	4247.614	2941	1.3
Albania	2011	Developing	76.6	88	0	5.37	437.0621	99	28	55.1	1	99	5.71	99	0.1	4437.179	295195	1.4
Albania	2010	Developing	76.2	91	1	5.28	41.82276	99	10	54.3	1	99	5.34	99	0.1	494.3588	291321	1.4
Albania	2009	Developing	76.1	91	1	5.79	348.056	98	0	53.5	1	98	5.79	98	0.1	4114.137	2927519	1.5
Albania	2008	Developing	75.3	1	1	5.61	36.62207	99	0	52.6	1	99	5.87	99	0.1	437.5396	2947314	1.6

Table 1: Life Expectancy dataset used in the project

We have both qualitative independent variable and quantitative independent variables in the dataset.

- The dependent variable is “Lifeexpectancy” in years
- The qualitative independent variable is “Status” (Developing and Developed)

- The quantitative independent variables to be included in analysis are
  - "Adult Mortality" – Probability of dying between 15 and 60 years per 1000 population
  - "infant deaths" – Number of Infant Deaths per 1000 population
  - "Hepatitis B" – Immunization coverage amongst 1 year olds (%)
  - "Measles" – Number of reported cases per 1000 population
  - "BMI" – Average BMI of entire population
  - "under-five deaths" – Number of under-five deaths per 1000 population
  - "Polio" – Immunization coverage amongst 1 year olds (%)
  - "Diphtheria – DTP3 Immunization coverage amongst 1 year olds (%)
  - "HIV/AIDS" – Deaths per 1000 live births 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 ages 10-19 years (%)
  - "thinness 5-9 years" – prevalence of thinness among children ages 5-9 years (%)
  - "income composition of resources" – Human Development Index (0 – 1)
  - "schooling" – Number of years of schooling (years)
- The remaining variables not listed above have incomplete information and therefore were not used in the analysis. They are "Alcohol", "percentage expenditure" and "total expenditure".

Based on the dataset it is possible to verify in the boxplot below (**Figure 2:** Life Expectancy Boxplot variance by year) that there is an increment of life expectancy year by year.

```
boxplot(Life.expectancy ~ Year, data = life)
```

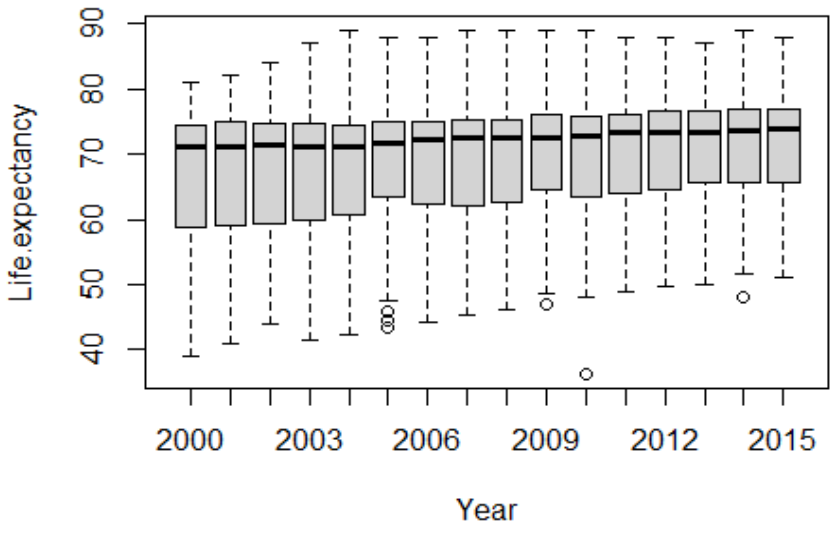


Figure 2: Life Expectancy Boxplot variance by year

In terms of Countries, we can see significant variance between countries every year in **Figure 3:** Life Expectancy by Country year by year.

```
ggplot(data = life) +
  geom_point(mapping = aes(x = Country, y = Life.expectancy)) + facet_wrap(~ Year)
## Warning: Removed 10 rows containing missing values (geom_point).
```

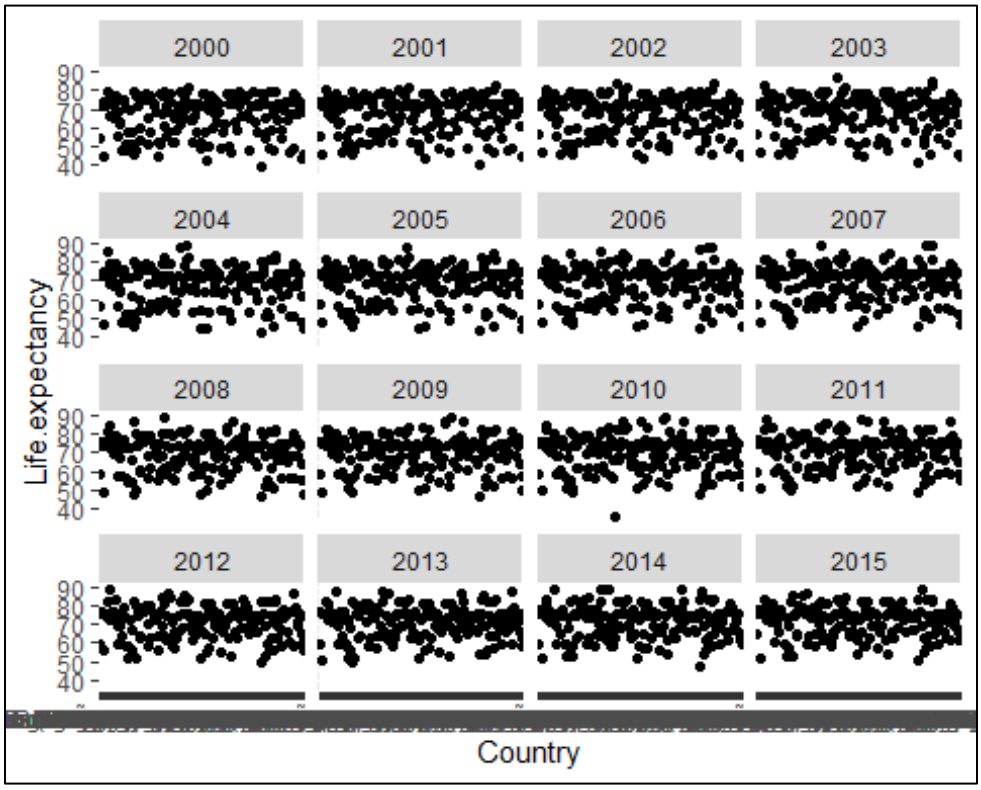


Figure 3: Life Expectancy by Country year by year

The data investigation includes analyzing all these independent variables to find out the ones that are particularly relevant in predicting life expectancy. Filtering of the dataset to extract information for the year 2015 was carried out prior to the analysis. Subsequently we developed linear models and tested them for the various assumptions which includes the following:

Linearity Assumptions:

$$H_0 : \textit{There is straight line relationship between the predictors and the response}$$
$$H_a : \textit{There is no straight line relationship between the predictors and the response}$$

Various residual vs fitted plots were generated throughout the project execution phase.

Equal Variance Assumptions:

$$H_0: \textit{Heteroscedasticity is not present (Homoscedasticity)}$$
$$H_a: \textit{Heteroscedasticity is present}$$

Residual and scale-location plots were analyzed and evaluation confirmed with Breusch-Pagan tests.

Normality Assumptions:

$$H_0: \textit{The sample data are significantly normally distributed}$$
$$H_a: \textit{The sample data are not significantly normally distributed}$$

Histograms and Q-Q plots were used to carry out analysis for normality of residuals and confirmed with Shapiro-Wilk normality test.

Individual t-test:

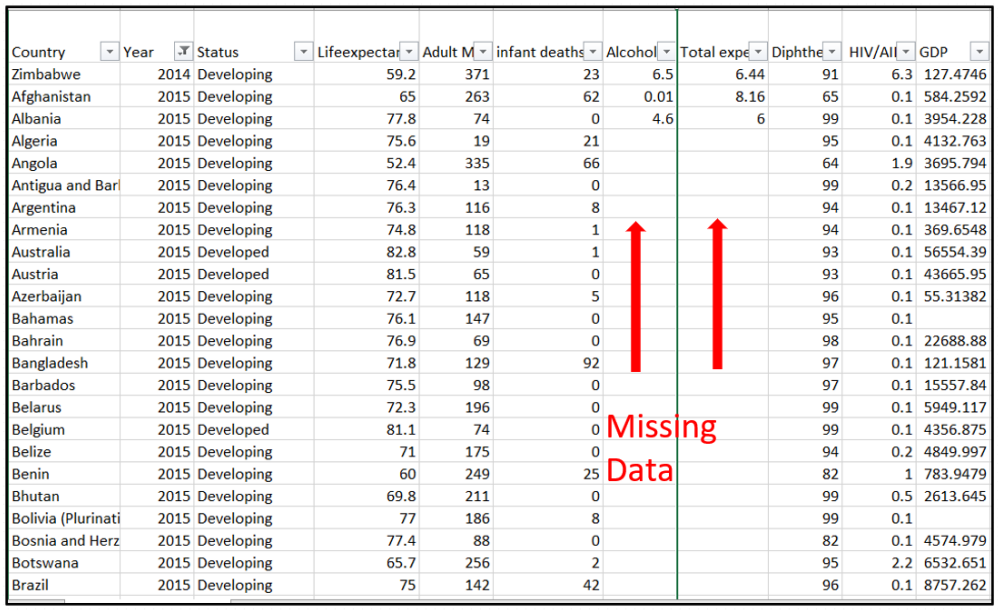
$$H_0: \beta_i = 0, \textit{ in the model for the independent parameters}$$
$$H_a: \beta_i \neq 0, \textit{ in the model for the independent parameters}$$

This was based on analysis of t-values and p-values to determine relevance.

The linear models were thereafter tested for multicollinearity using Variance Inflation Factors (VIF) and outlier effects by checking Cook's distance and leverage and ways were found to fix the issues to produce better statistical results for predicting life expectancy.

2.1 Modelling Workflow

The data in our dataset was acquired between the years 2000 and 2015. Some of the variables had several severe missing data that we needed to drop out from the analysis. The variables removed are: Alcohol(only has 6 records), Percentage Expenditure(only has two records) and Total Expenditure(only has two records).



