STATISTICAL ANALYSIS ON FACTORS INFLUENCINGLIFE EXPECTANCY

Adeel Qureshi
Cleveland Johnson
Nathan Deinlein
Puri Rudick

GitHub - https://github.com/nedeinlein/Project_1_Applied_Stats.git
Group 4 | MSDS 6372 Project 1

Introduction

Life expectancy is one of the most used summary indicators of the overall health and wellbeing of a population. A nation's life expectancy reflects its social and economic conditions, as well as the quality of healthcare policies, among other factors. The two main objectives of this study are:

1) to identify key factors that associated with the population's life expectancy and
2) to build life expectancy predictive models using those pivotal factors.

Data Description

The data set that we used in this study collected from World Health Organization (WHO) and United Nations website with the help of Deeksha Russell and Duan Wang. The data set provided in Kaggle[1] titled "Life Expectancy (WHO): Statistical Analysis on factors influencing life expectancy". The data set contains 22 variables, which composes of 19 numerical variables and 3 categorical variables. 2,938 observations from 193 countries around the world. See appendix page 2 and 3 for the contents of individual variables.

Exploratory Data Analysis

Data Clean Up

The data collected between 2000 and 2015 for most of the countries, except for 10 countries that have only one year of data collection in 2013 with missing values for life expectancy. For key factors identifying and model building purposes, we decided to leave the observations from these countries out from the study, totaling in 2,928 observations left for the data set from 183 countries. We then added the column of continent based on country onto the data set.

Variables	# of Missing Value	% of Missing Value
Population	644	22.0%
Hepatitis B	553	18.9%
GDP	443	15.1%
Total Expenditure	226	7.7%
Alcohol	193	6.6%
Income Composition of Resources	160	5.5%
Schooling	160	5.5%
BMI	32	1.1%
Thinness 1-19 years	32	1.1%
Thinness 5-9 years	32	1.1%
Diphtheria	19	0.6%
Polio	19	0.6%

Table 1: Missing Values Information

Missing Values

Within 2,928 observations, 12 variables contain missing values. Numbers and percentage of missing values in each variable as shown in Table 1. We will discuss about how we handle these missing values in the later section of this report.

Life Expectancy Trend

Overall, the data shows increasing in life expectancy over years from mean value of 66.75 in 2000 to 71.62 in 2015 (see appendix page 44 and 45 for life expectancy versus years).

When we looked at life expectancy value for individual country, we found that the country in Africa has the lowest life expectancy compares to other continents, follows by Asia, Oceania, Americas, and Europe, as shown in Figure 1. On another word, we can conclude that developed countries have higher life expectancy compare to developing countries. All continents also show upward trend of life expectancy over years (see appendix page 45 and 46).

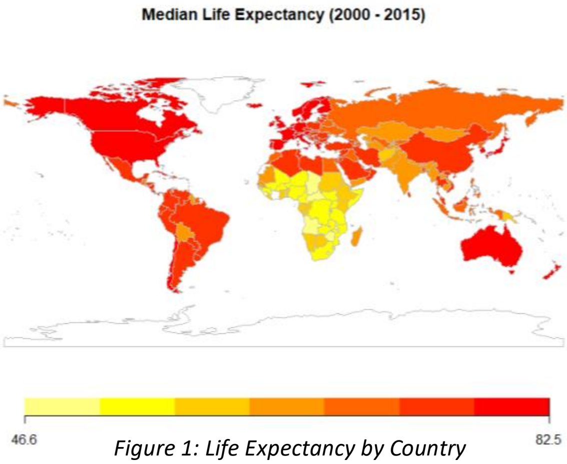


Figure 1: Life Expectancy by Country

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

Key Factors Influencing Life Expectancy

To identify key factors that associated with the population's life expectancy, we initiated the scatter plot matrix and histogram plot (see appendix) to examine the relationship each predictor has with the response (life expectancy). After that, we look in detail to pin down the variables that can be considered the cause of the outcome or related to the outcome. We then inspected to see if the response needed transformation, which it was not required. With no transformation necessitated for the response, we decided to replace missing values for each predictor with it means. Then we reviewed all possible predictors again to see if they needed to be transformed. Afterwards, we ran a single MLR fit using all predictors and then make a VIF call to obtain VIF value for each predictor.

