Data Analysis Project
MDA9159A

Runze Yi 251595005
Xianglin Jin 251028972

1. Introduction:

The word life expectancy intuitively refers to how long a person can live. It is a measurement strictly bound with specific circumstances such that a minor change in the features used to predict it may result in significantly different outcomes. Although there are plenty of previous studies on life expectance exploring factors like demographic variables, income composition, and mortality rates, the effect of immunization and human development are not fully taken into consideration. In this project, we aim to study and predict the life expectancy of people using one of the datasets, Life Expectancy (WHO) [1], that has been made public by the World Health Organization (WHO) under its Global Health Observatory (GHO) data repository. The dataset includes immunization, mortality, and economic and social factors spanning 193 countries from the years 2000 to 2015.

By the completion of this project, we aim to build a linear regression model that achieves a balance between prediction accuracy and the ability to generalize toward other data and not overfit. We will achieve that by answering the following questions:

- Are all the initially chosen features significant in terms of predicting life expectancy?
- Are there any correlated predictors?
- Which predictors/data points should be removed to build a better model?
- What will the model assumption be before/after the process?
- What is the predictive power of the final model, does it fail to generalize or overfit?

Throughout our project and research, previous works including *Statistical Modeling of Life Expectancy data (R) by MOHAMED EL-SAADANY* [2], *Life Expectancy Estimation based on Machine Learning and Structured Predictors by Khulood Faisal, Dareen Alomari, Hind Alasmari, Hanan Alghamdi and Kawther Saeedi* [3].

2. Data Description

The “Life Expectancy (WHO)” data contains 22 variables and 2938 observations, with Life expectancy as the response variable and the rest as predictors. The following are the descriptions of the features contained in the dataset.

- Country, name of countries
- Year, the time when the data collects, ranges from 2000 to 2015
- Status, whether the country is developing or developed
- Life expectancy, the average age of death, our **target response**
- Adult Mortality, probability of adults(age 15-60) deaths per 1000 population
- infant deaths, probability of infant deaths per 1000 population
- Alcohol, average liters of alcohol consumed
- Percentage expenditure, percentage of GDP that is used for expenditure on health
- Hepatitis B, percentage of HepB immunization coverage
- Measles, the number of cases of Measles each year per 1000 population
- BMI, the average body mass index of the entire population
- Under five deaths, the number of under-five deaths per 1000 population
- Polio, the percentage of Pol3 immunization coverage among 1-year-olds
- Total expenditure, the percentage of government expenditure on health
- Diphtheria, diphtheria, and pertussis immunization coverage among 1-year-olds
- HIV/AIDS, deaths per 1000 live births HIV/AIDS among 0~4-year-olds.
- GDP, the Gross Domestic Product per capita in USD

- Population, the population of the country
- Thinness 1-19 years, the percentage of prevalence of thinness among children and adolescents for age 10 to 19
- Thinness 5-9 years, the percentage of prevalence of thinness among children and adolescents for age 5 to 9
- Income composition of resources, the Human Development Index in terms of income composition of resources
- Schooling, the number of years of schooling

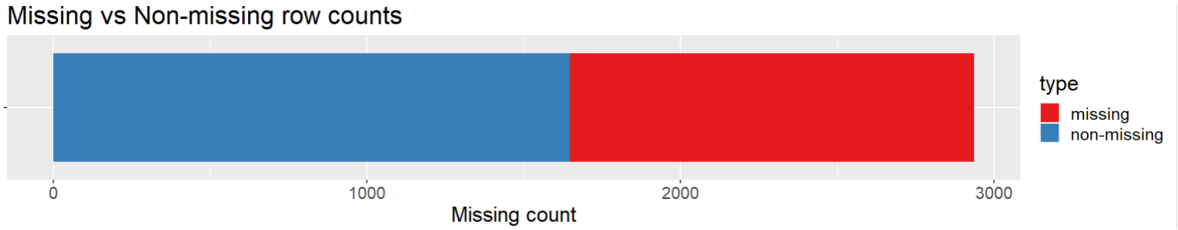
Among these features, Country, and Status are categorical data, but the feature Year will be considered as categorical as well, as the difference between each year will not necessarily be identical. The data will be better explained by the model with Year as a categorical variable.

3. Data Preparation and Statistics

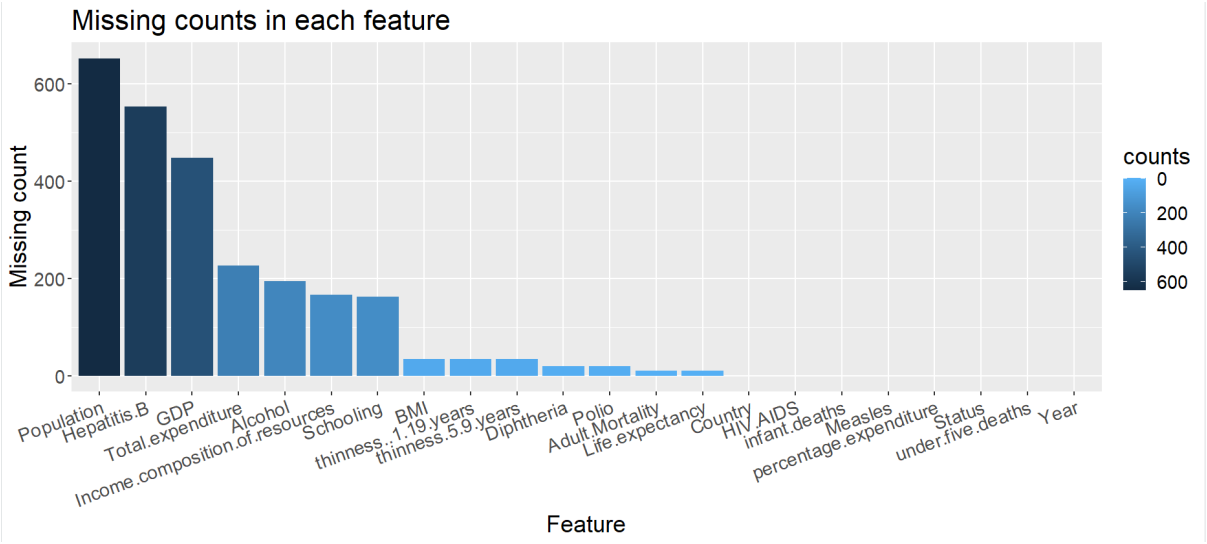
The data needs to be thoroughly understood in order to build a good predictive model.