Country	Year	Status	Lifeexpectar	Adult M	infant deaths	Alcohol	Total expe	Diphthe	HIV/AI	GDP
Zimbabwe	2014	Developing	59.2	371	23	6.5	6.44	91	6.3	127.4746
Afghanistan	2015	Developing	65	263	62	0.01	8.16	65	0.1	584.2592
Albania	2015	Developing	77.8	74	0	4.6	6	99	0.1	3954.228
Algeria	2015	Developing	75.6	19	21			95	0.1	4132.763
Angola	2015	Developing	52.4	335	66			64	1.9	3695.794
Antigua and Bar	2015	Developing	76.4	13	0			99	0.2	13566.95
Argentina	2015	Developing	76.3	116	8			94	0.1	13467.12
Armenia	2015	Developing	74.8	118	1			94	0.1	369.6548
Australia	2015	Developed	82.8	59	1			93	0.1	56554.39
Austria	2015	Developed	81.5	65	0			93	0.1	43665.95
Azerbaijan	2015	Developing	72.7	118	5			96	0.1	55.31382
Bahamas	2015	Developing	76.1	147	0			95	0.1	
Bahrain	2015	Developing	76.9	69	0			98	0.1	22688.88
Bangladesh	2015	Developing	71.8	129	92			97	0.1	121.1581
Barbados	2015	Developing	75.5	98	0			97	0.1	15557.84
Belarus	2015	Developing	72.3	196	0			99	0.1	5949.117
Belgium	2015	Developed	81.1	74	0			99	0.1	4356.875
Belize	2015	Developing	71	175	0			94	0.2	4849.997
Benin	2015	Developing	60	249	25			82	1	783.9479
Bhutan	2015	Developing	69.8	211	0			99	0.5	2613.645
Bolivia (Plurinati	2015	Developing	77	186	8			99	0.1	
Bosnia and Herz	2015	Developing	77.4	88	0			82	0.1	4574.979
Botswana	2015	Developing	65.7	256	2			95	2.2	6532.651
Brazil	2015	Developing	75	142	42			96	0.1	8757.262

Figure 4: Extreme missing data identified in Alcohol and Total Expenditure

In addition, for building the multilinear regression model, we just focused on creating a model for 2015 year because one of the assumptions for applying linear regression is the independence assumption and consequently time-series data should not be applied. Another variable also removed from the model was country as we just have one data observation per country per year and our goal is to build an overall worldwide multilinear regression model for just 2015.

The methodology adopted for this project followed the Multilinear Regression Modelling covered in DATA 603 winter 2021 class based on R generated code, which the steps are as following:

1) Data Preparation:

- We filtered out the data for 2015 (in order to meet the Independence Assumption of Linear regression) and removed Country since we have only one data observed per country per year.

- Removed variables with considerable missing data. They are Alcohol, Percentage Expenditure and Total Expenditure.

2) Statistical Analysis of the Full Model:

- Initially a Partial F test was carried out to verify if at least one of the independent variables has significant influence (p-value< $\alpha = 0.05$) in Life Expectancy;
- t-test was thereafter carried out and we checked the p-value of the coefficients to identify the significant variables at p-value < $\alpha = 0.05$
- After that, a model selection with Stepwise Regression Procedures were carried out using both, forward and backward methods.
- A regression subset was carried out to also verify the significant variables as well, but based on minimum values of Marlow's Cp and AIC.
- Finally using the full model, model diagnostics to verify if the required linear regressions assumptions of Linearity, Equal Variance of the Residuals (Homoscedasticity) and Normality were checked to see if some variables are redundant (Multicollinearity) and if there are outliers with significant influence in the response variable (Life Expectancy).

3) Statistical Analysis with reduced model (First Order Model):

- We considered only the variables identified in mainly stepwise regressions as being significant from the full model.
- Like in previous analysis, a t-test was done to identify the significant variables at p-value< $\alpha = 0.05$
- A Partial F Test was carried out to verify if the reduced model is better in predicting the Life Expectancy than the full model.
- Finally, all the model selection and diagnostics as in the full model were carried out on the reduced model. If the model selection indicated consistently to reduce some variables, all the steps above were repeated with this new first order model.

4) Statistical Analysis of First Order Model with Interaction Terms:

- The same steps above were followed as in previous analysis (model selection and model diagnostics), but also considering all the interaction terms between the significant variables.

5) Statistical Analysis of Interaction Terms Model with Quadratic Terms:

- The same steps above were followed as in previous analysis (model selection and model diagnostics), but also considering the higher order quadratic terms between the significant variables. Adding the quadratic terms did not produce any significant relationship with Life Expectancy.

6) Box-Cox Transformation of Best Model obtained from previous Models:

- This was implemented to address heteroscedasticity
- We determine the best Box-Cox lambda.
- We thereafter applied the Box-Cox transformation with the best lambda in the best model until this moment and checked the significance of coefficients at $\alpha = 0.05$ level with 95% confidence interval.
- We also carried out model diagnostics to verify if the transformation eliminated some problems identified in the previous analysis.

For all the analysis described above, the adjusted coefficient of determination, RMSE and linear regression assumptions were verified to estimate the best models to be used as starting point in each statistical analysis step.

3. Key Questions

We investigated the answers to the following key questions related to the data being analyzed.

- What are the best predictors of Life Expectancy from all the independent variables investigated?
- What are the differences if any between the evaluation from individual t-tests and stepwise regression analysis in determining the best predictors of Life Expectancy?
- Are there multicollinearity or outlier effects in the data being analyzed?
- Are there interaction or higher order effects that needs to be addressed?

The answers to some of these questions will be provided through exploratory data analysis in Chapter 4.

4. Main Results of the Analysis (Exploratory Data Analysis)

Using data wangling, data visualization and statistical analysis of our data, the flow chart that has led us to develop the final model is shown below.

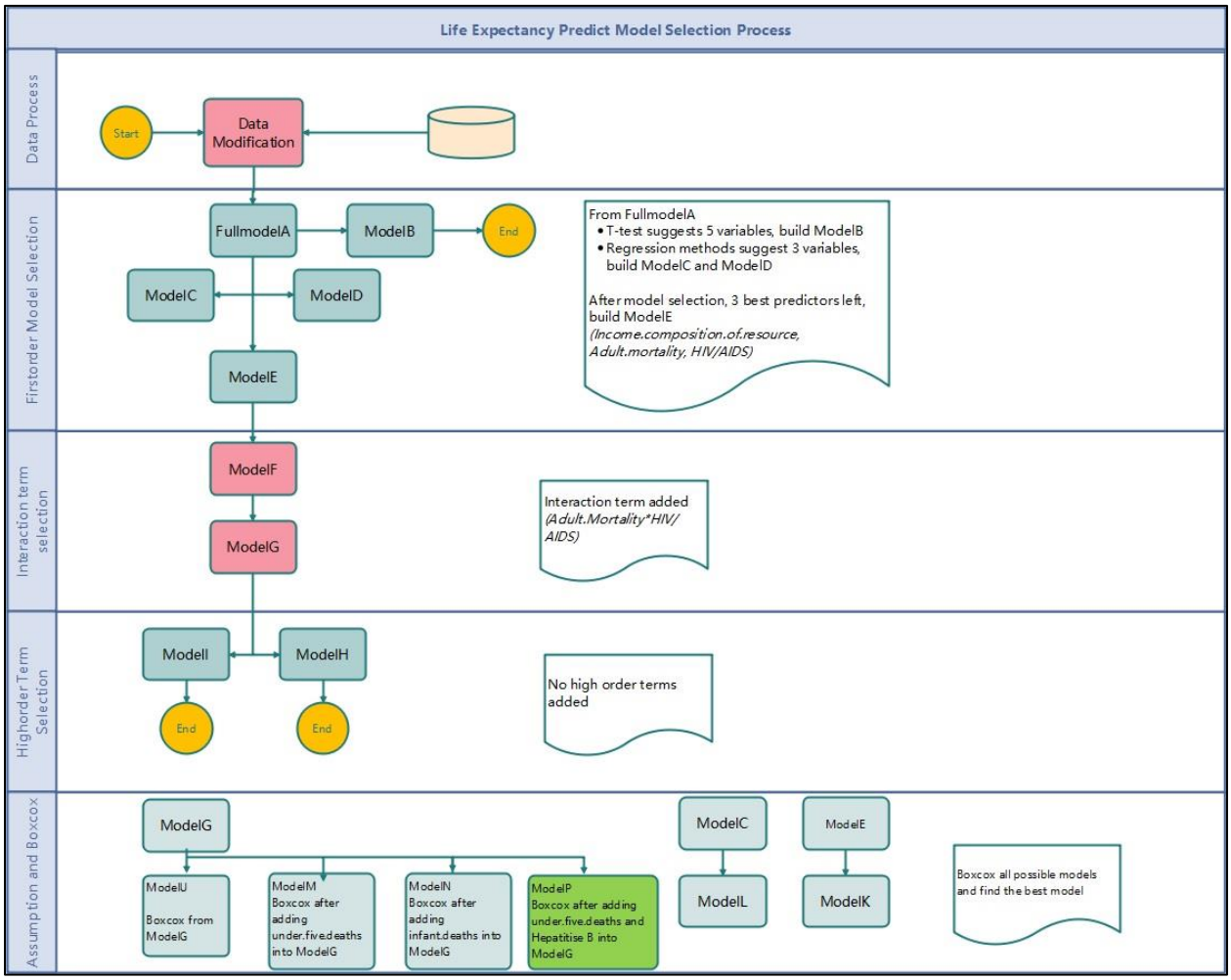


Figure 5: Model Development Flow Chart

The first step of building the foundation of the model was to develop the appropriate first order model. Initial full model which has all the variables except the ones removed for the purpose of missing significant amount of data is the following.

$$\hat{Y}_{\text{Life Expectancy}} = \hat{\beta}_0 + \hat{\beta}_1 X_{\text{Status}} + \hat{\beta}_2 X_{\text{Adult Mortality}} + \hat{\beta}_3 X_{\text{Infant Death}} + \hat{\beta}_4 X_{\text{Hepatitis B}} + \hat{\beta}_5 X_{\text{Measles}} + \hat{\beta}_6 X_{\text{BMI}} + \hat{\beta}_7 X_{\text{Under Five Death}} + \hat{\beta}_8 X_{\text{Polio}} + \hat{\beta}_9 X_{\text{Diphtheria}} + \hat{\beta}_{10} X_{\text{HIV}}\text{AID} + \hat{\beta}_{11} X_{\text{GDP}} + \hat{\beta}_{12} X_{\text{Population}} + \hat{\beta}_{13} X_{\text{Thiness 1-19 years}} + \hat{\beta}_{14} X_{\text{Thhinness 5-9 years}} + \hat{\beta}_{15} X_{\text{Income Composition of Resource}} + \hat{\beta}_{16} X_{\text{Schooling}} + \varepsilon$$

In total, we have 16 independent variables as part of the data set. In order to develop the foundation of significant first order model, we have decided to first test out the overall F-test to determine if there is at least one significant independent variable to the Life Expectancy.