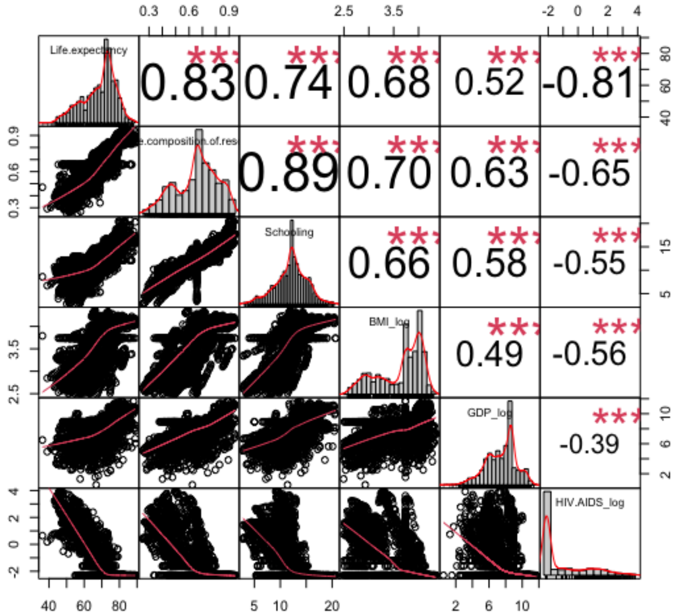


Figure 2: Scatter Plot Matrix for Selected Predictors

The variables that we identified to be key factors influencing life expectancy are Income Composition of Resources, Schooling, BMI (log transformation), GDP (log transformation), HIV/AIDS (log transformation), Country Status, and Continent. For Income Composition of Resources and Schooling variables, we replaced both null and zero value with their means because the zero values in the data set appear to not be real data collection. For BMI, mean BMI of 12 is the lower limit for human survival, so we decided to replace any BMI below 12 and null value with its mean then log transform the variable. Figure 2 shows the scatter plot matrix for selected predictors versus the response. While, Table 2 shows the statistical summary table after missing values replacement and transformations, along with VIF values of the key factors that we verified.

Variables	Mean	Std Dev	Min	Max	VIF (for Predictors)
Life Expectancy	69.22	9.52	36.30	89.00	
Numerical Variables					
Income Composition of Resources	.66	.15	.25	.95	2.767
Schooling	12.11	3.05	2.80	20.70	2.331
BMI (Log)	3.64	.48	2.49	4.35	1.621
GDP (Log)	1.69	1.81	.52	11.69	1.308
HIV, AIDS (Log)	-1.22	1.61	-2.30	3.92	1.691
Categorical Variables					
Status					1.451
- Developed	79.20		69.90	89.00	
- Developing	67.11		36.30	89.00	
Continent					1.256
- Africa	58.61		39.00	79.00	
- Americas	73.49		36.30	87.00	
- Asia	71.19		54.80	87.00	
- Europe	77.43		64.60	89.00	
- Oceania	71.21		58.90	89.00	

Table 2: Statistical Summary Table for the Response and the Key Predictors

Objective 1: Build Simple Regression Model

Goal and Approach

The goal of Objective 1 is to identify variables with key relationships to Life Expectancy and build a simple multiple linear regression model to predict Life Expectancy for a given country. The approach followed to solve this problem entailed leveraging the key variables from the Exploratory Data Analysis, leverage Forward and LASSO model selection methods, perform cross validation with an 85/15 train-test split, and finally validate the final model using the full data set and checking residuals/assumptions.

