Predicting Life Expectancy for Individual Countries

Introduction

The average life expectancy of a country is the average lifespan of its citizens. There is a lot of variability between countries. For example, the life expectancy in Norway is 72 years while the life expectancy in Mali is only 26 years [2]. The motivation for creating this model is to provide countries with a way of understanding what factors most impact their country's life expectancy. Predicting life expectancy from multiple variables can help country leaders know what they need to do to help increase their country's life expectancy. Then, they can focus their attention on those areas. For example, if the HIV/AIDS rates have a big negative impact on a country's life expectancy, they would know to focus on helping their citizens have access to medicine for HIV/AIDS. If their Polio rates are not significantly impacting their life expectancy, then the country does not need to focus more on preventative measures for Polio. The model we are trying to create will help administrations help their citizens increase their life expectancy through policies.

Data Exploration

Our data set was retrieved from kaggle.com [3]. There are 2,938 observations recorded from around the world. We are trying to predict life expectancy while taking into account 15 different variables, including our response variable. Some of the variables include schooling, alcohol-consumption rates, etc. A full list of variables and their descriptions are shown in Table 1. This data was recorded from The Global Health Observatory (GHO) under The World Health Organization (WHO). We chose data from the year 2014. We also decided not to use the variables that included mortality statistics like infant death because life expectancy is calculated from those statistics, and we are trying to find a statistical relationship, not a functional relationship. We also removed all data entries with missing values and ended with a total sample size of 131.

Variable Name	Description
Life Expectancy	Life Expectancy in age
Country	Country
Polio	Polio (Pol3) immunization coverage among 1-year-olds (%)
Diphtheria	Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)
Hepatitis B	Hepatitis B (HepB) immunization coverage among 1-year-olds (%)
HIV/AIDS	Deaths per 1 000 live births HIV/AIDS (0-4 years)
Measles	Measles - number of reported cases per 1000 population
Alcohol	Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol)
BMI	Average Body Mass Index of entire population
Schooling	Number of years of Schooling(years)
Thinness 5-9 years	Prevalence of thinness among children for Age 5 to 9(%)
Thinness 10-19 years	Prevalence of thinness among children and adolescents for Age 10 to 19 (%)
GDP	Gross Domestic Product per capita (in USD)
Population	Population of the country
Income Composition of Resources	Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
Total Expenditure	General government expenditure on health as a percentage of total government expenditure (%)
Percentage Expenditure	Expenditure on health as a percentage of Gross Domestic Product per capita(%)

Table 1: Definition of Variables

Scatter plot matrices for the variables are shown below. Figure 1 shows a scatter plot matrix of life expectancy against immunization rates. It shows that immunization rates are positively correlated with life expectancy as well as with each other. Figure 2 shows a scatter plot matrix of life expectancy against disease rates. It reports that HIV/AIDS is negatively correlated with life expectancy. Figure 3 shows a scatter plot matrix of life expectancy against adult lifestyle factors. Alcohol appears to have a lot of zero values. BMI appears to have some sort of non-linear association with life expectancy. Figure 4 shows a scatter plot matrix of life expectancy against child lifestyle factors like schooling and thinness. Schooling shares a remarkably clear linear relationship with life expectancy. The thinness data seems to be highly multicollinear and shows a negative correlation with life expectancy. There seems to be two points that do not fit into the trend that the rest of the points follow. Figure 5 shows a scatter plot matrix of life expectancy against several more general country data points. Of note, income composition and other spending related data points share a clear linear relationship with life expectancy, but GDP does not. Here we can also see an abnormally high value in population that might lead to high leverage.