4.1 Full Model Analysis (Part A)

A1. Hypothesis on Full Model Test

$$H_O\ (Null\ Hypothesis): \beta_1 = \beta_2 = \beta_3 = .. = \beta_i = 0$$
$$H_a\ (Alternative\ Hypothesis): \text{At least one } \beta_i \neq 0$$
$$\text{Where i is number from 1 to 16}$$

Full Model Test

$$P_{Full\ Model} = \ 2.2 \times 10^{-16} < 0.05$$

The P value is below 0.05 and this rejects the null hypothesis. The full model test suggests that there is at least one variable out of 16 that is significant in affecting the life expectancy.

As a next step, we have used the individual T test to screen out the variables that holds no significance to the life expectancy. 0.05 was used as the alpha value or limit on the P value of each individual T test. Below is the summary of the partial test or individual coefficients test (t-test).

A2. Hypothesis on Individual Coefficient Test (T Test)

$$H_O\ (Null\ Hypothesis): \beta_i = 0$$
$$H_a\ (Alternative\ Hypothesis): \text{At least one } \hat{\beta}_i \neq 0$$
$$\text{Where i is number from 1 to 16}$$

Full Model Individual Coefficient Test

Below variables had the p value smaller than the $\alpha = 0.05$ limit which indicates that they have significant impact on the Life Expectancy at $\alpha = 0.05$ level.

$$P_{Income\ Composition\ of\ Resources} = \ 9.59 \times 10^{-10}$$
$$P_{Adult\ Mortality} = \ 6.15 \times 10^{-8}$$
$$P_{HIV\ AIDS} = \ 0.0325$$
$$P_{Under\ 5\ Death} = \ 0.0445$$
$$P_{Infant\ Death} = \ 0.0469$$

To confirm the individual coefficient test, we also checked for the 95% confidence interval.

A3. 95% Confidence Interval

$$Income\ Composition\ of\ Resources: (2.338 \times 10^{1}\ to\ 4.312 \times 10^{1})$$
$$Adult\ Mortality: (-2.806 \times 10^{-2}\ to\ -1.377 \times 10^{-2})$$
$$HIV\ AIDS: (-9.283 \times 10^{-1}\ to\ -4.122 \times 10^{-2})$$
$$Under\ 5\ Death: (-9.446 \times 10^{-1}\ to\ -1.193 \times 10^{-3})$$
$$Infant\ Death: (9.329 \times 10^{-4}\ to\ 1.311 \times 10^{-1})$$

Confirming the individual coefficient test, the Income Composition of Resources, Adult Mortality, HIV AIDS, Under 5 Deaths and Infant Death are the only variables that contains 0 in the 95% confidence interval.

Using the significant variables, we developed a reduced model consisting of 5 variables. In effort to check the significance of each 5 variables, we then move on the part B of the analysis which is the stepwise regression of the model. Before the stepwise regression, we would need to determine that the reduced model used as part of the stepwise regression is more significant than the full model. ANOVA test comparing the full model to reduced model is performed.

4.2 Stepwise Regression Analysis (Part B)

To test that a particular subset of q of the coefficients are zero, the hypotheses are:

B1. ANOVA Test

$$H_O: (Null\ Hypothesis): \beta_{p-q+1} = \beta_{p-q+2} = \cdots = \beta_p = 0$$
$$H_a: (Alternative\ Hypothesis): \text{At least one } \beta_i \neq 0$$

The analysis considers the coefficients of the models as:

$$Full\ Model: \beta_1, \beta_2, \beta_3, \beta_4, \beta_5$$
$$Reduced\ Model: \beta_1, \beta_2, \beta_3$$

And the assumption is that:

$$H_O: (Null\ Hypothesis): \beta_4 = \beta_5 = 0$$
$$H_a: (Alternative\ Hypothesis): \text{At least one } \beta_4 or\ \beta_5 \neq 0$$
$$anova(reduced, full)$$

$$P_{Reduced\ model\ over\ Full\ model} = 0.1521 > 0.05$$

With the P value above 0.05, we fail to reject null hypothesis which concludes that the reduced model is a better predictor of Life Expectancy than the full model. Utilizing the reduced model, we have performed the step wise regression.

B2. Stepwise Regression (Both) Model

$$Variable\ 1: Income\ Composition\ of\ Resources$$
$$Variable\ 2: Adult\ Mortality$$
$$Variable\ 3: HIV\ AIDS$$

B3. Stepwise Regression (Forward) Model

$$Variable\ 1: Income\ Composition\ of\ Resources$$
$$Variable\ 2: Adult\ Mortality$$
$$Variable\ 3: HIV\ AIDS$$

B4. Stepwise Regression (Backward) Model

$$Variable\ 1: Income\ Composition\ of\ Resources$$
$$Variable\ 2: Adult\ Mortality$$
$$Variable\ 3: HIV\ AIDS$$

All three stepwise regressions on the reduced model had resulted in same variables of Income Composition of Resources, Adult Mortality and HIV AIDS. Additional criteria were also analyzed in confirming the accuracy of the stepwise model developed.

B5. Mallow's CP/ Akaike's Information Criterion (AIC)/Adjusted R Squared Criterion

	CP	AIC	$AdjustedR2$
[1,]	85.01800	700.9612	0.8049981
[2,]	18.11205	651.5686	0.8676401
[3,]	6.865340	640.8483	0.8790252
[4,]	7.644146	641.6105	0.8792130
[5,]	6.000000	639.8451	0.8817151

CP values of the full model has been analyzed to support the stepwise regression model. Based on the results above, CP value of 3 variables is showing the lowest values. AIC values are also analyzed, and 3 variables are showing the lowest AIC value. Adjusted R squared value is highest for the 5 variables but 3 variables is still showing a high Adjusted R squared value. As a result of analyzing the criterion of stepwise model, we can proceed with three variable model.

In reducing two additional variables of Infant Death and Under Five Death from the already reduced model, we took the effort to reduce a single variable at a time. We first reduce the Infant Death variable and utilized the ANOVA test to compare the reduced model to the full model. The reduced model has four variables and they are (1) Income Composition of Resources (2) Adult Mortality, (3) HIV AIDS and (4) Under Fiver Death.

4.3 Reduced Model Analysis, With Under Five Death (Part C)

To test that a particular subset of q of the coefficients are zero, the hypotheses take the form:

C1. ANOVA Test

$$H_O : (Null\ Hypothesis): \beta_{p-q+1} = \beta_{p-q+2} = \cdots = \beta_p = 0$$
$$H_a : (Alternative\ Hypothesis): \text{At least one } \beta_i \neq 0$$
$$P_{Reduced\ model\ over\ Full\ model} = 0.08401 > 0.05$$

With the P value above 0.05, we can conclude the reduced model is a better predictor of Life Expectancy than the full model. With the conclusion, we proceed with the rest of the test we have done in step A2 to A3.

C2. Hypothesis on Individual Coefficient Test (T Test)

$$H_O\ (Null\ Hypothesis): \beta_1 = \beta_2 = \beta_3 = .. = \beta_i = 0$$
$$H_a\ (Alternative\ Hypothesis): \text{At least one } \beta_i \neq 0$$
$$\text{Where i is number from 1 to 4}$$

Full Model Individual Coefficient Test

Below variables had the p value smaller than the $\alpha = 0.05$ limit which indicates that they have significant impact on the Life Expectancy at $\alpha = 0.05$ level.

$$P_{Income\ Composition\ of\ Resources} = 2 \times 10^{-16}$$
$$P_{Adult\ Mortality} = 7.06 \times 10^{-7}$$
$$P_{HIV\ AIDS} = 0.000413$$

From the individual coefficient test, we can see that the P value of the Under Five Death exceeds the 0.05 limit and should be removed.

C3. 95% Confidence Interval

$$Income\ Composition\ of\ Resources: (32.215441432\ to\ 40.370134940)$$
$$Adult\ Mortality: (-0.025409931 \text{ to } -0.011452783)$$
$$HIV\ AIDS: (-1.188319764\ to\ -0.349738453)$$
$$Under\ Five\ Death: (-0.006253899\ to\ 0.001817052\ )$$

Confirming the individual coefficient test, the Under Five Death is the only variables that contains 0 in the 95% confidence interval.

For section D, we repeat section C but with Infant Death.

4.4 Reduced Model Analysis, With Infant Death, (Part D)

To test that a particular subset of q of the coefficients are zero, the hypotheses take the form:

D1. ANOVA Test

$$H_O : (Null\ Hypothesis): \beta_{p-q+1} = \beta_{p-q+2} = \cdots = \beta_p = 0$$
$$H_a : (Alternative\ Hypothesis): \text{At least one } \beta_i \neq 0$$
$$P_{Reduced\ model\ over\ Full\ model} = 0.07539 > 0.05$$

With the P value above 0.05, we can conclude the reduced model is a better predictor of Life Expectancy than the full model. With the conclusion, we proceed with the rest of the test we have done in step A2 to A3.

D2. Hypothesis on Individual Coefficient Test (T Test)

$$H_O\ (Null\ Hypothesis): \beta_1 = \beta_2 = \beta_3 = .. = \beta_i = 0$$
$$H_a\ (Alternative\ Hypothesis): \text{At least one } \beta_i \neq 0$$
Where i is number from 1 to 4

Full Model Individual Coefficient Test

Below variables had the p value smaller than the $\alpha = 0.05$ limit which indicates that they have significant impact on the Life Expectancy at $\alpha = 0.05$ level.

$$P_{Income\ Composition\ of\ Resources} = 2 \times 10^{-16}$$
$$P_{Adult\ Mortality} = 6.69 \times 10^{-7}$$
$$P_{HIV\ AIDS} = 0.000422$$

From the individual coefficient test, we can see that the P value of the Infant Death exceeds the 0.05 limit and should be removed.

D3. 95% Confidence Interval

$$Income\ Composition\ of\ Resources: (32.299087402\ to\ 40.454939123)$$
$$Adult\ Mortality: (-0.025488178\ \text{to}\ -0.011511921)$$
$$HIV\ AIDS: (-1.189838655\ to\ -0.349242385)$$
$$Infant\ Death: (-0.007459032\ to\ 0.002880345)$$

Confirming the individual coefficient test, the Infant Death is the only variables that contains 0 in the 95% confidence interval. Arising from part C and part D of the analysis, we have decided to remove both Under Fiver Death and Infant Death variables to make a first order model with three independent variables. We would repeat the step C1 to C3 to confirm that our final first order model is significant.