Model Selection

The model selection started with the final data set from the Exploratory Data Analysis, summarized in Table 2 above. These variables included Income Composition of Resources, Schooling, BMI (log transformation), GDP (log transformation), HIV/AIDS (log transformation), Country Status, and Continent. Two model selections were used to find the best model, Forward and LASSO.

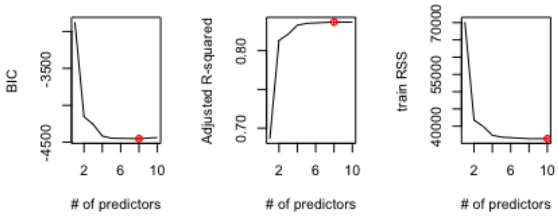


Figure 3: Forward Model Selection Plots for BIC, AdjR2 and RSS

The forward model selection process showed that the optimal model when looking at BIC and Adjusted R-squared was 8 predictions (Figure 3). The chosen predictors with coefficients can be found in Table 3. The cross validation with the test set validates the 8 predictors when viewing the Average Square Error (Figure 4).

Variables	Coefficient
Intercept	46.61
Status Developing	-3.10
Income Composition of Resources	19.53
Schooling	.22
Continent-Americas	3.43
Continent-Asia	1.13
Continent-Europe	1.01
BMI (Log)	1.51
HIV, AIDS (Log)	-2.57

Table 3: Forward Selection Predictors with Coefficients

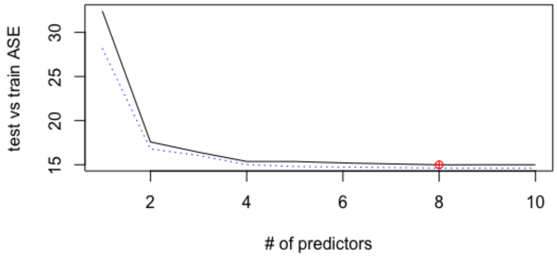


Figure 4: Forward Model Selection Cross Validation

The LASSO model selection process also landed on the same predictors as the forward selection. Figure 5 depicts those 6 predictors makes the optimal model. This output of the LASSO is slightly different than Table 3 above as it contains the Continent-Oceania variable to finish out the continent categorical variables.

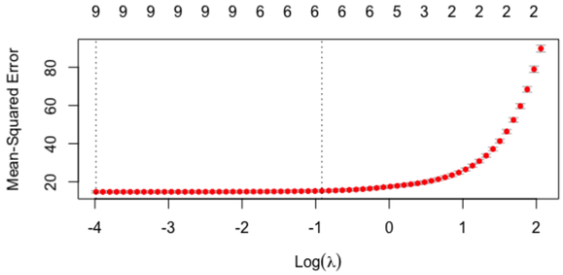


Figure 5: LASSO Optimal Model Output

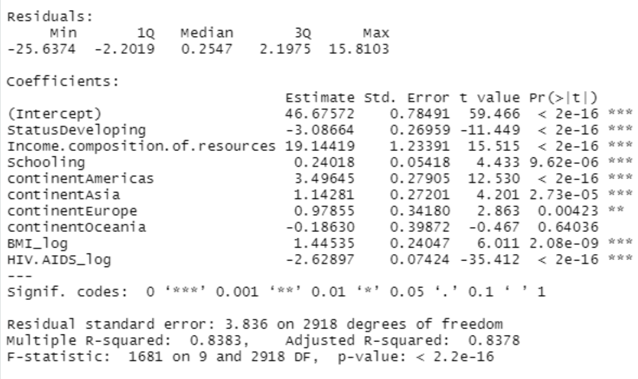
The full data set was then fit to a model containing the following predictors: Income Composition of Resources, Schooling, BMI (log transformation), HIV/AIDS (log transformation), Country Status, and Continent. Essentially only the GDP (log transformation) variable was removed.

The final model output has an adjusted R-squared of 83.78 (Figure 6). The only coefficient not statistically significant is the Continent-Oceania.