1) Handling Null Values

There are 1289 rows that contain at least 1 missing value, which occupies 43.87% of the entire dataset.



The distribution of these data entries in all features is:



The majority of missing values are in features Population, Hepatitis.B, and GDP, which certain overlaps.

As nearly half of the observations contain at least one missing value, while lots of them contain more than one missing value, simply replacing the missing values with the median or mean of features will introduce a loss of information rather than a gain.

2) Summary Statistic

Summary statics before removing null values(Country not included):

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
Year	2938						
... 2000	183	6.2%					
... 2001	183	6.2%					
... 2002	183	6.2%					
... 2003	183	6.2%					
... 2004	183	6.2%					
... 2005	183	6.2%					
... 2006	183	6.2%					
... 2007	183	6.2%					
... 2008	183	6.2%					
... 2009	183	6.2%					
... 2010	183	6.2%					
... 2011	183	6.2%					
... 2012	183	6.2%					
... 2013	193	6.6%					
... 2014	183	6.2%					
... 2015	183	6.2%					
Status	2938						
... Developed	512	17.4%					
... Developing	2426	82.6%					
Life.expectancy	2928	69.225	9.524	36.3	63.1	75.7	89
Adult.Mortality	2928	164.796	124.292	1	74	228	723
infant.deaths	2938	30.304	117.927	0	0	22	1800
Alcohol	2744	4.603	4.052	0.01	0.878	7.703	17.87
percentage.expenditure	2938	738.251	1987.915	0	4.685	441.534	19479.912
Hepatitis.B	2385	80.94	25.07	1	77	97	99
Measles	2938	2419.592	11467.272	0	0	360.25	212183
BMI	2904	38.321	20.044	1	19.3	56.2	87.3
under.five.deaths	2938	42.036	160.446	0	0	28	2500
Polio	2919	82.55	23.428	3	78	97	99
Total.expenditure	2712	5.938	2.498	0.37	4.26	7.492	17.6
Diphtheria	2919	82.324	23.717	2	78	97	99
HIV.AIDS	2938	1.742	5.078	0.1	0.1	0.8	50.6
GDP	2490	7483.158	14270.169	1.681	463.936	5910.806	119172.742
Population	2286	12753375.12	61012096.508	34	195793.25	7420359	1293859294
thinness..1.19.years	2904	4.84	4.42	0.1	1.6	7.2	27.7
thinness.5.9.years	2904	4.87	4.509	0.1	1.5	7.2	28.6
Income.composition.of.resources	2771	0.628	0.211	0	0.493	0.779	0.948
Schooling	2775	11.993	3.359	0	10.1	14.3	20.7

As can be observed, Year, as transformed into a categorical variable, is very evenly spread, with the exception of the year 2013 occupying 0.4% than other years.

Status as a categorical variable has only 2 ver imbalance categories, which appeals to be not valuable to the model.

The features infant.deaths, percentage.expenditure, Measles, under.five.deaths, HIV.AIDS, GDP, and Population all have a higher standard deviation compared to their mean. Population and Measles have a standard deviation almost 5 times as large as the mean.

The number of observations in each feature does not align, indicating some of them have missing values.

Summary statics after removing null values(Country not included):

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
Year	1649						
... 2000	61	3.7%					
... 2001	66	4%					
... 2002	81	4.9%					
... 2003	95	5.8%					
... 2004	103	6.2%					
... 2005	110	6.7%					
... 2006	114	6.9%					
... 2007	120	7.3%					
... 2008	123	7.5%					
... 2009	126	7.6%					
... 2010	128	7.8%					
... 2011	130	7.9%					
... 2012	129	7.8%					
... 2013	130	7.9%					
... 2014	131	7.9%					
... 2015	2	0.1%					
Status	1649						
... Developed	242	14.7%					
... Developing	1407	85.3%					
Life.expectancy	1649	69.302	8.797	44	64.4	75	89
Adult.Mortality	1649	168.215	125.31	1	77	227	723
infant.deaths	1649	32.553	120.847	0	1	22	1600
Alcohol	1649	4.533	4.029	0.01	0.81	7.34	17.87
percentage.expenditure	1649	698.974	1759.229	0	37.439	509.39	18961.349
Hepatitis.B	1649	79.218	25.605	2	74	96	99
Measles	1649	2224.494	10085.802	0	0	373	131441
BMI	1649	38.129	19.754	2	19.5	55.8	77.1
under.five.deaths	1649	44.22	162.898	0	1	29	2100
Polio	1649	83.565	22.451	3	81	97	99
Total.expenditure	1649	5.956	2.299	0.74	4.41	7.47	14.39
Diphtheria	1649	84.155	21.579	2	82	97	99
HIV.AIDS	1649	1.984	6.032	0.1	0.1	0.7	50.6
GDP	1649	5566.032	11475.9	1.681	462.15	4718.513	119172.742
Population	1649	14653625.889	70460393.403	34	191897	7658972	1293859294
thinness..1.19.years	1649	4.851	4.599	0.1	1.6	7.1	27.2
thinness.5.9.years	1649	4.908	4.654	0.1	1.7	7.1	28.2
Income.composition.of.resources	1649	0.632	0.183	0	0.509	0.751	0.936
Schooling	1649	12.12	2.795	4.2	10.3	14	20.7