4.5 Reduced Model Analysis, 3 Variables, (Part E)

To test that a particular subset of q of the coefficients are zero, the hypotheses take the form:

E1. ANOVA Test

$$H_O : (Null\ Hypothesis): \beta_{p-q+1} = \beta_{p-q+2} = \cdots = \beta_p = 0$$
$$H_a : (Alternative\ Hypothesis): \text{At least one } \beta_i \neq 0$$
$$P_{Reduced\ model\ over\ Full\ model} = 0.08575 > 0.05$$

With the P value above 0.05, we can conclude the reduced model is a better predictor of Life Expectancy than the full model. With the conclusion, we proceed with the rest of the test we have done in step A2 to A3.

E2. Hypothesis on Individual Coefficient Test (T Test)

$$H_O\ (Null\ Hypothesis): \beta_1 = \beta_2 = .. = \beta_i = 0$$
$$H_a\ (Alternative\ Hypothesis): \text{At least one } \beta_i \neq 0$$
$$\text{Where i is number from 1 to 3}$$

Full Model Individual Coefficient Test

Below variables had the p value smaller than the $\alpha = 0.05$ limit which indicates that they have significant impact on the Life Expectancy at $\alpha = 0.05$ level.

$$P_{Income\ Composition\ of\ Resources} = 2 \times 10^{-16}$$
$$P_{Adult\ Mortality} = 4.59 \times 10^{-7}$$
$$P_{HIV\ AIDS} = 0.000488$$

From the individual coefficient test, we can see that all variables have their individual P valued less than 0.05 limit.

E3. 95% Confidence Interval

$$Income\ Composition\ of\ Resources: (32.59829485\ to\ 40.66465030)$$
$$Adult\ Mortality: (-0.02568627 \text{ to } -0.01176048)$$
$$HIV\ AIDS: (-1.17745844\ \ to\ -0.33922883)$$

None of the variables have 0 in their 95% confidence interval which indicates that this model E stands. Assumptions were checked, but this model did not pass the Shapiro-Wilks test and the Breusch-Pagan test.

4.6 Interaction Model Analysis (Part F)

Full interaction terms have been developed to include the following variables.

(1) Income Composition of Resources, (2) Adult Mortality, (3) HIV AIDS, (4) Income Composition of Resources* Adult Mortality, (5) Income Composition of Resources* HIV AIDS , (5) Adult Mortality * HIV AIDS. The full interaction term will undergo ANOVA test, individual coefficient test and 95% confidence interval confirmed to determine inf further reduction is required.

F1. ANOVA Test

$$\mathrm{H_O}: (Null\ Hypothesis): \beta_{p-q+1} = \beta_{p-q+2} = \cdots = \beta_p = 0$$
$$\mathrm{H_a}: (Alternative\ Hypothesis): \text{At least one } \beta_i \neq 0$$
$$P_{Reduced\ model\ over\ Full\ model} = 0.9956 > 0.05$$

With the P value above 0.05, we can conclude the full interaction model is a better predictor of Life Expectancy than the full model. With the conclusion, we proceed with the rest of the test.

F2. Hypothesis on Individual Coefficient Test (T Test)

$$\mathrm{H_O}\ (Null\ Hypothesis): \beta_1 = \beta_2 = .. = \beta_i = 0$$
$$\mathrm{H_a}\ (Alternative\ Hypothesis): \text{At least one } \beta_i \neq 0$$
$$\text{Where i is number from 1 to 6}$$

Full Model Individual Coefficient Test

Below variables had the p value smaller than the $\alpha = 0.05$ limit which indicates that they have significant impact on the Life Expectancy at $\alpha = 0.05$ level.

$$P_{Income\ Composition\ of\ Resources} = 2 \times 10^{-16}$$
$$P_{Adult\ Mortality*HIV\ AIDS} = 0.000997$$

From the individual coefficient test, we can see that only two variables have their individual P valued less than 0.05 limit. This signals that interaction model needs to be revised.

F3. 95% Confidence Interval

$$Income\ Composition\ of\ Resources: (30.334174183\ \ to\ 42.801060260\ )$$
$$Adult\ Mortality: (-0.041164108\ \text{ to } 0.011623260\ )$$
$$HIV\ AIDS: (-5.068231925\ \ to\ 1.708439258\ )$$
$$Income\ Composition\ of\ Resources * Adult\ Mortality: (-0.063503967\ \ to\ 0.023173393\ )$$
$$Income\ Composition\ of\ Resources * HIV\ AIDS: (-6.656579993\ \ to\ 4.145374454\ )$$
$$Adult\ Mortality * HIV\ AIDS: (0.001962741\ \ to\ 0.007537992\ )$$

All except the Income Composition of Resources and interaction of Adult Mortality and HIV AIDS have 0 in their 95% confidence interval which indicates that the interaction model needs to be

reduced. In determining which variables to remove, we turn back to the stepwise regression technique. Below outlines the result of the stepwise model of the interaction model.

F4. Stepwise Regression (Both) Model

$$Variable\ 1: Income\ Composition\ of\ Resources$$
$$Variable\ 2: HIV\ AIDS$$
$$Variable\ 3: Adult\ Mortality$$
$$Variable\ 4: Income\ Composition\ of\ Resources * Adult\ Mortality$$

F5. Stepwise Regression (Forward) Model

$$Variable\ 1: Income\ Composition\ of\ Resources$$
$$Variable\ 2: HIV\ AIDS$$
$$Variable\ 3: Adult\ Mortality$$
$$Variable\ 4: Income\ Composition\ of\ Resources * Adult\ Mortality$$
$$Variable\ 5: Adult\ Mortality * HIV\ AIDS$$

F6. Stepwise Regression (Backward) Model

$$Variable\ 1: Income\ Composition\ of\ Resources$$
$$Variable\ 2: Adult\ Mortality$$
$$Variable\ 3: HIV\ AIDS$$
$$Variable\ 4: Adult\ Mortality * HIV\ AIDS$$

As the results show, we have different model selection for all three stepwise model developed. In determining the best interaction, we use the CP, AIC and Adjusted R squared value to assess the criterion.

F7. Mallow's CP/ Akaike's Information Criterion (AIC)/Adjusted R Squared Criterion

	CP	AIC	$AdjustedR2$
[1,]	110.8498	692.0198	0.8038996
[2,]	33.66209	641.3787	0.8689818
[3,]	14.10498	624.2984	0.8862158
[4,]	4.185469	614.3838	0.8954863
[5,]	5.211830	615.3598	0.8954693
[6,]	7.000000	617.1359	0.8947896

The best model would have low CP and AIC values and high Adjusted R squared values. CP values of the full model has been analyzed to support the stepwise regression model. Based on the results above, CP value of 4 variables is showing the lowest values. AIC values are also analyzed, and 4 variables are showing the lowest AIC value. Adjusted R squared value is highest for the 4 variables. As a result of analyzing the criterion of stepwise model, we can proceed with four variable model.

In utilizing the selection algorithm, the four variable that can be choose for the interaction model is (1) Income Composition of Resources, (2) Adult Mortality, (3) Income Composition of Resources*Adult Mortality and (4) Income Composition of Resources*HIV AIDS. We have decided however that the main variable HIV AIDS should be part of the model so we have used the individual T test of the full interaction model to determine the best fit. As a result, the revised interaction model contains the following variables.

(1) Income Composition of Resources, (2) Adult Mortality, (3) HIV Aids, (4) Adult Mortality*HIV AIDS.

4.7 Revised Interaction Model Analysis (Part G)

In determining whether our final interaction model is significant, we undergo the ANOVA test to compare it with the full model, individual coefficient test and it's 95% confidence interval to make sure all variables are significant as well run the stepwise regression to make sure no further change is required.

G1. ANOVA Test

$$H_O : (Null\ Hypothesis): \beta_{p-q+1} = \beta_{p-q+2} = \cdots = \beta_p = 0$$
$$H_a : (Alternative\ Hypothesis): \text{At least one } \beta_i \neq 0$$
$$P_{Reduced\ model\ over\ Full\ model} = 0.9663 > 0.05$$

With the P value above 0.05, we can conclude the revised interaction model is a better predictor of Life Expectancy than the full model. With the conclusion, we proceed with the rest of the test.

G2. Hypothesis on Individual Coefficient Test (T Test)

$$H_O\ (Null\ Hypothesis): \beta_1 = \beta_2 = .. = \beta_i = 0$$
$$H_a\ (Alternative\ Hypothesis): \text{At least one } \beta_i \neq 0$$
Where i is number from 1 to 5

Full Model Individual Coefficient Test

Below variables had the p value smaller than the $\alpha = 0.05$ limit which indicates that they have significant impact on the Life Expectancy at $\alpha = 0.05$ level.

$$P_{Income\ Composition\ of\ Resources} = 2 \times 10^{-16}$$
$$P_{Adult\ Mortality} = 6.92 \times 10^{-11}$$
$$P_{HIV\ AIDS} = 1.98 \times 10^{-7}$$
$$P_{Adult\ Mortality*HIV\ AIDS} = 3.37 \times 10^{-5}$$

From the individual coefficient test, we can see that all variables have their individual P valued less than 0.05 limit. This signals that interaction model is a good fit.

G3. 95% Confidence Interval

$$Income\ Composition\ of\ Resources: (29.216340098\ to\ 37.35891546)$$

$$Adult\ Mortality: (-0.034606774\ \text{to}\ -0.01960134\ )$$
$$HIV\ AIDS: (-3.405744334\ \ to -1.60673210\ )$$
$$Adult\ Mortality * HIV\ AIDS: (0.002935985\ \ to\ 0.00793218\ )$$

None of the variables have 0 in their 95% confidence interval which indicates that all the variables of the revised interaction model is significant.

G4. Stepwise Regression (Both) Model

$$Variable\ 1: Income\ Composition\ of\ Resources$$
$$Variable\ 2: Adult\ Mortality$$
$$Variable\ 3: HIV\ AIDS$$
$$Variable\ 4: Adult\ Mortality * HIV\ AIDS$$

G5. Stepwise Regression (Forward) Model

$$Variable\ 1: Income\ Composition\ of\ Resources$$
$$Variable\ 2: Adult\ Mortality$$
$$Variable\ 3: HIV\ AIDS$$
$$Variable\ 4: Adult\ Mortality * HIV\ AIDS$$

G6. Stepwise Regression (Backward) Model

$$Variable\ 1: Income\ Composition\ of\ Resources$$
$$Variable\ 2: Adult\ Mortality$$
$$Variable\ 3: HIV\ AIDS$$
$$Variable\ 4: Adult\ Mortality * HIV\ AIDS$$

All the models with different stepwise regression results in same variables. We use the CP, AIC and Adjusted R squared value to assess the criterion.