Checking Assumptions/Influential points:

The assumptions plots for the final model are met as can be seen in Figure 7 below. In reviewing influential points, observations 1126, 544 and 864 are highlighted in our residuals vs leverage plots. These observations are Haiti for 2010, Chad or 2000 and

Eritrea for 2000. In reviewing these observations, there were no anomalies within the variables themselves. Based on this review, observations were left in the data set.



```
Residuals:
     Min       1Q   Median       3Q      Max 
-25.6374  -2.2019   0.2547   2.1975  15.8103 

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)    
(Intercept)                    46.67572    0.78491  59.466  < 2e-16 ***
StatusDeveloping               -3.08664    0.26959 -11.449  < 2e-16 ***
Income.composition.of.resources 19.14419    1.23391  15.515  < 2e-16 ***
Schooling                       0.24018    0.05418   4.433 9.62e-06 ***
continentAmericas               3.49645    0.27905  12.530  < 2e-16 ***
continentAsia                   1.14281    0.27201   4.201 2.73e-05 ***
continentEurope                 0.97855    0.34180   2.863  0.00423 ** 
continentOceania               -0.18630    0.39872  -0.467  0.64036    
BMI_log                         1.44535    0.24047   6.011 2.08e-09 ***
HIV.AIDS_log                   -2.62897    0.07424 -35.412  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.836 on 2918 degrees of freedom
Multiple R-squared:  0.8383,	Adjusted R-squared:  0.8378 
F-statistic:  1681 on 9 and 2918 DF,  p-value: < 2.2e-16
```

Figure 6: Final Simple Regression Model Output

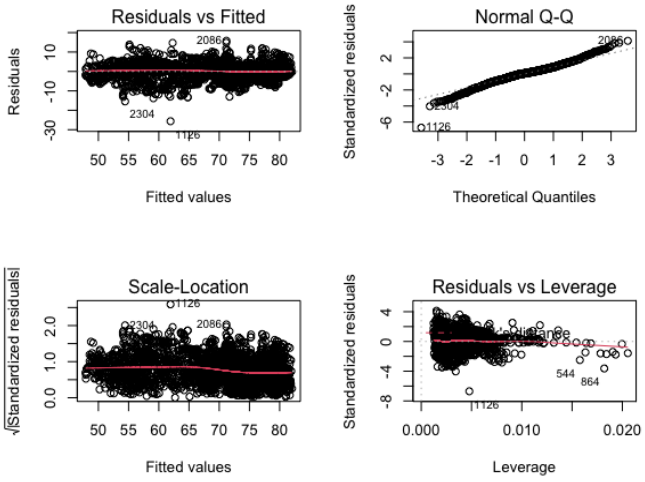


Figure 7: Final Model Residual Plots

Parameter Interpretation

The final model can be interpreted as follows:

- Intercept: The starting average life expectancy with all variables at 0 is 46.67 years.
- Status: If a country is considered Developing, the average life expectancy decreased by 3.09 years.
- Income Composition of resources: This variable is a human development index between 0 and 1. At an index of 1, the average life expectancy increases by 19.14 years. An index lower than one will only provide a fractional amount of the 19.14 years.
- Schooling: For every year of schooling the average life expectancy increases .24 years.
- Continents: Depends on the continent within which a country resides and is as follows
    - Americas – increase average life expectancy by 3.5 years
    - Asia – increases average life expectancy by 1.14 years
    - Europe – increases average life expectancy by .98 years
    - Oceania/Africa - no increase associated with these continents
- BMI: A doubling of average BMI results in a .44 increase in average life expectancy.
- HIV-AIDS: A doubling of death per 1000 live births results in a .79 decrease in average life expectancy.