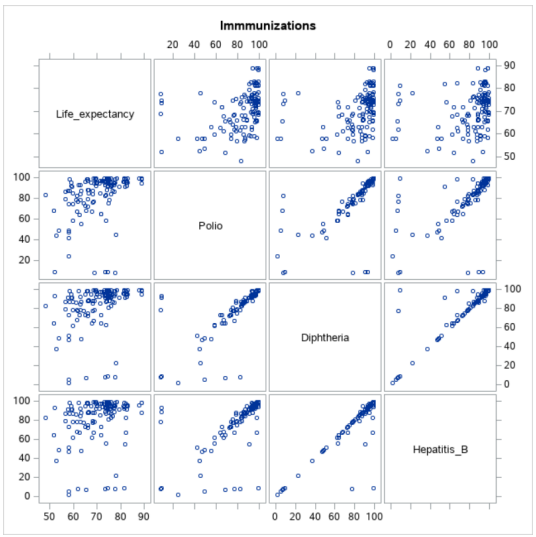

Figure 1: Scatter Plot Matrix of Life Expectancy with Immunization Data

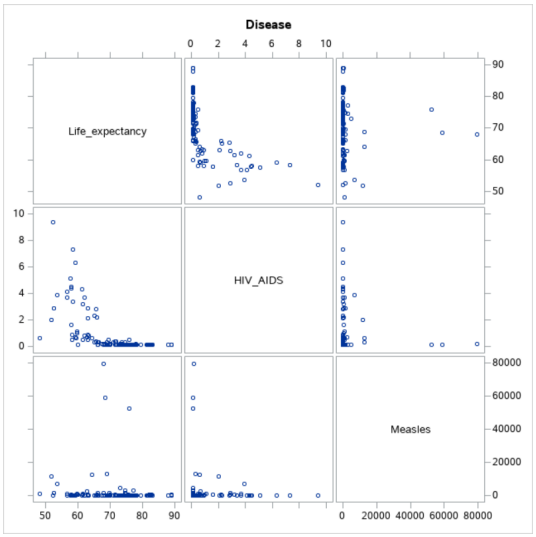

Figure 2: Scatter Plot Matrix of Life Expectancy with Disease Data

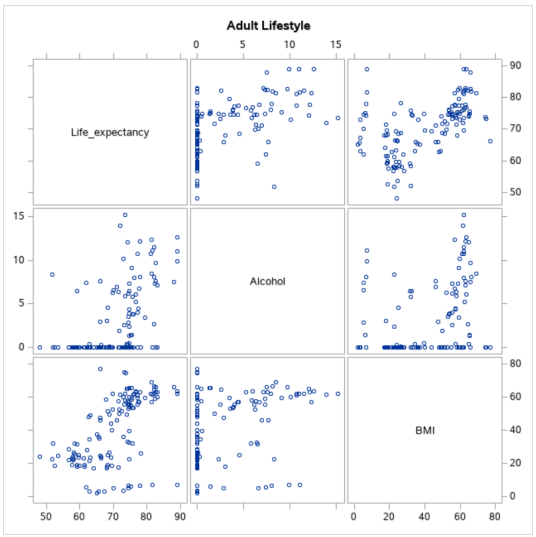

Figure 3: Scatter Plot Matrix of Life Expectancy with Adult Lifestyle Data

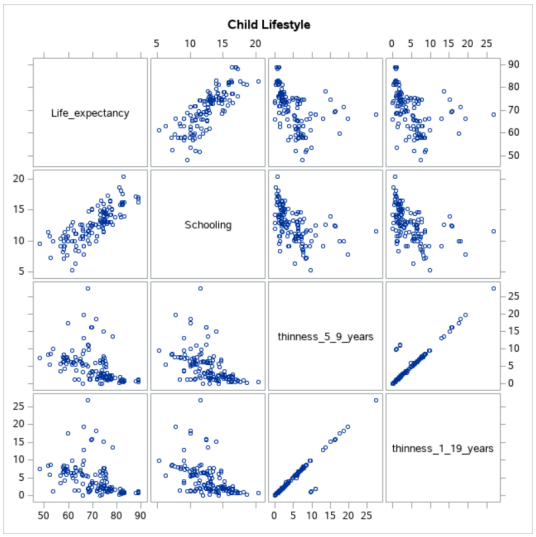

Figure 4: Scatter Plot Matrix of Life Expectancy with Child Lifestyle Data

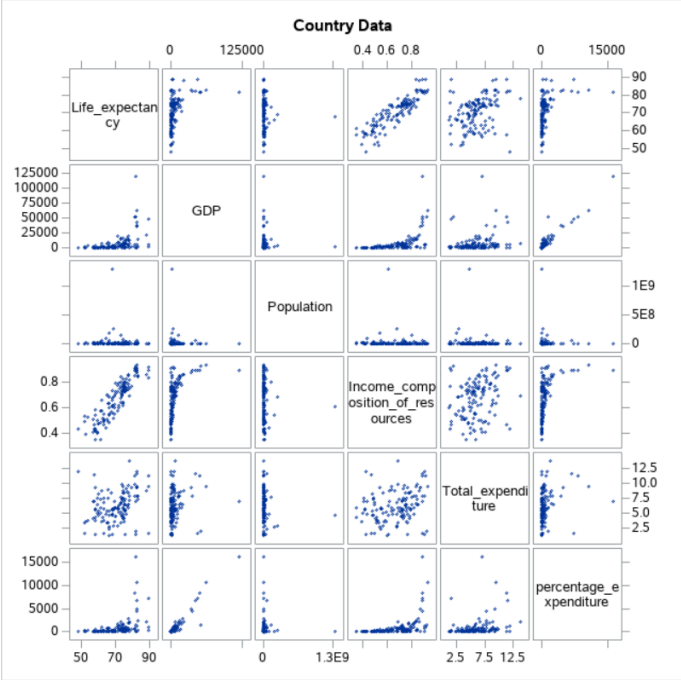

Figure 5: Scatter Plot Matrix of Life Expectancy with General Country Data

Model Assumptions for the Full Model

In the previous section, we found several variables that had visually linear relationships with life expectancy. This validates the model assumption that X and Y share a linear relationship. We also saw some potentially high leverage points like the high value in population. The R-student against leverage plot shown in Figure 6 shows that India has high leverage. It also shows that Sierra Leon could be an outlier. These two observations will be looked at in more detail in the remedial measures section. Figure 7 shows the normal probability plot for the residuals. It looks linear, so we conclude that the residuals are normally distributed. All of the data is taken in 2014, so we don't need to worry about time dependence. However, there still might be spatial dependence because of the geographical location of the countries. We are going to ignore that for this analysis and move forward with the assumption of independent observations. The residual vs predicted value plot in Figure 8 shows no signs of heteroskedasticity (non-constant variance in the residuals), so we'll take this model assumption to be satisfied.