G5. Mallow's CP/ Akaike's Information Criterion (AIC)/Adjusted R Squared Criterion

	CP	AIC	$AdjustedR2$
[1,]	109.6179	687.3854	0.8031101
[2,]	33.67568	637.5609	0.8680253
[3,]	21.54336	627.3925	0.8791086
[4,]	5.00000	611.4226	0.8941989

Based on the CP, AIC and adjusted R squared values, the best model with lowest CP value, lowest AIC value and highest Adjusted R squared value is obtained with the 4 variable model which contains all the base variables with the interaction term of Adult Mortality and HIV AIDS. All assumptions tests carried out passed except the Breusch-Pagan test for homoscedasticity. In confirming the final interaction model, we now assess for the possibility of higher order terms.

4.8 Higher Order Model Analysis Part 1 (Part H)

We first plot the variables against each other to determine if any higher order term may exist.

H1. Variable Plot Analysis

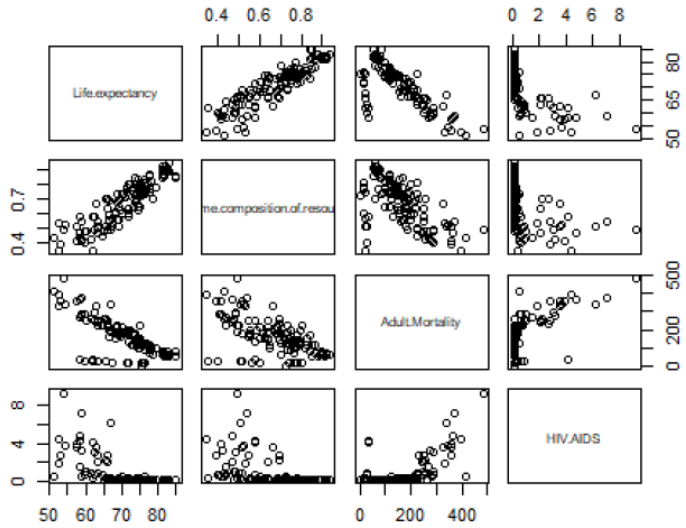


Figure 6: Variable Plot

From the plot, we can see that the Income Composition of Resources and HIV AIDS have nonlinear relationship with Life Expectancy. We would then develop a new model to incorporate these higher order terms. First higher order model developed will have HIV AIDS as quadratic term. All the appropriate testing including ANOVA test, T test and its confidence interval is performed again to determine the significance of the higher order model.

H2. ANOVA Test

$$H_O : (Null\ Hypothesis): \beta_{p-q+1} = \beta_{p-q+2} = \cdots = \beta_p = 0$$
$$H_a : (Alternative\ Hypothesis): \text{At least one } \beta_i \neq 0$$
$$P_{Reduced\ model\ over\ Full\ model} = 0.9442 > 0.05$$

With the P value above 0.05, we can conclude the higher order model is a better predictor of Life Expectancy than the full model. With the conclusion, we proceed with the rest of the test.

H3. Hypothesis on Individual Coefficient Test (T Test)

$$H_O\ (Null\ Hypothesis): \beta_1 = \beta_2 = .. = \beta_i = 0$$
$$H_a\ (Alternative\ Hypothesis): \text{At least one } \beta_i \neq 0$$
Where i is number from 1 to 5

Full Model Individual Coefficient Test

Below variables had the p value smaller than the $\alpha = 0.05$ limit which indicates that they have significant impact on the Life Expectancy at $\alpha = 0.05$ level.

$$P_{Income\ Composition\ of\ Resources} = 2 \times 10^{-16}$$
$$P_{Adult\ Mortality} = 4.23 \times 10^{-9}$$
$$P_{HIV\ AIDS} = 8.77 \times 10^{-7}$$
$$P_{Adult\ Mortality*HIV\ AIDS} = 0.00421$$

From the individual coefficient test, we can see that quadratic coefficient has its individual P value greater than 0.05 which indicates that with 95% confidence, the quadratic term of HIV AIDS isn't significant with Life Expectancy.

H4. 95% Confidence Interval

$$Income\ Composition\ of\ Resources: (29.193192634 \quad to\ 37.420373658\ )$$
$$Adult\ Mortality: (-0.035795200 \quad \text{to} \quad -0.018748194\ )$$
$$HIV\ AIDS: (-3.445655721 \quad to -1.541508554\ )$$
$$Adult\ Mortality * HIV\ AIDS: (0.001784496\ \ to\ 0.009320143\ \ )$$
$$HIV\ AIDS^2: (-0.191146489 \quad to\ 0.175717357\ )$$

Confirming the T test, we see that quadratic term of HIV AIDS has a 0 in its 95% confidence interval. We cannot use such quadratic term. The next model will analyze the quadratic term of Income Composition of Resources.

4.9 Higher Order Model Analysis Part 2 (Part I)

Part H is repeated for Income Composition of Resources quadratic term.

I1. ANOVA Test

$$\text{H}_\text{O}: (Null\ Hypothesis): \beta_{p-q+1} = \beta_{p-q+2} = \cdots = \beta_p = 0$$
$$\text{H}_\text{a}: (Alternative\ Hypothesis): \text{At least one } \beta_i \neq 0$$
$$P_{Reduced\ model\ over\ Full\ model} = 0.946 > 0.05$$

With the P value above 0.05, we can conclude the higher order model is a better predictor of Life Expectancy than the full model. With the conclusion, we proceed with the rest of the test.

I2. Hypothesis on Individual Coefficient Test (T Test)

$$\text{H}_\text{O}\ (Null\ Hypothesis): \beta_1 = \beta_2 = .. = \beta_i = 0$$
$$\text{H}_\text{a}\ (Alternative\ Hypothesis): \text{At least one } \beta_i \neq 0$$
$$\text{Where i is number from 1 to 5}$$

Full Model Individual Coefficient Test

Below variables had the p value smaller than the $\alpha = 0.05$ limit which indicates that they have significant impact on the Life Expectancy at $\alpha = 0.05$ level.

$$P_{Adult\ Mortality} = 1.14 \times 10^{-10}$$
$$P_{HIV\ AIDS} = 2.59 \times 10^{-7}$$
$$P_{Adult\ Mortality*HIV\ AIDS} = 4.79 \times 10^{-5}$$
$$P_{Income\ Composition\ of\ Resources} = 0.0288$$

From the individual coefficient test, we can see that quadratic coefficient has its individual P value greater than 0.05 which indicates that with 95% confidence, the quadratic term of Income Composition of Resources isn't significant with Life Expectancy.

I3. 95% Confidence Interval

$$Income\ Composition\ of\ Resources: (3.174217575 \quad to\ 57.237342780)$$
$$Adult\ Mortality: (-0.034586755 \quad \text{to} \quad -0.019435521\ )$$
$$HIV\ AIDS: (-3.403914671 \quad to -1.592076980\ )$$
$$Adult\ Mortality * HIV\ AIDS: (0.002862208\ \ to\ 0.007926009\ )$$
$$Income\ Composition\ of\ Resources^2: (-18.094488962\ to\ 22.812517109)$$

Confirming the T test, we see that quadratic term of Income Composition of Resources has a 0 in its 95% confidence interval. We cannot use such quadratic term. No higher order term is utilized for our model.

4.10 Box Cox Transformation (Part J)

After confirming that no higher order is required for the final model, we look to add box cox transformation due to the heteroscedasticity of the revised interaction model. Using the "boxcox" function, the lambda value required for the transformation is found to be 1.7474. Once the new model is developed, we revisit the individual T test as well find the 95% confidence interval to confirm that new model is significant with Life Expectancy.

J1. Hypothesis on Individual Coefficient Test (T Test)

$$\text{H}_\text{O}\ (Null\ Hypothesis): \beta_1 = \beta_2 = .. = \beta_i = 0$$
$$\text{H}_\text{a}\ (Alternative\ Hypothesis): \text{At least one } \beta_i \neq 0$$
$$\text{Where i is number from 1 to 5}$$

Full Model Individual Coefficient Test

Below variables had the p value smaller than the $\alpha = 0.05$ limit which indicates that they have significant impact on the Life Expectancy at $\alpha = 0.05$ level.

$$P_{Income\ Composition\ of\ Resources} = 2 \times 10^{-16}$$
$$P_{Adult\ Mortality} = 6.42 \times 10^{-11}$$
$$P_{HIV\ AIDS} = 2.15 \times 10^{-7}$$

$$P_{Adult\ Mortality*HIV\ AIDS} = 1.41 \times 10^{-5}$$

From the individual coefficient test, we can see that all the P values are less than 0.05 which indicates that the model from box cox transformation is significant.

J2. 95% Confidence Interval

$$Income\ Composition\ of\ Resources: (708.40678683\ to\ 901.1455019)$$
$$Adult\ Mortality: (-0.82045870\ \text{to}\ -0.4652729)$$
$$HIV\ AIDS: (-80.41612024\ to\ -37.8326265)$$
$$Adult\ Mortality * HIV\ AIDS: (0.07605862\ to\ 0.1943210)$$

None of the confidence interval has 0 values in between which does support the individual coefficient test we performed.

Subsequently, various Box-Cox transformations were carried out until we arrived at the best model that has been selected as the representative model for estimating life expectancy. This final model is discussed below.

4.11 Revised Interaction Model With Box-Cox Transformation (Part P)

We later decided to carry out other box cox transformations due to the heteroscedasticity of the revised interaction model. However, we included "under.five.deaths" and "Hepatitis.B" to the linear model G. Data points 119, 98 and 53 were removed as a result of Cook's distance check of the data.