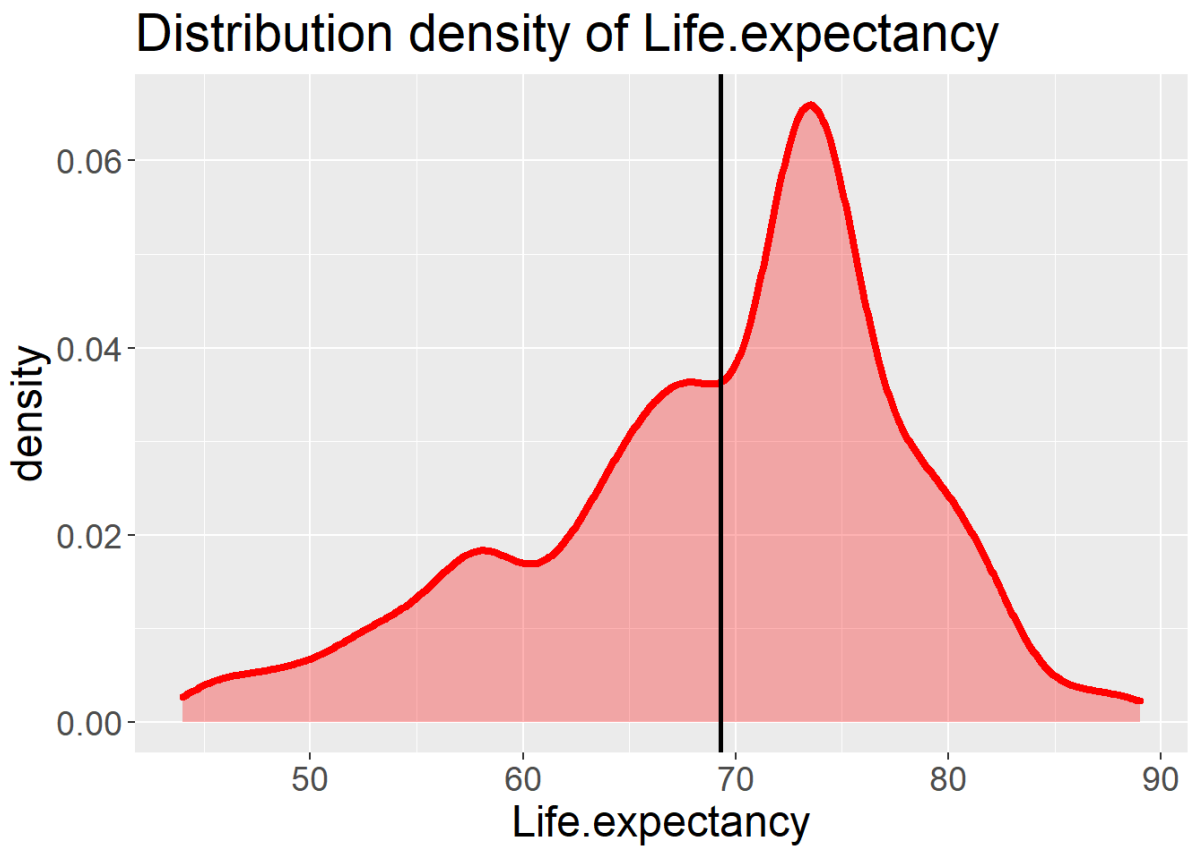
After removing all rows with null values, the years are not as balanced as before, with the lowest of the year 2015, occupying only 0.1% of the data.

Status becomes slightly less imbalanced than it once was but still appeals to be invaluable to the model as it has only 2 categories and fails to bring indicative information to the model.

The features infant.deaths, percentage.expenditure, Measles, under.five.deaths, HIV.AIDS, GDP, and Population are still having a higher standard deviation compared to their mean. The Normality assumption is very likely to fail.

The number of observations is now aligned to 1649, all entries now have no missing values in them.

The distribution of the response variable Life Expectancy is as follows:



Life expectancy has a mean of 69.302 and a standard deviation of 8.797.
This distribution clearly skews towards the right, meaning it is not in a shape of a normal distribution. This further supports that the Normality assumption is likely to be violated.

Also, a scatter plot matrix is generated to check on the relationships between all variables including the response variable. It is sad to see that there is no clear linear relationship between the response variable and any predictor.

4. Analysis

Linear Regression[4], is a linear approach for modeling the relationship between a scalar response variable and one or more explanatory variables. The case of one explanatory variable is called *simple linear regression*; for more than one, the process is called **multiple linear regression**[5].

It assumes that there exists a function y = f(x) + **ε** where **ε** is the random error.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_{p-1} x_{p-1} + \epsilon$$

Here p-1 is the number of features to be used in prediction, x are data values of each feature, and y is the response variable life expectancy.

In building a multiple linear regression model, we try to explain the linear relationship between the features and the response variable by finding estimated values of the **β**, such that

$$\hat{\beta} = (X^T X)^{-1} XY$$

Where beta_hat is the vector of estimated **β** coefficients, X is the data matrix, and Y is a vector of the response variable.