We conducted tests for multicollinearity, looking for variables that have a high linear relationship with one another. There were a couple of variables that showed multicollinearity. Amongst those variables were population and schooling. Both of these variables reported a condition index of 10 or higher and at least two or more proportions of variation were more than 50%. We conclude that these two variables show multicollinearity amongst our data. This multicollinearity may affect any inference we try to make from the coefficients, but it does not affect the predictive power of the model.

In summary, we have decided that our model adequately meets all of the model assumptions except we found potential outliers / influential points. We will address the potentially problematic observations in the remedial measures section.

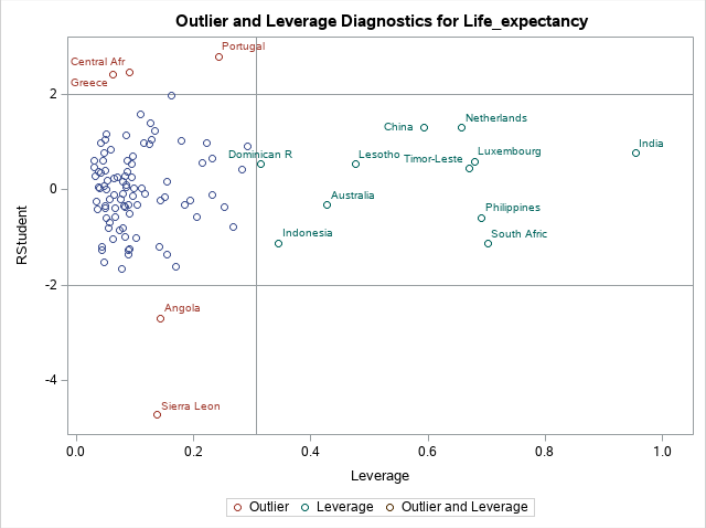

Figure 6: Rstudent vs Leverage Plot

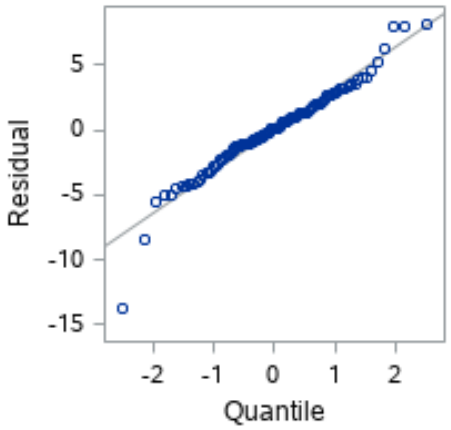

Figure 7: Normal Probability Plot

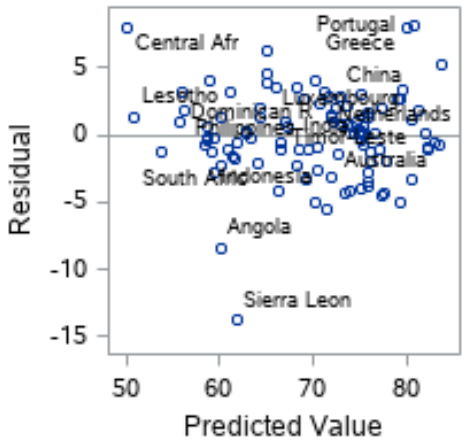

Figure 8: Residuals vs Predicted Values

Remedial Measures

In the previous section, we found that there were two problematic observations: India and Sierra Leon. Figure 9 shows the Cook's D plot. We can see that India has extremely high influence, this may be due to having a large population. After log-transforming population, we retrieved the Cook's D plot shown in Figure 10. This proved to fix our model assumption violation. We also performed a log transformation on a few variables; percentage expenditure, GDP, thinness 5-9 years, and thinness 10-19 years to better center these observations around their mean.

The problem with Sierra Leon is that life expectancy was severely over-predicted by the model. There is nothing in its X-profile that gives an explanation for this. However, according to Gire et. al. in *Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak*, Sierra Leon was affected by the Ebola virus in 2014 [1]. In this analysis, we tried to include data on infection rate as a control measure, but unfortunately, we don't have any data on the Ebola virus. Because this is such a unique situation, we're going to remove Sierra Leon from our data set with the understanding that our model shouldn't be used to predict life expectancy for countries that have been impacted by unexpected viruses.