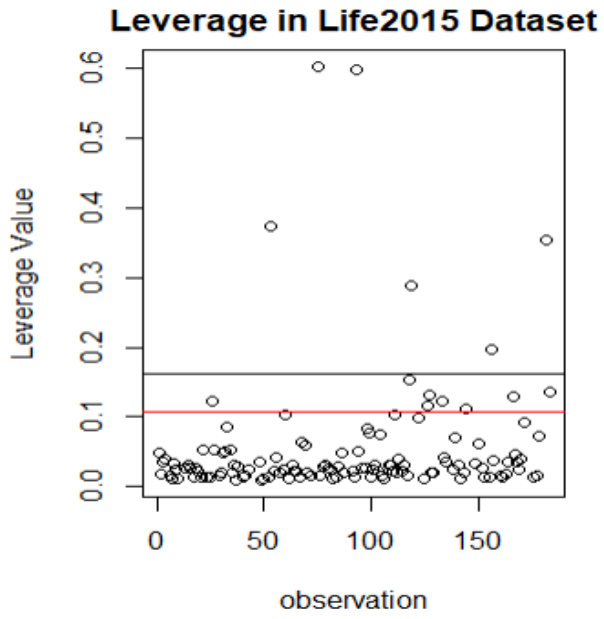


Figure 7: Leverage Plot for model P

The individual t-tests were carried out to identify the independent variables that were significant in predicting life expectancy.

P1. Hypothesis on Individual Coefficient Test (T Test)

$$H_O\ (Null\ Hypothesis): \beta_1 = \beta_2 = .. = \beta_i = 0$$
$$H_a\ (Alternative\ Hypothesis): \text{At least one } \beta_i \neq 0$$
Where i is number from 1 to 6

Full Model Individual Coefficient Test

Below variables had the p value smaller than the $\alpha = 0.05$ limit which indicates that they have significant impact on the Life Expectancy at $\alpha = 0.05$ level.

$$P_{Income\ Composition\ of\ Resources} = 2 \times 10^{-16}$$
$$P_{Adult\ Mortality} = 5.54 \times 10^{-11}$$
$$P_{HIV\ AIDS} = 4.34 \times 10^{-5}$$
$$P_{Adult\ Mortality*HIV\ AIDS} = 3.47 \times 10^{-4}$$
$$P_{Hepatitis.B} = 2.18 \times 10^{-2}$$

From the individual coefficient test, we can see that all the P values except "under.five.deaths" are less than 0.05 which indicates that the model from box cox transformation is significant.

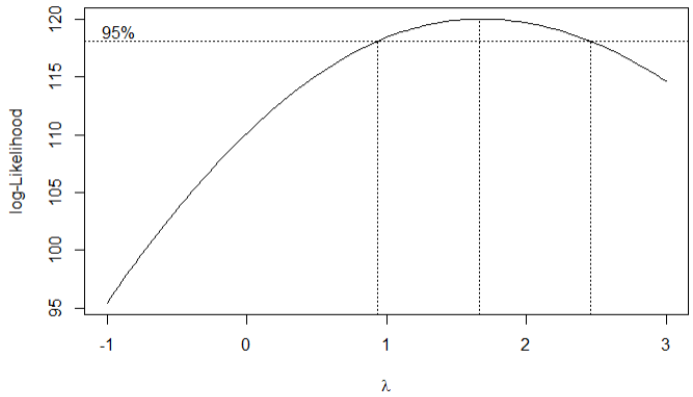


Figure 8: Box-Cox plot for model P

```
[1] "t-test of previous Box-Cox transformation indicated no significance of
interaction term after this transformation"
[1] 1.666667
[1] "The best lambda for Box-Cox Transformation is:  1.66666666666667"
[1] "From the output, as the best lambda would be 1.66667."
```

P2. 95% Confidence Interval

$$Income\ Composition\ of\ Resources: (492.41829099\ \ to\ 628.26121039\ )$$
$$Adult\ Mortality: (-0.57528254\ \text{to}\ -0.32729041\ )$$
$$HIV\ AIDS: (-50.74805455\ \ to\ -18.46407368\ )$$
$$Adult\ Mortality * HIV\ AIDS: (0.03709493\ to\ 0.12335954\ )$$
$$Hepatitis.B: (0.06213597\ to\ 0.77581510)$$

The 95% confidence interval for "under.five.deaths" crossed the zero point between the upper and lower values and seems not to be suitable for the analysis, but had passed in the first order model (Refer to part A).

Various plots were created to check the assumptions made in the Methodology section. They are displayed below:

Linearity Assumptions:

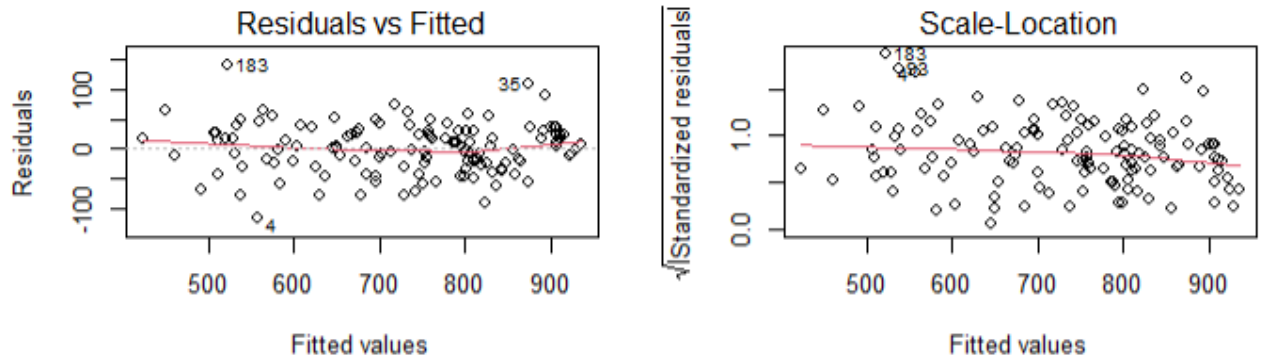


Figure 9: Residual and Scale-Location plots confirm that linearity assumption is met.

The above residual and scale-location plots clearly shows that the residuals seem to be similarly spaced on either sides of the red line with no prominent shape observed. Therefore the linearity assumption is met.

Independence Assumptions:

Since we did not analyze a time series data, for the purpose of this study, the independence assumption holds.

Normality Assumptions:

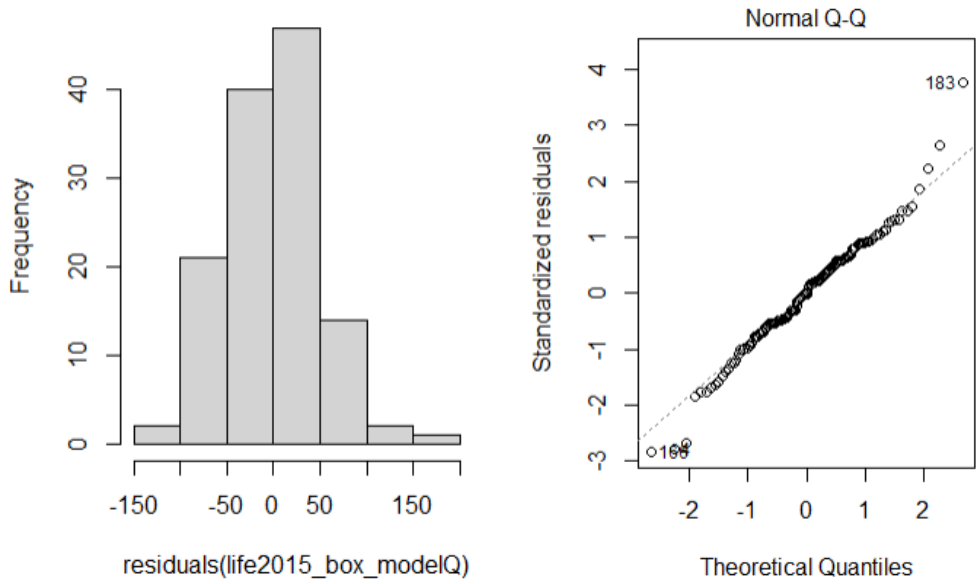


Figure 10: Histogram of residuals and the Normal Q-Q Plot of modelP

The histogram plot of the residuals and the Normal Q-Q plot shows that the assumption holds for our dataset. This was confirmed by the Shapiro-Wilks test (W= 0.99142 , p-value = 0.6263), so we fail to reject the null hypothesis.

Equal Variance Assumptions:

In this case the residual plot is utilized again and confirmed with the Breusch-Pagan test. The BP test gave a result of 0.05693 > α =0.05 which indicates that heteroscedasticity is not present.

Our final, best-fitted model includes main effects, interactions and Box-Cox transformation and the summary of which is:

```
lm(formula = (((Lifeexpectancy^1.66667) - 1)/1.66667) ~ Income.composition.of.resources +
    Adult.Mortality + HIV.AIDS + under.five.deaths + Hepatitis.B +
    Adult.Mortality * HIV.AIDS, data = life2015[-c(119, 53, 93),
    ])

Residuals:
     Min       1Q   Median       3Q      Max
-116.902  -25.632    2.449   27.219  141.008

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                   401.02873   33.47769  11.979  < 2e-16 ***
Income.composition.of.resources 560.33975   34.30497  16.334  < 2e-16 ***
Adult.Mortality                -0.45129    0.06263  -7.206 5.54e-11 ***
HIV.AIDS                      -34.60606    8.15281  -4.245 4.34e-05 ***
under.five.deaths              -0.03015    0.03156  -0.955 0.341398
Hepatitis.B                     0.41898    0.18023   2.325 0.021769 *
Adult.Mortality:HIV.AIDS        0.08023    0.02178   3.683 0.000347 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 43.1 on 120 degrees of freedom
Multiple R-squared:  0.9032,    Adjusted R-squared:  0.8984
F-statistic: 186.7 on 6 and 120 DF,  p-value: < 2.2e-16
```

Figure 11: Summary of final modelP

Therefore, the relationship will take the form:

$$(\hat{Y}^{1.667}_{\text{Life Expectancy}} - 1)/1.667 = \hat{\beta}_0 + \hat{\beta}_1 X_{\text{Income Composition of Resource}} + \hat{\beta}_2 X_{\text{Adult Mortality}} + \hat{\beta}_3 X_{\text{HIV-AIDS}} + \hat{\beta}_4 X_{\text{Under Five Death}} + \hat{\beta}_5 X_{\text{Hepatitis B}} + \hat{\beta}_6 X_{\text{Adult Mortality}} * X_{\text{HIV-AIDS}} + \varepsilon$$

4.12 Revised Interaction Model With Box-Cox Transformation (Part R)

We also decided to carry out other box cox transformations due to the heteroscedasticity of the revised interaction model. However, we only included “Hepatitis.B” and “the interaction term between Adult.Mortality and Hepatitis.B” to the linear model E. Data points #53, #93, #119, were removed as a result of Cook’s distance check of the data.