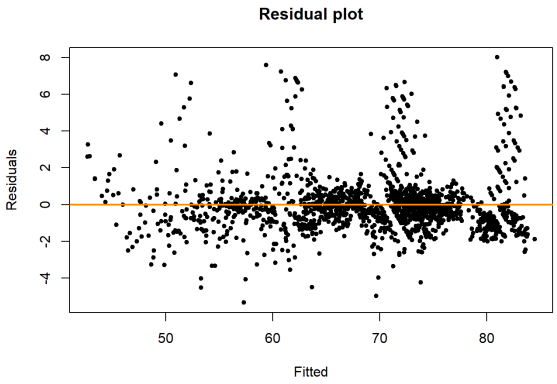
The estimated life expectancy or the prediction of life expectancy using the data we have is obtained as

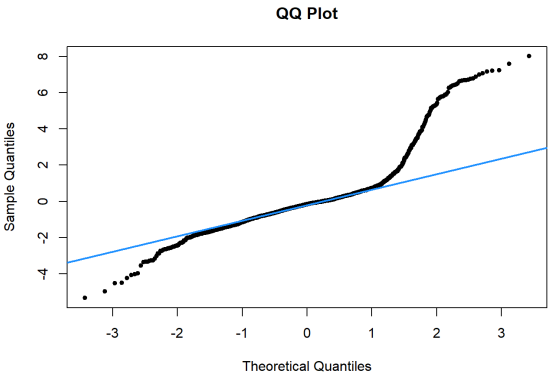
$$\hat{Y} = X\hat{\beta}$$

To better understand the model we are building and the data we use, we will analyze the behavior of our model in the following order:

- Check model assumption with prepared data, taking into consideration influential points and box-cox transformation
- Check model assumption with prepared data but removing Country, taking into consideration influential points and box-cox transformation. The column country has 193 categories within 2938 observations, and 133 categories with 1649 observations after removing the missing values. It is crucial that we study the effect of this predictor.
- Variable selection, including the **Variance Inflation Factor** (VIF), **Akaike Information Criterion** (AIC), and the **Bayesian Information Criterion** (BIC). ANOVA tables are also used to check the significance of the categorical predictors individually.
- Model comparison, compare obtained models through **Predicted Residual Error Sum of Squares** (PRESS) and **Rooted Mean Squared Error** (RMSE) to find the single desired model. These metrics apply **Cross Validation** to the models to prevent overfitting.

- **Model Assumptions(before)**

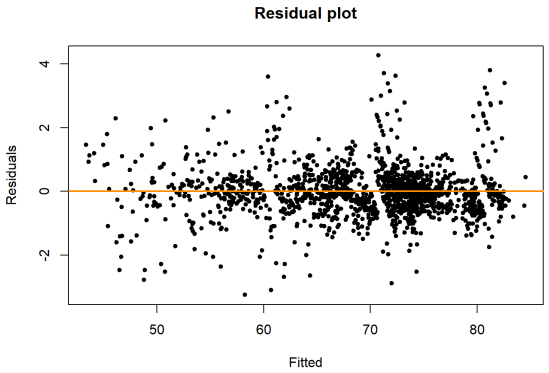


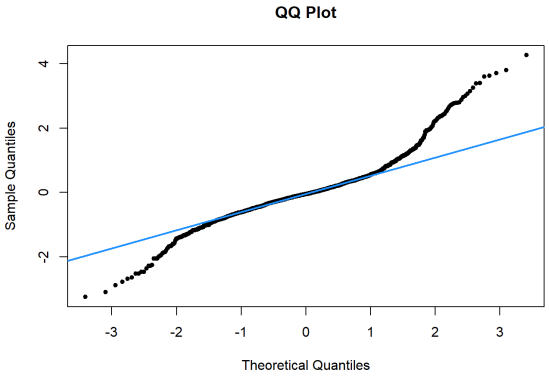


These are residual and QQ plots on data without missing values. In the residual plot, we may see that the Linearity assumption might hold as E(e) is around 0 for all regions, but it is clearly observed that the Equal Variance assumption and Normality assumption are violated.

BP test returns p = 2.889e-08, Shapiro test returns p = 2.2e-16

Using the cook's distance, we obtained 116 data points that are identified to be influential with a criterion of 4/n. After removing these points:

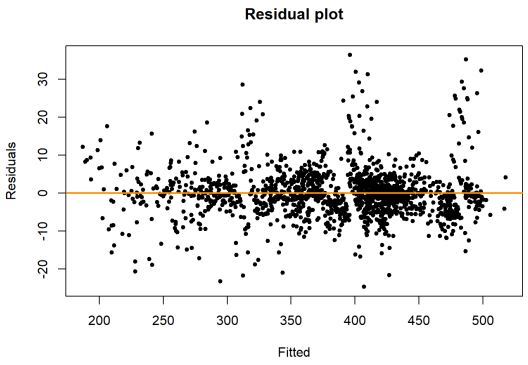


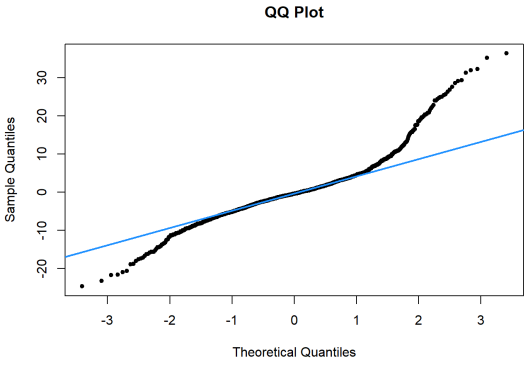


Similar results are suggested by the plots, where the Linearity assumption might hold and others violated., but the QQ plot looks a little bit better.