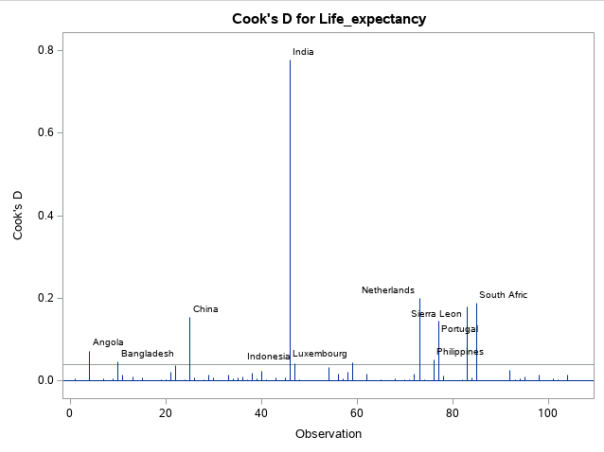

Figure 9: Cook's Distance Plot Before Transformation

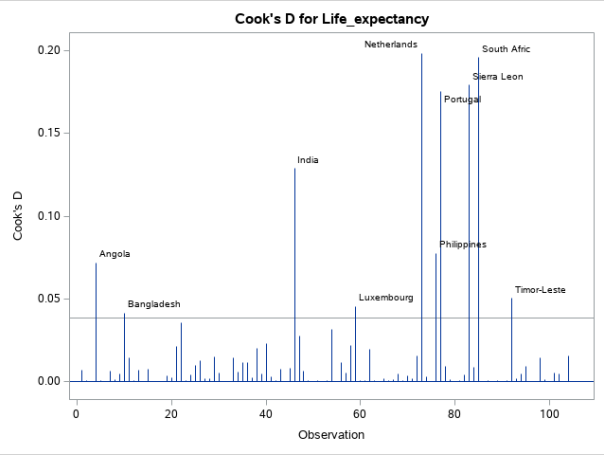

Figure 10: Cook's Distance Plot After Transformation

Variable Selection

We want to perform variable selection so we can isolate the variables that are most important in predicting a country's life expectancy. Ideally, we want to narrow it down to just a handful of variables that a country can put their focus on to increase their life expectancy. We will attempt variable selection using three different techniques: backward selection with $slstay = .1$, stepwise selection with $slentry = .05$ and $slstay = .1$, and elastic net with 10 cross-validation groups. Also, we will compare these models by looking at their Adjusted R-Squared value, their MSPR, and the number of variables included in each model.

The results of each method are shown Table 2. Of note, stepwise selection and elastic net selected the same variables. Backward selection selected all of these and then some. Both stepwise selection and elastic net settled on the same variables which is promising. They also settled on fewer variables than backward selection, which is another plus. Table 3 shows the

R-squared and MSPR evaluations of each model. The adjusted R-squared for the backward-selected model was slightly higher than the adjusted R-squared for the stepwise model, but the stepwise-selected model did significantly better on the test set, so we are going to favor the stepwise-selected model over the backward-selected model. In the same way, the elastic net had a lower adjusted R-squared than either but did the best on the test set.

Variable Name	Backward Selection	Stepwise Selection	Elastic Net
Hepatitis B	✓	✓	✓
Total Expenditure	✓	✓	✓
HIV/AIDS	✓	✓	✓
Log Thinness 5-9 years	✓	✓	✓
Income Composition of Resources	✓	✓	✓
Alcohol	✓		
Log GDP	✓		
Log Percentage Expenditure	✓		
Infant Deaths			
Adult Mortality			
Measles			
BMI			
under five deaths			
Polio			
Diphtheria			
Log Population			
Log Thinness 10-19 years			
Schooling			

Table 2: Variable Selection Table

	backward Selection	Stepwise Selection	Elastic Net
Adj R-Squared	0.885	0.882	0.877
MSPR	12.675	11.383	10.728

Table 3: Variable Selection Measures

Interaction Terms

We tested three different interaction terms to see if some of our variables should be considered jointly. First, we considered the variables Schooling and Total Health Expenditure to see if they should be included in the model together. We hypothesize that a country which spends a high amount of money on health initiatives, would also spend a high amount of money on schooling. Also, we considered Schooling and Thinness from ages 5-9 because schooling involves children ages 5-9. Lastly, we decided to test for Schooling and Income because the income composition of resources should reasonably play a part in how schooling is funded.

The p-values for each of these interaction terms came out as 0.097 for Schooling and Total Expenditure, 0.271 for Schooling and Thinness, and 0.201 for Schooling and Income. These p-values are quite large which means that there is not significant evidence to suggest that the variables we chose need to be considered jointly. We decided not to include any of these interaction terms in our model.

Model Assumptions for the Final Model

As we did in the original modeling assumptions and remedial measures, we conducted various tests to check the variables within our model. In Figure 11, we included the Cook's D plot for Life Expectancy for our final model. It shows that after our original residual measures and variable selection, our variables have somewhat balanced and there are no countries that significantly skew the results. The values in Cook's D plot for the countries with the most effect are significantly lower than they were originally, so we do not need to worry much about them.

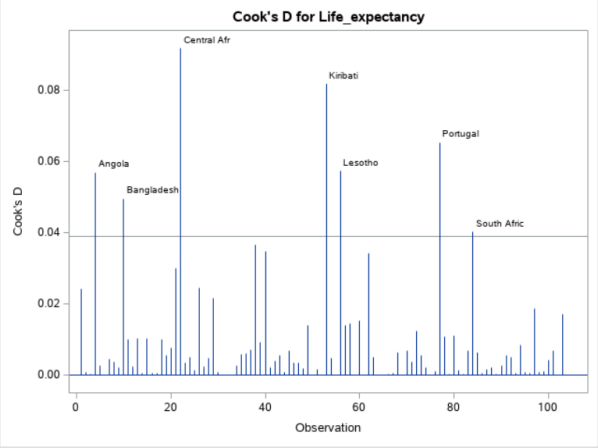

Figure 11: Cook's D Plot for the Final Model

Figure 12 shows the leverage and outlier plot for our final model. There are a few countries that still have some leverage over our model, but compared to our original leverage plot in Figure 9, the amount of leverage those countries have has decreased drastically.