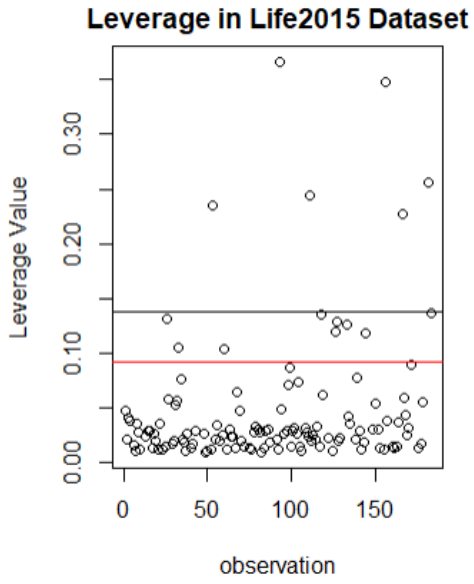


Figure 12: Leverage Plot for model R

The individual t-tests were carried out to identify the independent variables that were significant in predicting life expectancy.

P1. Hypothesis on Individual Coefficient Test (T Test)

$$H_O\ (Null\ Hypothesis): \beta_1 = \beta_2 = .. = \beta_i = 0$$
$$H_a\ (Alternative\ Hypothesis): \text{At least one } \beta_i \neq 0$$
$$\text{Where i is number from 1 to 6}$$

Full Model Individual Coefficient Test

Below variables had the p value smaller than the $\alpha = 0.05$ limit which indicates that they have significant impact on the Life Expectancy at $\alpha = 0.05$ level.

$$P_{Income\ Composition\ of\ Resources} = 2 \times 10^{-16}$$
$$P_{Adult\ Mortality} = 0.957337$$
$$P_{HIV\ AIDS} = 0.06667$$
$$P_{Adult\ Mortality*Hepatitis.B} = 0.000644$$
$$P_{Hepatitis.B} = 2.45 \times 10^{-6}$$

From the individual coefficient test, we can see that all the P values except “Adult.Mortality” and “HIV.AIDS” are less than 0.05 which indicates that the model from box cox transformation is significant.

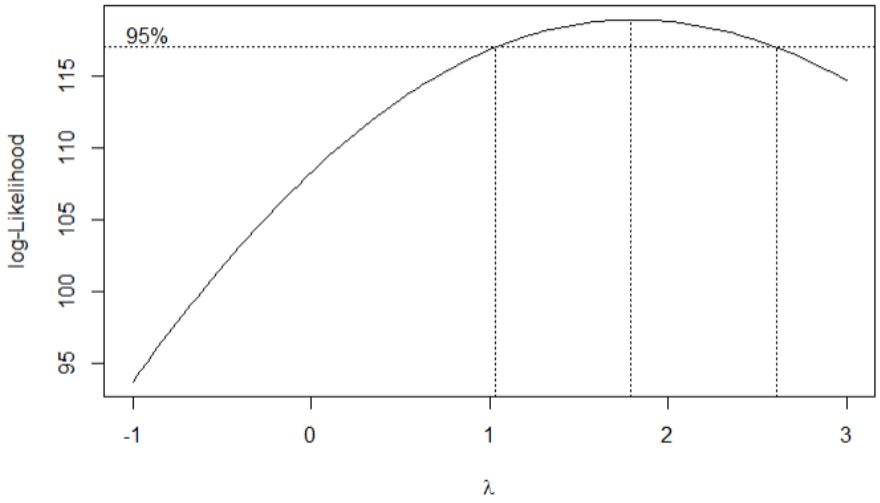


Figure 13: Box-Cox plot for model R

```
[1] "t-test of previous Box-Cox transformation indicated no significance of
interaction term after this transformation"
[1] "The best lambda for Box-Cox Transformation is:  1.78787878787879"
[1] "From the output, as the best lambda would be 1.78787."
```

P2. 95% Confidence Interval

$$Income\ Composition\ of\ Resources: (873.37802078\ \ to\ 1.092566e+03\ )$$
$$Adult\ Mortality: (-0.38865292\ \ \text{to}\ 3.681605e-01\ )$$
$$HIV\ AIDS: (-22.19896293\ \ to\ \ 7.499017e-01\ )$$
$$Adult\ Mortality * Hepatitis.B: (-0.01302083\ \ to\ \ -3.618418e-03\ )$$
$$Hepatitis.B: (1.49086362\ \ to\ \ 3.479816)$$

The 95% confidence interval for "Adult.Mortality" and "HIV.AIDS" crossed the zero point between the upper and lower values and hence seems not suitable for the analysis, but had passed in the first order model (Refer to Appendix ModelR and Part A).

Various plots were created to check the assumptions made in the Methodology section. They are displayed below:

Linearity Assumptions:

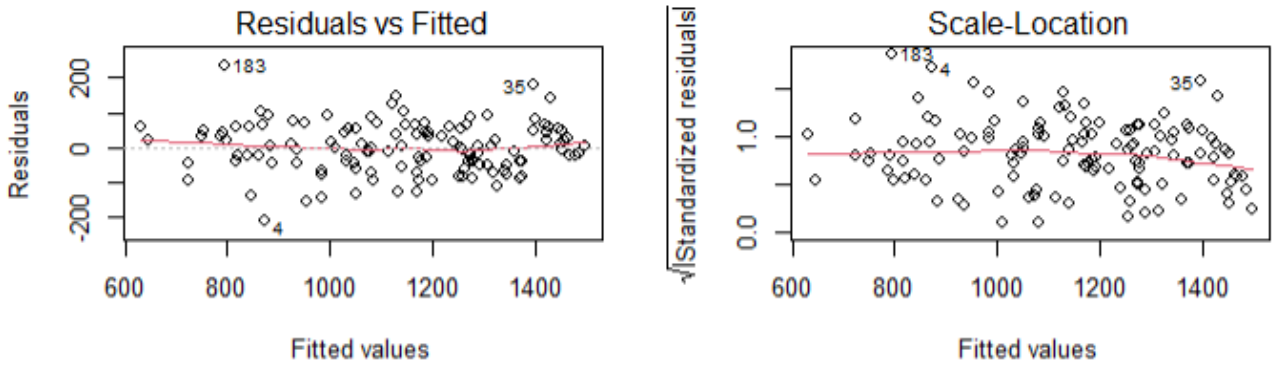


Figure 14: Residual and Scale-Location plots confirm that linearity assumption is met.

The above residual and scale-location plots clearly shows that the residuals seem to be similarly spaced on either side of the red line with no prominent shape observed. Therefore, the linearity assumption is met.

Independence Assumptions:

Since we did not analyze a time series data, for the purpose of this study, the independence assumption holds.

Normality Assumptions:

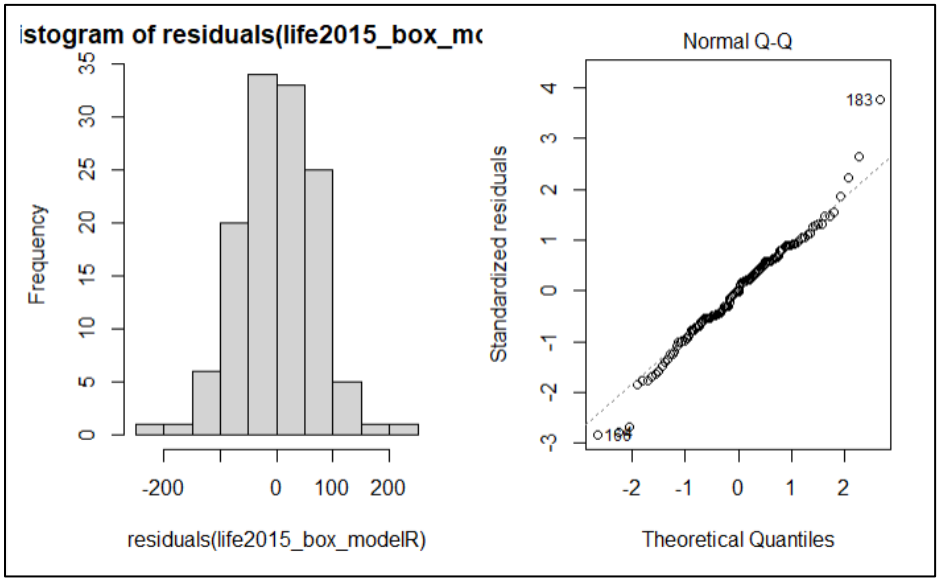


Figure 15: Histogram of residuals and the Normal Q-Q Plot of modelR

The histogram plot of the residuals and the Normal Q-Q plot shows that the assumption holds for our dataset. This was confirmed by the Shapiro-Wilks test (W= 0.99243, p-value = 0.7269), so we fail to reject the null hypothesis.

Equal Variance Assumptions:

In this case the residual plot is utilized again and confirmed with the Breusch-Pagan test. The BP test gave a result of 0.1213 > α =0.05 which indicates that heteroscedasticity is not present.

Our final, best-fitted model includes main effects, interactions and Box-Cox transformation and the summary of which is:

```
Call:
lm(formula = (((Lifeexpectancy^1.78787) - 1)/1.78787) ~
Income.composition.of.resources +
    Adult.Mortality + HIV.AIDS + Hepatitis.B + Adult.Mortality *
    Hepatitis.B, data = life2015[-c(119, 53, 93), ])

Residuals:
     Min       1Q   Median       3Q      Max
-208.919  -45.093    1.978   50.646  234.400

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                   392.661654  55.300722   7.100 9.24e-11 ***
Income.composition.of.resources 982.971792  55.356994  17.757  < 2e-16 ***
Adult.Mortality                -0.010246   0.191137  -0.054 0.957337
HIV.AIDS                      -10.724531   5.795859  -1.850 0.066698 .
Hepatitis.B                     2.485340   0.502321   4.948 2.45e-06 ***
Adult.Mortality:Hepatitis.B    -0.008320   0.002375  -3.504 0.000644 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 72.55 on 121 degrees of freedom
Multiple R-squared:  0.9011,    Adjusted R-squared:  0.897
F-statistic: 220.4 on 5 and 121 DF,  p-value: < 2.2e-16
```

Figure 16: Summary of final modelR

Therefore, the relationship will take the form:

$$(\hat{Y}^{1.788}_{\text{Life Expectancy}} - 1)/1.788 = \hat{\beta}_0 + \hat{\beta}_1 X_{\text{Income Composition of Resource}} + \hat{\beta}_2 X_{\text{Adult Mortality}} + \hat{\beta}_3 X_{\text{HIV-AIDS}} + \hat{\beta}_4 X_{\text{Hepatitis.B}} + \hat{\beta}_5 X_{\text{Adult Mortality}} * X_{\text{Hepatitis.B}}$$

5. Model Prediction

After arriving at our model for this project, it was thereafter applied to try and predict the average life expectancy for the year 2015 and the following results were obtained with Model P:

```
# MODEL P
predict(life2015_box_modelP, meanData, interval = "predict")

##        fit      lwr    upr
## 1 722.5345 636.7589 808.31

# Converting to Life.Expectancy (Box-Cox Lambda)
pred_Mean_Life_P_fit = (722.5345 * 1.66667 + 1)^(1/1.66667)
pred_Mean_Life_P_lwr = (636.7589 * 1.66667 + 1)^(1/1.66667)
pred_Mean_Life_P_upr = (808.31 * 1.66667 + 1)^(1/1.66667)
```

Figure 17: Model P prediction of world Life Expectancy for 2015 using mean data

Using the mean data of significant variables of Model P as input, the Life Expectancy around the World in 2015 based on Model P prediction is on the average 70.5728282292507 years at α = 0.05 level. For 95% prediction interval, the life expectancy estimated by Model P is between 65.4239214422687 and 75.4824759745119".