BP test returns p = 2.2e-16, Shapiro test returns p = 2.2e-16, even worse than before

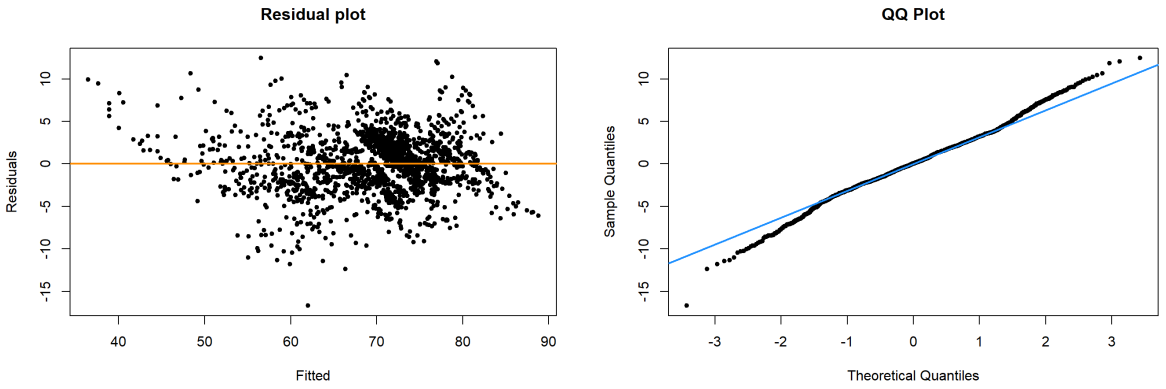
Using Box-cox transformation, we obtained λ = 1.5, and after the transformation:





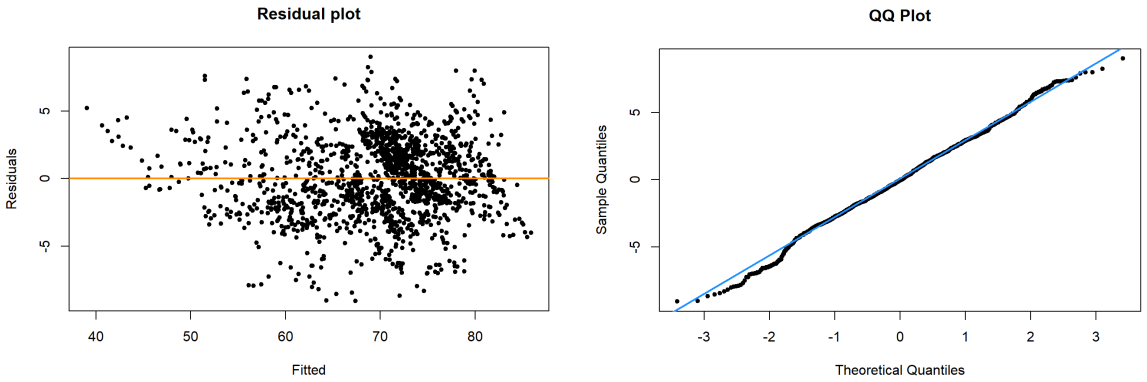
It still suggests similar results.
BP test returns p = 2.534e-11, Shapiro test returns p = 2.2e-16

- **Model Assumptions(removing country)**
  If we take Country out of consideration and repeat the above process:



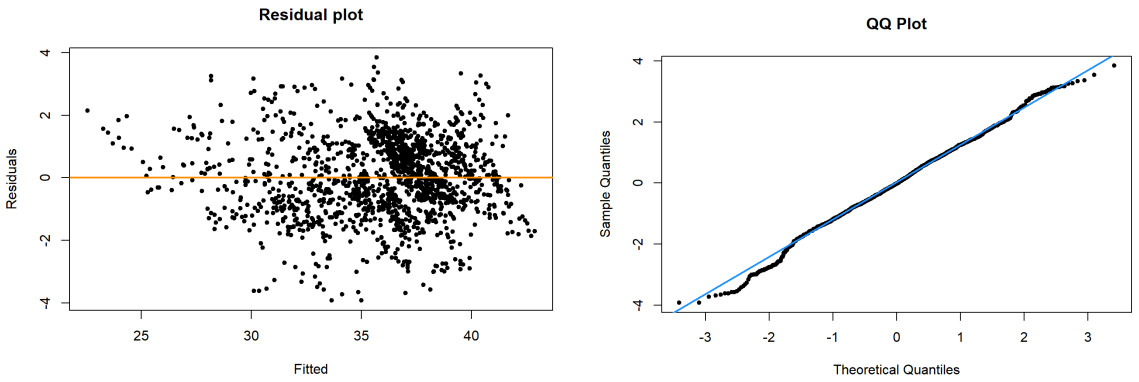
Both Linearity and Equal Variance assumptions are violated. The normality assumption seems also violated but much better than before.
BP test returns p = 2.2e-16, Shapiro test returns p = 7.87e-08

Removing 103 influential points using the cook's distance:



The residual plot looks almost the same, and the QQ plot looks better.
BP test returns p = 2.2e-16, Shapiro test returns p = 0.003814

Using Box-cox transformation, we obtained $\lambda = 0.8$, and after the transformation:



It still suggests similar results. This is as expected since the box-cos transformation returns a lambda value so close to 1.