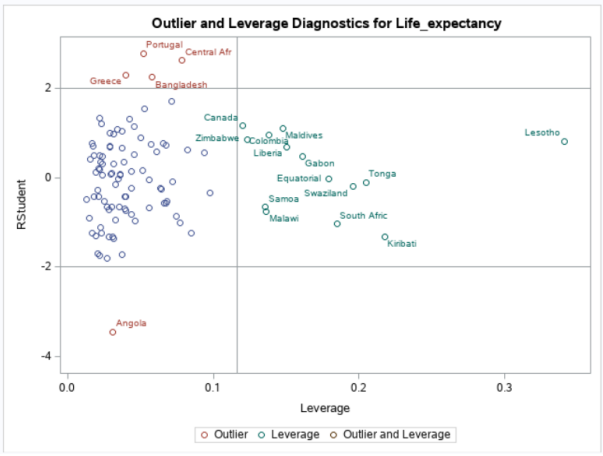

Figure 12: Rstudent against Leverage Plot for the Final Model

In Figure 13, the residual information for the final model is included. The plot that is to the left, the residual plot, shows that the countries within our data are equally surrounding the value 0 and have constant variance. The residual plot against a normal distribution, found in the middle of Figure 13, shows that the data is mostly following the normal curve. The plot farthest to the right, the histogram with a normal curve, shows that the majority of the data fits under the normal curve. However, we realize that Angola is still appearing as an outlier, but after researching, we could not find a justification to remove the point. We decided to keep the point in and realize that this may impact the parameter estimates.

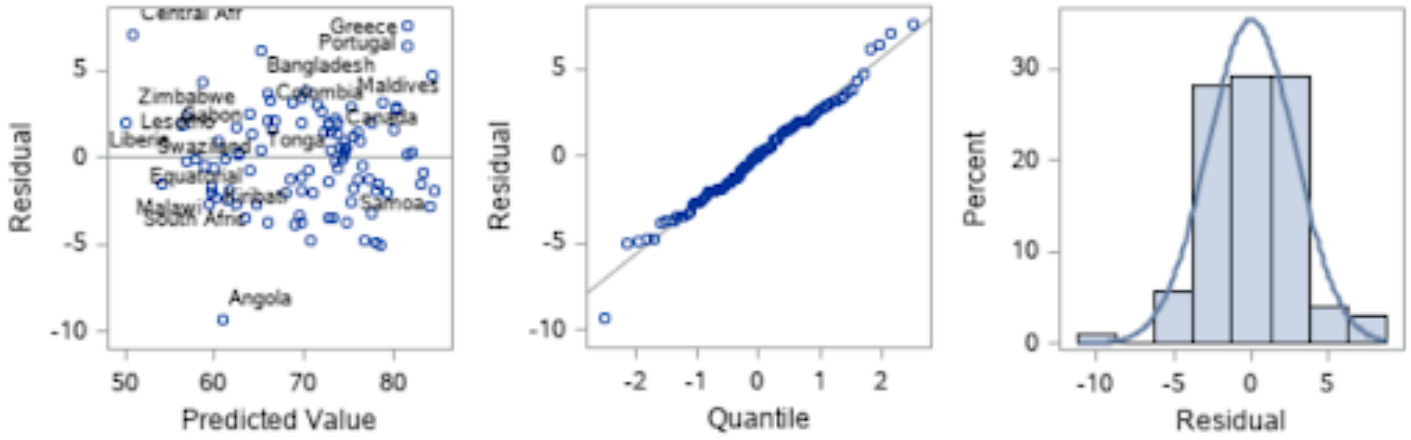

Figure 13: Residual Information for The Final Model

We tested for multicollinearity in our final model and nothing stood out as concerning. The variance inflation factors are shown in Figure 14. Because none of them are greater than 10, we conclude that none of the variables in our model are multicollinear. This means that our beta estimates should be interpretable.

Final Model Equation and Inference

Our final model equation is shown in Equation 1. The definitions of each X variables are shown, in order, in Figure 14, along with additional parameter information. Each β value is related to the variables in our model; for example, as Income Composition of Resources increases by one unit, our predicted life expectancy increases by 36.125 years as long as all other variables are held constant. Income Composition of Resources ranges from zero to one, which means that our prediction of life expectancy can vary by 36.125 years, which is highly significant. Another example would be, if Total Expenditure increases by 1 unit while all other variables are held constant, life expectancy is expected to increase by .484 years. Total Expenditure is the percentage of a government's expenditure that's spent on health, so practically what this means is that we'd predict a 1 year increase in life expectancy for a country that increases their Total Expenditure by 2-3 percent. On the other hand, as HIV/AIDS increases by one, the predicted average life expectancy decreases by 1.352 years, as all other variables are held constant. HIV/AIDS is the number of deaths per 1,000 live births due to AIDS/HIV (0-4 years). As can be seen visually in Figure 15, observed values fell between 0.1 and 9.4, with most observations being below 1. For countries in which infant deaths due to AIDS/HIV is a problem, addressing this issue would be a good method for bringing expected life expectancy up, as 1.352 years per 1 value decrease is significant.