The interpretation of BMI and HIV-AIDs using logs caused a review of the original data to understand the ability for the doubling of BMI or HIV-AIDSs to occur. For BMI, the range of values was from 12 to a max of 78. This only provided 3 opportunities to double, not marking a practically significant increase in average life expectancy. HIV-AIDs on the other hand spread from .1 to 50.6. This spread would provide a great number of opportunities for doubling making it a practically significant variable.

Objective 2: Building More Complex Models

Goal and Approach

The goal of Objective 2 of this exercise is to ascertain whether multiple various models could be built to predict life expectancy based on various factors. This data, collected by the WHO, gave multiple possible factors that could explain and then predict average life expectancy. In this, we looked at three separate model types to determine which might give us the best balance between bias and error, these were a simple multiple linear regression model, a more complex multiple linear model, and kNN.

Complex Multiple Linear Model

When we looked at building a more complex MLR, the first thing we wanted to add in were interaction terms. Every permutation of continuous variables was added to the initial model. This was then run with a backwards feature selection feature. The goal was to try to make this a model that would error on the side of bias a little more than the initial model, but still be usable.

Complex Multiple Linear Model with Interaction Terms

```
fwdcpmodel1 <- regsubsets(Life.expectancy ~ Status + continent + Income.composition.of.resources + Schooling
                          + BMI_log + GDP_log + HIV.AIDS_log + Income.composition.of.resources*Schooling
                          + Income.composition.of.resources*BMI_log + Income.composition.of.resources*GDP_log
                          + Schooling*BMI_log + Schooling*GDP_log + BMI_log*GDP_log,
                          data = train, method = 'backward', nvmax=13)
```

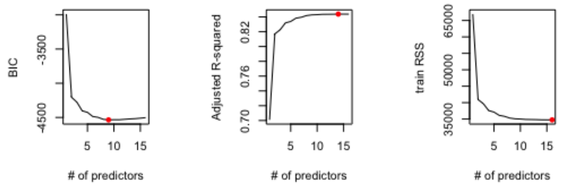


Figure 8: Complex MLR Model Plots for BIC, AdjR2 and RSS

After running the feature selection, output of the BIC showed that around 9 was probably ideal for the number of variables for the model (Figure 8).

At this point a cross validation piece was run to check the ASE for Train versus Test set. Eleven here was the minimum value. This suggests that there is a risk of overfitting if 11 variables are used instead of 9 (Figure 9).

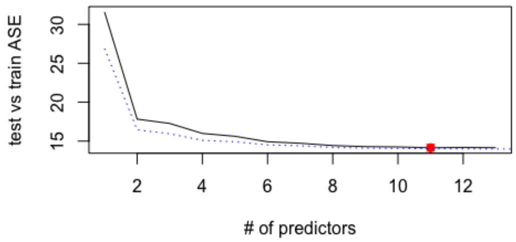


Figure 9: Complex MLR Model Cross Validation

Checking Assumptions/Influential points:

At the risk of overfitting, it was determined to start with the initial 11 variables. Using this information an intermediate model was built. This model was run to show that the model met assumptions (Figure 10) as well as running VIF (Table 4).

Variables	GVIF	DF	GVIF^1/(2*DF)
Status	2.38305	1	1.543714
Continent	7.45802	4	1.285517
Income Composition of Resources	218.626	1	14.786002
Schooling	202.708	1	14.247567
BMI (Log)	46.7168	1	6.83497
HIV/AIDS (Log)	*2.89619*	1	1.701818
Income Composition of Resources : Schooling	121.313	1	11.014196
Income Composition of Resources : BMI (Log)	528.451	1	22.988052
Schooling : BMI (Log)	478.92	1	21.884251