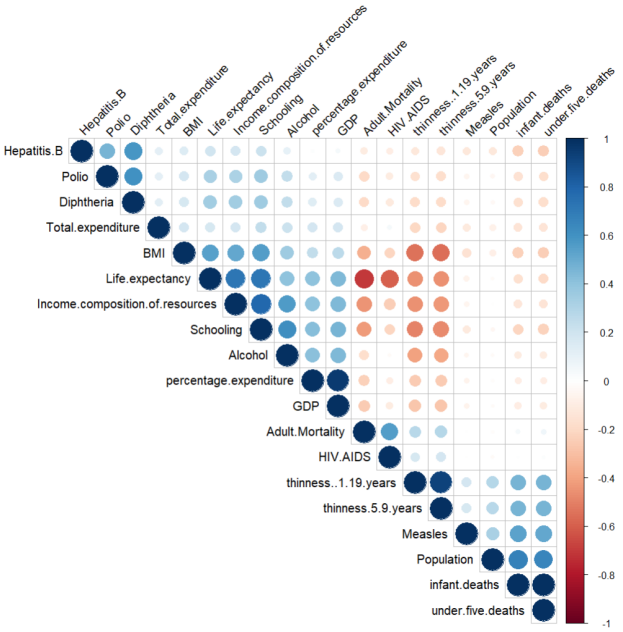
BP test returns p = 2.2e-16, Shapiro test returns p = 0.001902

Removing the predictor Country increases the model's performance on the Normality assumption while still violating it, but makes the model violates the Linearity assumption as well. More tests need to be done to determine whether to keep this predictor.

- **Variable Selection**

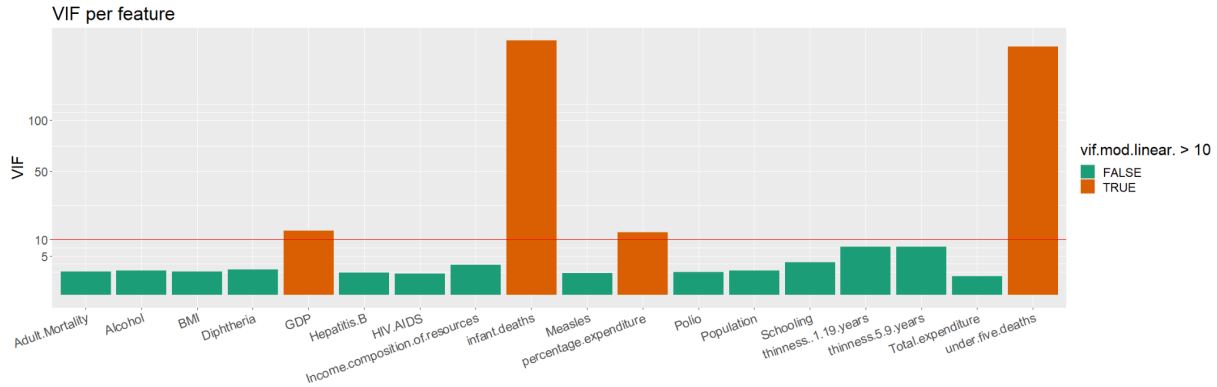
The first step in our variable selection is to check whether the categorical variables, Country, year, and status are significant by comparing the reduced model to the full model with ANOVA tables. As result, the variable Status is irrelevant but Country and Year are significant to the model.

The variable Country contains the names of 193 countries worldwide, which will generate 193 extra variables when we try to fit a model. These variables are non-identifiable parameters for the variance inflation factor test, and also cause the model to be extremely over-fitting when we conduct a PRESS test on the model. Therefore, we decide to remove the variable Country to avoid further issues.



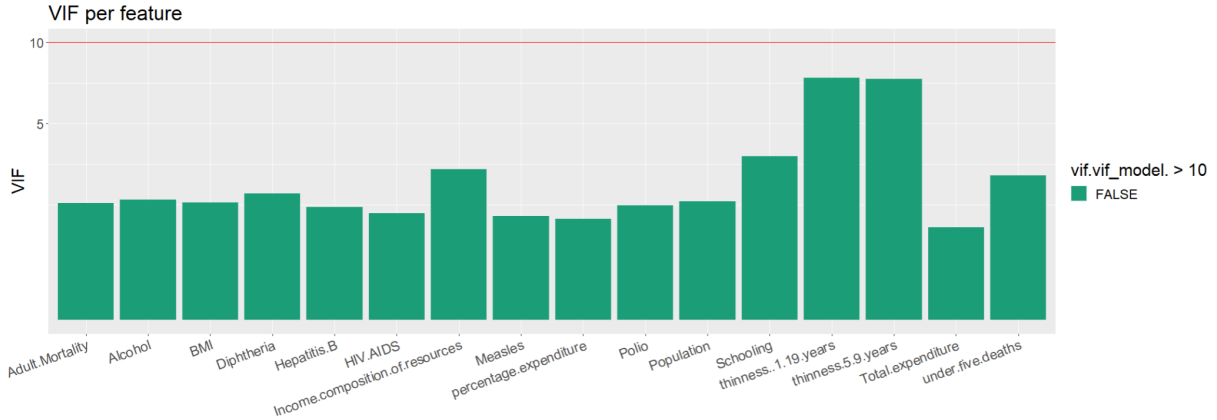
During the data analysis, we figured that some of the variable predictors are highly correlated. When colinearity occurs, it directly affects the standard errors and undermines independent variables' statistical significance. From the graph below, we can notice a very strong correlation between [infant death] and [under five death], [Percentage expenditure] and [GDP], [Thinness 1-19 years] and [Thinness 5-9 years].

Therefore, for a better variable selection, we continuously run through the variance inflation factor test and drop the variable that has the highest vif value each time. In this stage, after we dropped infant deaths and GDP, all the variables have a vif less than 10.



This graph above illustrate the vif value for each variable initially, and we can see that [infant death] has the largest vif between [infant death] and [under five death], and [GDP] has the largest vif between [Percentage expenditure] and [GDP].

The graph below shows the vif after we dropped [infant death] and [GDP] from the dataset. As the result, all the variables are now having vifs less than 10.