$$Y = 42.963 + 0.0321X_1 + 0.484X_2 - 1.352X_3 - 0.917X_4 + 36.125X_5 \quad (1)$$

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	42.96330	2.04876	20.97	<.0001	0
Hepatitis_B	1	0.03212	0.01335	2.41	0.0180	1.16543
Total_expenditure	1	0.48425	0.12200	3.97	0.0001	1.11067
HIV_AIDS	1	-1.35214	0.19661	-6.88	<.0001	1.33640
log_thinness_5_9_years	1	-0.91695	0.31115	-2.95	0.0040	1.45130
Income_composition_of_resources	1	36.12531	2.64833	13.64	<.0001	1.85761

Figure 14: Final Model Parameter Estimates

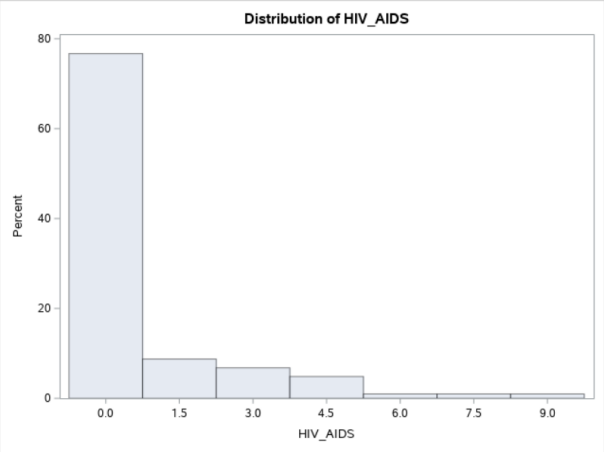

Figure 15: Histogram of infant AIDS/HIV Deaths

OLS Alternative

Due to the remaining possible outliers after remedial measures were performed, it was decided to perform a robust regression using the M-estimation. The M-estimation controls for the outliers in the data. The resulting R-square value of the robust regression was 0.767. This indicates that the robust regression model accounts for approximately 77% of the variance in the data. By observing the p-values we obtained from the Chi-square test, it was concluded that the variables which were significant using a significance level $\alpha = .05$ were the Intercept, Total Expenditure, HIV-AIDS, and Income Composition of Resources. All parameter estimates and significance test statistics are shown in Figure 16. To compare this method to the linear regression model, the validation error for each model will be calculated and compared in the next section.

Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	40.8246	2.5715	35.7845	45.8647	252.04	<.0001
Alcohol	1	-0.1540	0.1110	-0.3715	0.0635	1.93	0.1653
percentage_expenditu	1	0.0003	0.0005	-0.0007	0.0013	0.31	0.5772
Hepatitis_B	1	0.0333	0.0409	-0.0468	0.1134	0.66	0.4154
Measles	1	0.0000	0.0000	-0.0001	0.0001	0.00	0.9834
BMI	1	0.0008	0.0226	-0.0435	0.0451	0.00	0.9718
Polio	1	-0.0006	0.0215	-0.0426	0.0415	0.00	0.9795
Total_expenditure	1	0.4456	0.1499	0.1519	0.7394	8.84	0.0029
Diphtheria	1	-0.0078	0.0469	-0.0997	0.0841	0.03	0.8683
HIV_AIDS	1	-1.3684	0.2269	-1.8132	-0.9236	36.36	<.0001
GDP	1	-0.0000	0.0001	-0.0002	0.0001	0.33	0.5652
Population	1	0.0000	0.0000	-0.0000	0.0000	0.14	0.7044
thinness_1_19_years	1	0.0071	0.2888	-0.5589	0.5730	0.00	0.9805
thinness_5_9_years	1	-0.1121	0.2867	-0.6740	0.4498	0.15	0.6957
Income_composition_o	1	42.9573	6.1447	30.9138	55.0008	48.87	<.0001
Schooling	1	-0.1400	0.2850	-0.6987	0.4186	0.24	0.6232
Scale	1	2.8043					

Figure 16: Robust Regression Parameter Estimates

In addition to the robust regression model containing all of the original variables, a robust regression model was fit using the subset of variables obtained from variable selection. It was found that the resulting R-square value was 0.756 for the robust regression model using the described subset of variables. All parameter estimates and significance test statistics are shown in Figure 17. To compare this method to the other models described, the validation error will be calculated and compared in the next section.

Parameter Estimates							
Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	43.1204	2.0329	39.1360	47.1048	449.92	<.0001
Hepatitis_B	1	0.0309	0.0133	0.0048	0.0570	5.37	0.0205
Total_expenditure	1	0.4160	0.1170	0.1867	0.6454	12.64	0.0004
HIV_AIDS	1	-1.3637	0.1948	-1.7455	-0.9820	49.02	<.0001
log_thinness_5_9_yea	1	-0.7982	0.3105	-1.4068	-0.1895	6.61	0.0102
Income_composition_o	1	36.4570	2.5825	31.3954	41.5185	199.29	<.0001
Scale	1	2.7288					