Table 4: Complex MLR Model VIF Summary Table

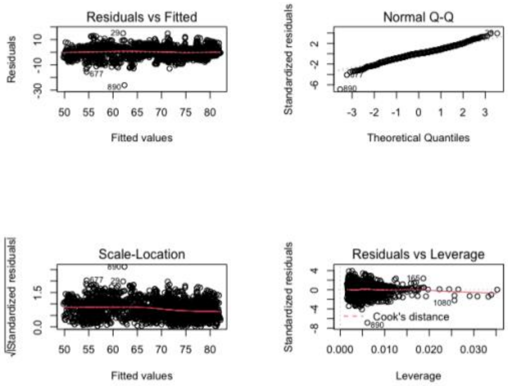


Figure 10: Complex MLR Model Residual Plots

Looking at the remaining factors, VIF was high for a number of them, for very obvious reasons as most are interaction term of each other. A final model was created to simplify the interaction terms down to eliminate Variance inflation factor. This resulted in the below model. This left interaction terms in the model while removing Inflation from the model. Overall diagnostics for the model were good and are also shown in Figure 11 as well as VIF on Table 5.

```
finalcpmodelb <- lm(Life.expectancy  ~  Status  +  continent  +  Income.composition.of.resources*HIV.AIDS_log
                    + Schooling*BMI_log, data = train)
```

Variables	GVIF	DF	GVIF^1/(2*DF)
Status	2.144139	1	1.464288
Continent	6.820763	4	1.271245
Income Composition of Resources	7.498905	1	2.738413
HIV/AIDS (Log)	2.889566	1	1.699872
Schooling	74.506796	1	8.631732
BMI (Log)	17.514491	1	4.185032
Schooling : BMI (Log)	139.552841	1	11.813249

Table 5: Final Complex MLR Model VIF Summary Table

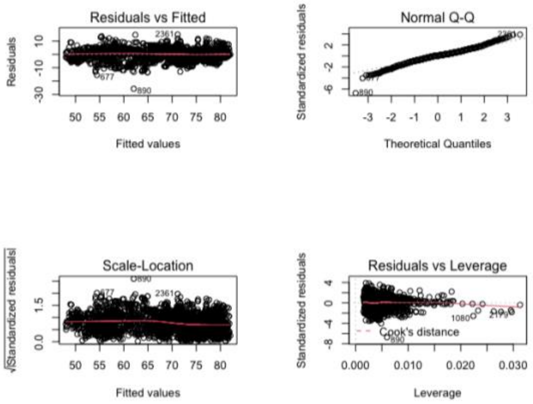


Figure 11: Final Complex MLR Model Residual Plots

Parameter Interpretation

The final model can be interpreted as follows if all other factors are held equal:

- Intercept: Average life expectancy with no other factors would be 45.58 years.
- Status-Developing: If a country is considered developing, life expectancy would be reduced by 3.11 years.
- Continent-Americas: If a person resides in the American continents, life expectancy would increase by 3.41 years.
- Continent-Asia: If a person resides in the Asian continent, life expectancy would increase by 1.1 years
- Continent-Europe: If a person resides in Europe, life expectancy would increase by 1.01 years.
- Continent-Oceania: If a person resides in Oceania, life expectancy would be reduced by .04 years.
- Income Composition of Resources: If a person were to see an increase of one in Income Composition of Resources life expectancy would rise by 19.57 years.
- Schooling: An increase in schooling by 1 year would see and increase to life expectancy of .31 years.
- BMI: An increase of 1 to the log of a person's BMI would see an increase in life expectancy of 1.81 years.
- HIV/AIDS: And increase of 1 to the log of infant deaths per 1000 births due to HIV/AIDs, would see life expectancy drop by 2.57 years.
- Schooling*BMI: This interaction term is created by multiplying years of schooling by the log of the BMI of an individual. When this value is increased by 1, life expectancy drops by .03 years.

kNN - Nonparametric Model

We chose the kNN regression as our non-parametric model. The kNN model approximates the association between independent variables and continuous response variable by averaging the observations in the same neighborhood. It uses the Euclidean distance between the nearest neighbors to make the prediction. It does not have any complex way of predicting. As you create the model you have the choose to choose the number of K's. Which means that k=1, will choose the closest neighbor 1 neighbor, while k=5 will choose 5 close neighbors to make the prediction.