A similar estimate was carried out with Model R prediction of the average Life Expectancy of the world in 2015.

```
# MODEL R
predict(life2015_box_modelR, meanData, interval = "predict")

##        fit      lwr    upr
## 1 1139.689 995.4774 1283.9

# Converting to Life.Expectancy (Box-Cox Lambda)
pred_Mean_Life_R_fit = (1139.689 * 1.78787 + 1)^(1/1.78787)
pred_Mean_Life_R_lwr = (995.4774 * 1.78787 + 1)^(1/1.78787)
pred_Mean_Life_R_upr = (1283.9 * 1.78787 + 1)^(1/1.78787)
```

Figure 18: Model R prediction of world Life Expectancy for 2015 using mean data

Using the mean data of significant variables of Model R as input, the Life Expectancy around the World in 2015 based on Model R prediction is on the average 70.9566810821455 years at α =

0.05 level. For 95% prediction interval, the life expectancy estimated by Model P is between 65.7881166846681 and 75.844148896526".

Both models are pretty good in predicting the average life expectancy of the world in the year 2015 and the estimated values of these predictions are very close to one another.

6. Conclusions

The analysis done in the WHO dataset for 2015 indicated that the variables Income Composition of Resources, Adult Mortality and HIV/AIDS present significant influence in the life expectancy irrespective of the statistical method used at the beginning of the analysis. However, Stepwise Selection Procedure and Best Subset indicated Hepatitis B and important variable at 5 % level (Model R), whereas t-test indicated to consider Under Five deaths (Model M) or Infants deaths (Model N). A mixed model (Model P) also considering the interaction term of Adult.Mortality * HIV.AIDS presented in the model and was tested to be a good model as well. This model passed all the linear regression assumptions of linearity and normality. For equal variance of the residuals assumption (homoscedasticity), this was only possible for it to pass after considering Box Cox transformation in all the models. Based on this and considering that Model P as the one of the valid models that presented the highest adjusted coefficient of 0.8984 and the lowest AIC (1325). We recommend this model as the final model to be used in predicting the Life Expectancy in 2015. The summary of the various models that were being considered are displayed below in Table 2:

Model	VARIABLES	MODEL DIAGNOSTICS					R^2	adj. R^2	cP	AIC
		Linearity	Normality	Homoscedasticity	Multicolinearity	Outliers				
A	All Variables		0.2905	0.333	max 178		0.9005	0.8865		
B	Life ~ Income + Adult.Mortality + Infant.deaths + under.five + HIV		0.01248	0.0157	max 90	0.8 (#119)	0.8863	0.8817	6	639.8
C	Life ~ Income + Adult.Mortality + under.five + HIV						0.8829	0.8791		
D	Life ~ Income + Adult.Mortality + Infant.deaths + HIV						0.8825	0.8787		
E	Life ~ Income + Adult.Mortality + HIV		0.02022	0.001146	max 2.02	0.52 (#53)	0.8817	0.8789		
F	Life ~(Income + Adult.Mortality + HIV)^2		0.01935	0.06245	All 3 interaction coef	1.2 (#93)	0.8998	0.8948		
G	Life ~ Income + Adult.Mortality + HIV + Adult.Mortality * HIV		0.05849	0.01695	Int. term witm main		0.8976	0.8942	5	611.4
H	Life ~ Income + Adult.Mortality + HIV + Adult.Mortality * HIV + HIV^2						0.8976	0.8933		
I	Life ~ Income + Adult.Mortality + HIV + Adult.Mortality * HIV + Income^2						0.8976	0.8934		
J	Box Cox Transformation for Model G (Lambda = 1.7474)		0.1559	0.04099	Int. term witm main		0.8977	0.8944		
K	Box Cox Transformation for Model E (Lambda = 1.5858)		0.07322	0.0005274			0.8812	0.8783		
L	Box Cox Transformation for Model C (Lambda = 1.5454)		0.04811	0.002435			0.8824	0.8786		
M	Box Cox Transformation for Model G with under.deaths (Lambda = 1.70707)		0.1425	0.08614	Int. term witm main	12 points high leverage	0.8988	0.8946	6	1551.9
N	Box Cox Transformation for Model G with infant.deaths (Lambda = 1.70707)		0.1517	0.08454	Int. term witm main		0.8984	0.8942		
P	Box Cox Transformation for Model G with under.deaths and Hepatitis.B (Lambda = 1.66667)		0.5664	0.05693	Int. term witm main	11 points high leverage	0.9032	0.8984	7	1325
Q	Box Cox Transformation for Model G with Hepatitis.B (Lambda = 1.70707)		0.6263	0.02948	Int. term witm main	11 points high leverage	0.9024	0.8984	6	1367
R (alternative)	Box Cox Transformation for Daniel's variables (Lambda = 1.78787)		0.7269	0.1213	9.08	12 points high leverage	0.9011	0.897	6	1456

Table 2: Summary of comparison between all the models considered in analysis

7. Discussions

From the first meeting in our team, we decided to use alternative technique to push forward our project topic, that is, we develop two version models back-to-back, in this way, we could find the errors and select from the best model.

After several analysis, the final models that met all the required assumptions for linear regression that our group found were:

The model mentioned above modelP is

$$\frac{(Life.expectancy^{1.66667}-1)}{1.66667} = 390.1042+568.279*Income.composition.of.resources-0.4378*Adult.Mortality-33.6332*HIV.AIDS+0.4581*Hepatitis.B-0.0294*under.five.deaths+0.0775*Adult.Mortality*HIV.AIDS$$

The other modelR is

$$\frac{(\widehat{Lifeexpectancy})^{1.767677}-1}{1.767677} = 416.786+873.469*Incomecompositionofresources-0.067*AdultMortality-14.768*HIV/AIDS+1.842*HepatitisB-0.006*AdultMortality*HepatitisB$$

At the end ,we compared these two models below, ModelP presented a better adjusted R^2, but slightly compared with Model R. For the criteria of minimum Marlow's Cp, the ModelR was the best .

	adj r square	Cp	AIC
ModelP	0.8984	7	1325
ModelR	0.897	6	1456

Table 3: Summary of comparison of statistical parameters for modelP and modelR

After analysis, we found out thatthe model selection methods and their sequence of usage are very important in influencing the final result.

At the start of building ModelP , we mainly used t-test to select the best predictors. Considering this criteria, we determined that the significant variables (at 5 % level) that influences Life Expectancy are: Income.composition.of.resources, Adult.Mortality, HIV.AIDS, and one of the variables infant.deaths or under.five.deaths (Multicolinearity indicated that they are redundant), as well as the interaction term Adult.Mortality * HIV.AIDS.

For building ModelR , we used stepwise regression, backward regression and forward regression mainly since the beginning (foundation), the best predictors obtained were: Income.composition.of.resources , Adult.Mortality, HIV.AIDS, Hepatitis B and the interaction term Adult.Mortality * Hepatites.B These model does not contains infant.deaths or under.five.deaths, but contains Hepatitis B. The use ofHepatitis B in the t-test resulted in ap-value of 0.0577, little higher than $\alpha = 0.05$. This implied that we should drop it, but using stepwise regression analysis, it was suggested to keep this variable

As for infant.deaths, under.five.deaths, on the contrary, regression functions did not recommend them, but t-test showed a p-values that are significant, 0.0469 and 0.0445 respectively (less but close to 0.05).

So, the challenges for us in our project are, though we have learned how to deal with data and how to create model, but what are the best methods, criteria and the using sequences in building and select a model, especially when facing critical value, and conflicting results.

We think, there are several ways we could improve our model definitely. As mentioned above, if we have much more time and knowledge, we could test more solutions to polish out a suitable model.

In terms of coefficients of the selected variables if the best models (P and R), they indicated that IncomeComposition.of.Resources and Hepatites.B immunization influences positively in the Life Expectancy, whereas Adults.Mortality, HIV.AIDS and under.five .deaths impacts it negatively.

Regarding linear regression assumptions, the independence of variables was respected because we just analyzed one year (2015) data in the dataset. In terms of linearity and normality, the majority of models indicated that they were attained, but the homoscedasticity (equal variance of the residuals) was only obtained after applying Box-Cox transformation.

8. Recommendations

For future studies, it is recommended to redo the work considering other years (separately) in the dataset just to verify if the significant variable are the same from year to year, or to carry out

some statistical analysis with the full data (all the years) considering some regression using some time-series approach.

9. Appendix (Files)

The following are links to the R codes and the dataset in our analysis

1. DATA603_Team5_WHO-ModelP-RCode
2. DATA603_Team5_WHO - ModelR - RCode
3. Project Dataset

10. References

1. "Healthy Life Expectancy at 60 (years)"World Health Organization.
2. "Health Status Statistics: Mortality". World Health Organization.
3. "Life expectancy and Healthy life expectancy, data by country" . World Health Organization. 2020.
4. "Life expectancy and Healthy life expectancy, data by WHO region". World Health Organization. 2020.
5. "Human Development Report 2020 (statistical tables 1 and 4)" United Nations Development Programme.
6. "World Population Prospects 2015 Revision" United Nations, Department of Economic and Scocial Affairs. 2015.
7. "Health status - Life expectancy at birth - OECD Data". theOECD.
8. "Data for Canada is for 2012".
9. "Life Expectancy Dataset used in the analysis"
10. "World life expectancy information"