Furthermore, on variable selection, we chose to collect all the possible models with backward selection, forward selection, and stepwise selection with both AIC and BIC matrics based on the model after the vif process. As the result, both forward selections with either AIC or BIC did not reduce any variables, therefore the backward selection and the step-wise selection share the same result.

After the stepwise selection, we now have three models, the model after the vif process(model_vif), the model after AIC stepwise selection(step_AIC), and the model after BIC stepwise selection(step_BIC). By obtaining the test statistic about the step_AIC and step_BIC against model_vif, we can see that the variable reduced by step_AIC are insignificant, but some parameters removed are significant in the step_BIC model. To find a better model among these three, first, we conduct a PRESS test and a K-fold cross-validation test. We choose five as our K due to the large dataset, and as the result, model_vif has better results in the RMSE and PRESS tests.

Therefore, even though step_AIC only reduced variables that are not significant, model_vif has a better accuracy overall. Hence, throughout the entire variable selection process, the variables that we dropped are Country, status, infant deaths, and GDP.

5. **Results**

Results of the 5 fold cross validation are:

Model	mean(RMSE) by Cross-Validation	Number of Predictors
step_AIC	4.624697	7
step_BIC	4.82167	3
model_VIF	3.957291	16

These results are from models excluded of any categorical variables as they would prevent models from performing Cross Validation, as some categorical data can't appear in all folds.

The model_VIF achieved a much better RMSE than the other two models while having much more predictors.

The step_BIC achieved a slightly larger RMSE than step_AIC but has only 3 predictors compared to 7 of step_AIC.

Even though a model with more predictors is always expected to have better accuracy, models still risk of getting overfitting by having too many and failing to generalize outside of the data it uses to train on.

The PRESS score for the models are:

Model	PRESS	Number of Predictors
step_AIC	18.02758	7
step_BIC	20.43165	3
model_VIF	13.74599	16
step_AIC_year	17.75981	8
step_BIC_year	20.48571	4
model_VIF_year	13.6213	17
step_AIc_country	Inf	8

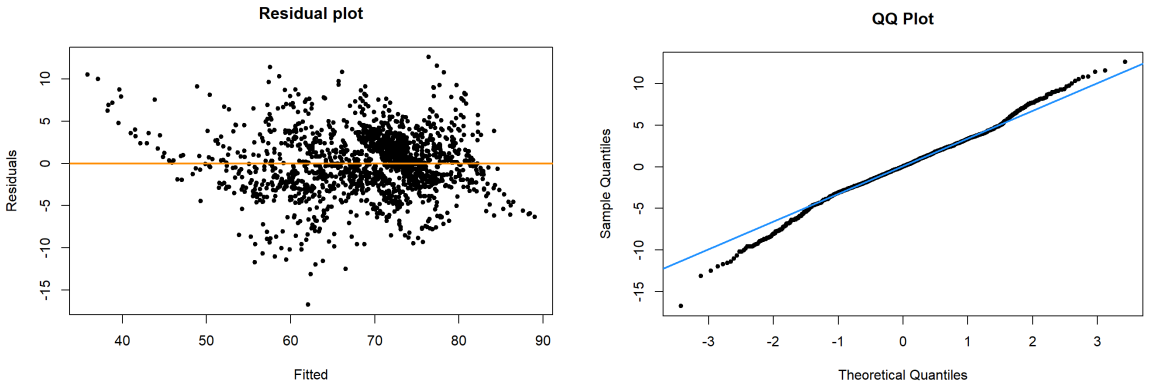
step_BIC_country	Inf	4
model_VIF_country	Inf	17
step_Aic_both	Inf	9
step_BIC_both	Inf	5
model_VIF_both	Inf	18

The model step_AIC_year and step_AIC_country indicates the model of adding year or country back.

The PRESS[6] measures the fit of a model to a sample of observations that were not themselves used to estimate the model, with the lowest values of PRESS indicating the best structures. This further supports our decision to choose model_VIF among the other 2 models and adds to it by including the predictor Year again.

The models with predictor Country all produce Inf as PRESS score, indicating that they have overfitted to the training data and fail to generalize towards other data.

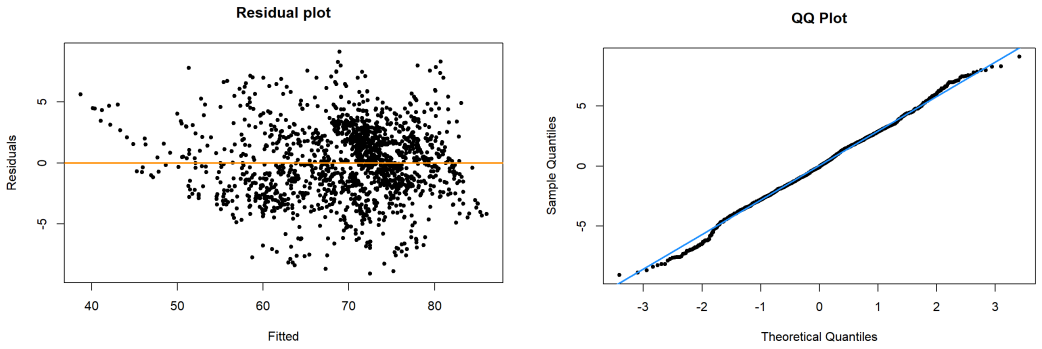
- **Model Assumptions(desired model)**



Both Linearity and Equal Variance assumptions are violated. The normality assumption seems also violated but much better than before.