The kNN model cannot be overfitted since you must check which value of k is the best. The emphasis is less on explanatory variables since they will end up becomes points for k to make the prediction of life expectancy. It reduces the risk of overfitting does not completely negate it. But your options reduce the K is again which choosing the number of k's. It should be noted that if you have a large dataset and you choose small k. You can run into risk of overfitting due to smaller k.

Nonparametric Model Result

The metric that was chosen to see how the model performs was the RMSE. This tells how well the model is performing and we would want a lower RMSE. The ASE for kNN does not have impact since choosing the number of k's is what is important. The model was tested from k=1 to k=50 and the result as shown in Figure 12. In the tests it showed the lower number of k's produced the best results. But overall predictions in the graphs below show that in application the models do not perform well when looking at the comparison between actual versus predicted regardless of the k value chosen.

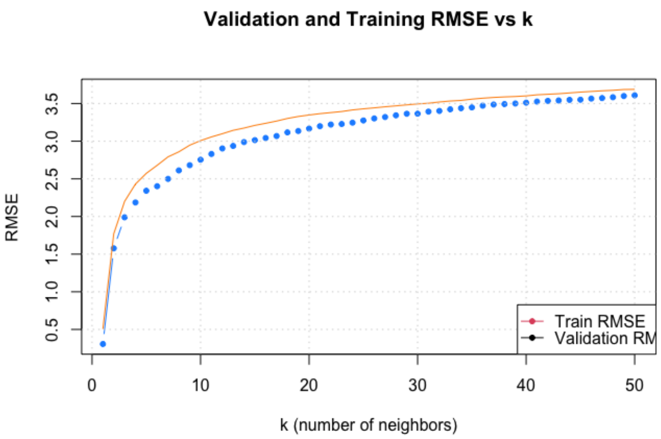


Figure 12: RMSE Results for k=1 to k=50

Model Performance: Actual versus Predicted with Different k Values:

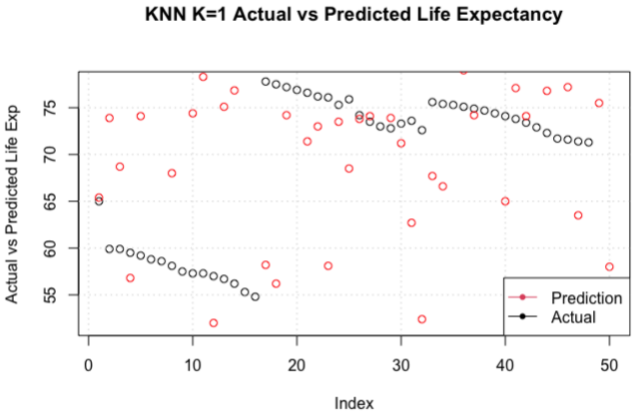


Figure 13: Actual vs. Predicted with k=1

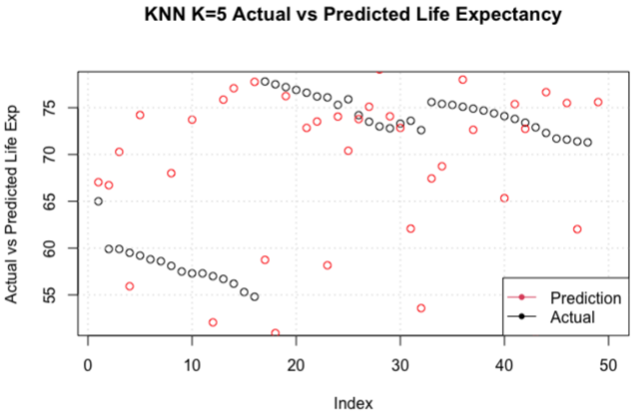


Figure 14: Actual vs. Predicted with k=5

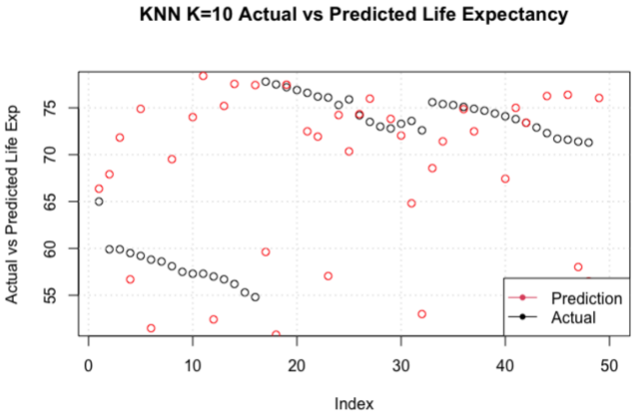


Figure 15: Actual vs. Predicted with k=10

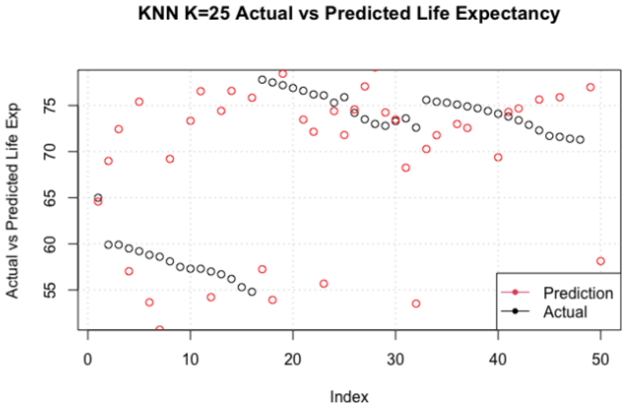


Figure 16: Actual vs. Predicted with k=25

Comparison of Model Results

To compare against the three models we are looking at ASE. We will also be discussing the issue of over fit for these three models. Below is a table of the ASE's compared against each other.

Model	Simple MLR	Complex MLR	KNN
ASE	14.71	14.97	.1013

From this comparison alone the KNN model looks to be the best model for predicting life expectancy based on the data given. Ruling out the complex MLR model because it is the highest ASE, we can dive deeper into the benefits and drawbacks of each of the remaining two models. With the KNN model, it has a high bias because of how the model functions. This leads to a situation of overfitting, where it will work very well with the data set given but will be questionable on performing as well with others. It's low ASE could also just be due to the small percentage of the test set that was used. Had a larger test set been used it might not have performed as well. The simple MLR model though has a higher ASE but it will also be more consistent on testing similar data sets. As this project is more about recognizing a trend than trying to predict something critical, I think the KNN model would probably be better for looking at these trends as overfitting should not cause too many issues for this historical data set. I discounted the complex MLR simply because it didn't make a lot of sense with what we were trying to accomplish. There was only one interaction term without high VIR. This would probably not have been the case had we not started with some manual feature selection. But we wanted to make sure our model made sense and was not just including interaction terms for the sake of being overly complex.

Final summary

Through Objective 1, a simple multiple linear regression model was found by handpicking variables that were highly correlated with Life Expectancy. The final simple regression model only removed one variable, GDP_log, and one Continent category, Oceania. In Objective 2, interaction terms were added to the multiple linear regression model. One interaction, BMI_log * Schooling, was found significant with a low variable inflation factor. Both the simple and complex multiple linear regression models had similar ASEs. The non-parametric model, kNN, had a good Root Mean Square Error, yet had trouble during the prediction phase.

The models within this study are statistically significant and have good R-square and Average Square Error/Root Mean Square Error values. This data was provided as an observational study, therefore there are no causal inferences that can be made against the population. We are able to make inferences in reference to the life expectancy of a specific country between 2000 and 2015.

If given more time, we would have added more of the original variables to the model selection process. In the process of Exploratory Data Analysis, many variables were removed due to low correlation or assumption violation. This presents a potential for the team to miss interactions that could assist with the error explanation.