Figure 17: Robust Regression Parameter Estimates with variable subset

Accuracy

Pertaining to the accuracy within our model, we have reported the validation error for the original model, the null, and the final model which is found in Table 4. We report the mean number for each model in hopes to find the lowest number. The validation error reported for the original model is 12.779. The validation error for the null hypothesis is 70.875. Lastly, the validation error for the final model is 11.383. From these procedures, we can conclude that our final model, after remedial measures and variable selection, was found to have the least amount of error amongst the three models. There was a vast difference in error between the null model and the initial model while the difference between the initial and final models was smaller. However, the error still decreased while also decreasing the number of variables, which is good. Our full Robust Regression model (including all the variables) has a validation error of 12.278 which is slightly less than our initial model error, but it is still higher than our final Ordinary Least Squares regression model. We also created a robust regression model using only the variables we selected in the variable selection section. It has a validation error of 11.093, which is the best of any of our tested models. We can see that the maximum error is very high for all of the models. They are all predicting Belgium poorly. Belgium has the highest life expectancy in our data set. The best models are predicting a life expectancy in the 70s for Belgium, which is very high, but Belgium actually has a life expectancy in the 80s.

Model	Mean	Std Dev	Min	Max
Null	70.875	85.948	0.091	349.617
Full	12.779	21.601	0.011	103.920
Final	11.383	20.257	9.031e-5	102.056
Robust Full	12.278	24.350	4.363e-4	123.640
Robust Final	11.093	19.689	0.066	97.640

Table 4: Accuracy Table

Conclusion

The final model achieved with Ordinary Least Squares regression after variable selection and remedial measures helps predict a country's average life expectancy based on variables such as Total Expenditure, Income Composition of Resources, and HIV/AIDS rates. Our regression model helps show that multiple factors affect life expectancy. This helps countries know what variables their country should focus on to increase average life expectancy for their citizens.

Our final regression model was the most accurate of our tested models based on validation error. By using variable selection and remedial measures, we created a model with lower validation error than the initial model with all variables and the null model. We also tested the validation error for our alternative method of regression, Robust Regression, which had a higher validation error than our final model, but had a lower error than our initial Ordinary Least Squares Model. In practice, this means that we're probably going to favor our final OLS model.

One future direction of research that we considered was doing a full time-series study. It would be interesting to see how the predicted life expectancy changes throughout multiple years rather than just 2014. We could use that to find possible long term patterns in a country that affect life expectancy. Also, it would be interesting to research possible case studies for a specific country to see if any of these variables we chose has a specific causal relationship. For example, we could find a country that has put specific measures to lower HIV/AIDS rates and see if their life expectancy changes as expected. Another suggestion we would make for future researchers is to conduct a study about the prevalence of the Coronavirus in each country and how it may impact the predicted life expectancy for that given country. This would help address the issue we ran into with Sierra Leon, where we weren't able to handle predictions for epidemics and pandemics.

References

[1] Gire SK et. al. Genomic surveillance elucidates ebola virus origin and transmission during the 2014 outbreak. *Science*, 2014. doi: 10.1126/science.1259657. Epub 2014 Aug 28. PMID: 25214632; PMCID: PMC4431643.

[2] Esteban Ortiz-Ospina Max Roser and Hannah Ritchie. Life expectancy. *Our World in Data*, 2013. https://ourworldindata.org/life-expectancy.

[3] Kumar Rajarshi. Life expectancy (who) statistical analysis on factors influencing life expectancy. 2017. https://www.kaggle.com/kumarajarshi/life-expectancy-who.

Appendix

Introduction and Data Exploration SAS Code

```
/* Group 1 Final Project */
/* Designate file path */
FILENAME REFFILE '<file path removed for to preserve anonymity>';

/* Import Data set */
PROC IMPORT DATAFILE=REFFILE replace DBMS=CSV OUT=WORK.life;
GETNAMES=YES;
RUN;

/* Create data set with only 2014 data and delete observations with any missing values *
data life2014; set life;
where year = 2014;
if nmiss(of _numeric_) > 0 then delete;
run;

/* ************************************************************************************
/* Make a scatter plot matrix of non-categorical variables */
proc sgscatter data=life2014;
matrix life_expectancy alcohol percentage_expenditure Hepatitis_B Measles;
title1 'Life Expectancy Data';
run;

/* Make a scatter plot matrix of non-categorical variables */
proc sgscatter data=life2014;
  matrix life_expectancy BMI polio total_expenditure diphtheria;
title1 'Life Expectancy Data';
run;

/* Make a scatter plot matrix of non-categorical variables */
proc sgscatter data=life2014;
matrix life_expectancy HIV_AIDS GDP population thinness_1_19_years;
title1 'Life Expectancy Data';
run;

/* Make a scatter plot matrix of non-categorical variables */
proc sgscatter data=life2014;
matrix life_expectancy thinness_5_9_years income_composition_of_resources schooling;
title1 'Life Expectancy Data';
run;
```