BP test returns p = 2.2e-16, Shapiro test returns p = 2.102e-08

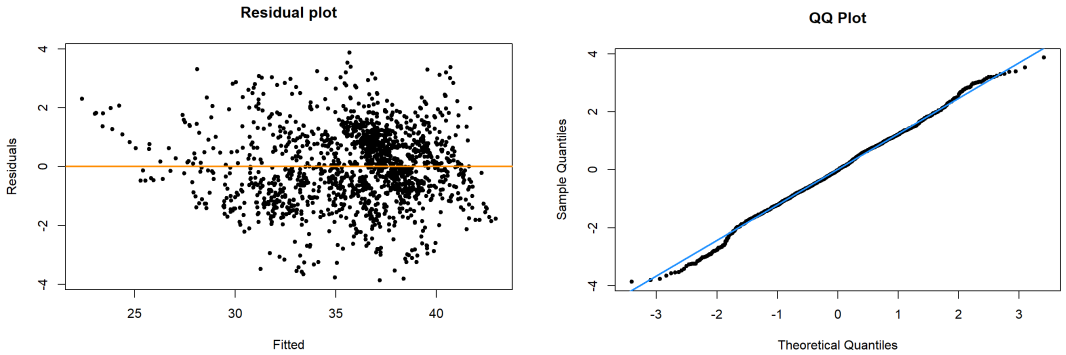
Using the cook's distance to remove 116 influential points:



The residual plot looks quite the same, but the QQ plot looks even better.
BP test returns p = 2.2e-16, Shapiro test returns p = 0.01018

Even though the Linearity and Equal Variance assumptions are still not good, the Normality assumption now almost holds.

Using Box-cox transformation, we obtained λ = 0.8, and after the transformation:



These plots suggest similar results. This is as expected since the box-cos transformation returns a lambda value so close to 1.

BP test returns p = 2.2e-16, Shapiro test returns p = 0.009024

After performing these tests, we can see the removal of influential points has the best attempt at fixing the model assumptions. The model trained like this has

RMSE(on self)	R2	Adjusted R2
2.920846	0.8733	0.8707

If we split the current 1649-116=1533 observations into 70-30 training and testing sets, the performance of the chosen model on the testing set will be

RMSE(on training)	RMSE(on testing)	R2	Adjusted R2
2.442954	1.62088	0.875	0.8713

6. Discussion

By doing this project, we explored the prediction of life expectancy with factors like immunization and human development. This also gives us experience in handling multi-featured, non-prepared data and successfully achieving the balance we desired between prediction accuracy and generalization toward unseen data. A low RMSE of approximately 1.62 is achieved on the testing set, by the final selected model.

Starting with the original data, we performed data exploration and preparation to have a better understanding of the dataset as well as the problem. We ended up eliminating 1289 entries that contain missing values instead of replacing them with median or mean. We then moved on to model assumptions with or without the predictor Country. Even though we had failed to make the model assumptions hold, our effort had been proven to be effective in correcting the Normality assumption to a p-value of 0.01018, which is significantly better than the unprepared model and data.

During variable selection, we explored the use of the Variance Inflation Factor(VIF), Akaike Information Criterion (AIC), and the Bayesian Information Criterion (BIC) as well as concepts of forward, backward and stepwise selection to find the most crucial predictor excluding any significant correlation, while penalizing the model with complexity to achieve the desired balance. After excluding predictors of Infant death and GDP using VIF, two more different models were selected with 9 and 5 predictors each. Cross-validation-based Predicted Residual Error Sum of Squares (PRESS) and Rooted Mean Squared Error (RMSE) are also used in terms of model selection to find the most desirable one from the obtained models and to reach a final result. Both the PRESS and RMSE of the selected model are the lowest amongst others in the comparison, indicating our success in arriving at the desired balance.

A train-test set splitting is conducted in the end, obtaining the testing RMSE of 1.62088. This demonstrates the model is not only accurate but also with the ability to generalize and make useful predictions on unseen data.

7. Limitations and future work

The dataset used in this research is incomplete as it has many missing values. After excluding the missing values, there are only 1649 observations left to train the model with. Besides that, all model assumptions for trained models are violated. We believe that a better performance would be achieved if a proper transformation strategy is used on the response variable.

In the three attempts of Box-cox transformation, two of them obtained lambda close to 1. Such lambda values lead to inefficient and invaluable transformations which resulted in even slightly worse cases in the model assumptions. Also since there is hardly any distinguishable linear relationship between the response variable and the predictors, we failed in adding any transformations of X values.

Another piece that we failed to accomplish in this research is interactions between predictors. Since we have more than 20 predictors, it is hard to do an interaction analysis and find out which among them are useful.

8. **Inference:**

[1] Life Expectancy (WHO) dataset from the World Health Organization (WHO) under its Global Health Observatory (GHO) data repository
Kaggle https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who
[2] Statistical Modeling of Life Expectancy data (R) by MOHAMED EL-SAADANY
Kaggle
https://www.kaggle.com/code/mohamedelsaadany/statistical-modeling-of-life-expectancy-data-r
[3]Life Expectancy Estimation based on Machine Learning and Structured Predictors by Khulood Faisal, Dareen Alomari, Hind Alasmari, Hanan Alghamdi and Kawther Saeedi
https://dl.acm.org/doi/fullHtml/10.1145/3503047.3503122
[4] Linear regression - Wikipedia, https://en.wikipedia.org/wiki/Linear_regression
[5] David A. Freedman (2009). *Statistical Models: Theory and Practice*. Cambridge University Press. p. 26. A simple regression equation has on the right-hand side an intercept and an explanatory variable with a slope coefficient. A multiple regression e right-hand sides, each with its own slope coefficient
[6] PRESS - Wikipedia, https://en.wikipedia.org/wiki/PRESS_statistic