Original Model Assumptions SAS Code

```
/* ********************************************************************************
/* Separate Into Training and Test Sets.
Only Fit Models to the Training Set. The variable
"Selected" separates training (0) from test (1) */
proc surveyselect data=life2014 seed=12345 out=life2014out
rate=0.2 outall;/* Withold 20% for validation */
run;

data train; set life2014out; if Selected=0; robusty = life_expectancy;
run;
data test; set life2014out; if Selected=1;
run;

data combined;
set train test;
run;

/* Look at crude initial model */
proc reg data=train plots(label)=(CooksD RStudentByLeverage DFFITS DFBETAS);
id country;
model life_expectancy = alcohol percentage_expenditure Hepatitis_B Measles
BMI polio total_expenditure diphtheria HIV_AIDS GDP population thinness_1_19_years
thinness_5_9_years income_composition_of_resources schooling/ vif collin;
store regModel;
output out=train_out residual=residual predicted=predicted;
run;

/* Load macro for BF-test of constant variance and correlation test of normality. */
%macro resid_num_diag(dataset,datavar,label='requested variable',predvar=' ',predlabel='

/* Run the BF-test for constant variance and the correlation test of normality */
%resid_num_diag(dataset=train_out, datavar=residual, label='residuals',
predvar=predicted, predlabel='predicted values');

/* Look at suspect observations for outliers and/or influential points */
proc print data=train;
where country = 'India' | country = 'Sierra Leon';
var country life_expectancy alcohol percentage_expenditure Hepatitis_B Measles
BMI polio total_expenditure diphtheria HIV_AIDS GDP population thinness_1_19_years
thinness_5_9_years income_composition_of_resources schooling;
title1 'Suspect observations';
run;
```

Remedial Measures SAS Code

```
/* Remove Sierra Leon (Due to Ebola) */
data train; set train;
if country = 'Sierra Leon' then delete;
run;

/* Look at all histograms before transformations */
proc univariate data=train noprint;
histogram life_expectancy alcohol percentage_expenditure Hepatitis_B Measles
BMI polio total_expenditure diphtheria HIV_AIDS GDP population thinness_1_19_years
thinness_5_9_years income_composition_of_resources schooling;
run;

/* Modify training set with log of: population, %expenditure GDP thinness_1-19, thinness
data train2; set train;
log_population = log(population);
log_percentage_expenditure = log(percentage_expenditure);
log_GDP = log(GDP);
log_thinness_1_19_years = log(thinness_1_19_years);
log_thinness_5_9_years = log(thinness_5_9_years);
run;

/* Look at histograms of transformed variables */
proc univariate data=train2 noprint;
histogram log_percentage_expenditure log_GDP log_population
log_thinness_1_19_years log_thinness_5_9_years;
run;

/* Look at model with log of: population, %expenditure GDP thinness_1-19, thinness_5-9 *
proc reg data=train2 plots(label)=(CooksD RStudentByLeverage DFFITS DFBETAS);
id country;
model life_expectancy = alcohol log_percentage_expenditure Hepatitis_B Measles
BMI polio total_expenditure diphtheria HIV_AIDS log_GDP log_population log_thinness_1_19
log_thinness_5_9_years income_composition_of_resources schooling/ vif collin;
output out=train2_out residual=residual predicted=predicted;
run;
```

Variable Selection SAS Code

```
/* Perform Backward Elimination for variable selection */
proc reg data=train2;
model life_expectancy = alcohol log_percentage_expenditure Hepatitis_B Measles
BMI polio total_expenditure diphtheria HIV_AIDS log_GDP log_population log_thinness_1_19
log_thinness_5_9_years income_composition_of_resources schooling / selection=backward sl
title1 'Backward Elimination';
run;

/* Stepwise Selection */
proc reg data=train2;
model life_expectancy = alcohol log_percentage_expenditure Hepatitis_B Measles
BMI polio total_expenditure diphtheria HIV_AIDS log_GDP log_population log_thinness_1_19
log_thinness_5_9_years income_composition_of_resources schooling / selection=stepwise sl
title1 'Stepwise Selection';
run;

/* Elastic Net */
proc glmselect data=train2 plots=(criterion ase) seed=42069;
    class status;
    model life_expectancy = alcohol log_percentage_expenditure Hepatitis_B Measles
    BMI polio total_expenditure diphtheria HIV_AIDS log_GDP log_population log_thinness_
    log_thinness_5_9_years income_composition_of_resources schooling / selection=elastic
    cvmethod=random(10);
    output out=elastic_net_out p=predelasticnet;
    title1 'Elastic Net';
run;

/* Look at tentative final model from variable selection */
proc reg data=train2 plots(label)=(CooksD RStudentByLeverage DFFITS DFBETAS);
id country;
model life_expectancy = Hepatitis_B total_expenditure HIV_AIDS
log_thinness_5_9_years income_composition_of_resources / vif collin;
output out=train2_out residual=residual predicted=predicted;
title1 'Validity check after variable selection';
run;
```

Interaction Terms SAS Code

```
/* Test Interaction Terms */
data train2; set train2;
school_expend = schooling*total_expenditure;
school_logthin59 = schooling*log_thinness_5_9_years;
school_income = schooling*income_composition_of_resources;
run;
proc reg data=train2;
id country;
model life_expectancy = Hepatitis_B total_expenditure HIV_AIDS
log_thinness_5_9_years income_composition_of_resources schooling school_expend;
title1 'Testing Interaction Term Schooling*total_expenditure';
run;
proc reg data=train2;
id country;
model life_expectancy = Hepatitis_B total_expenditure HIV_AIDS
log_thinness_5_9_years income_composition_of_resources schooling school_logthin59;
title1 'Testing Interaction Term Schooling*log_thinness_5_9_years';
run;
proc reg data=train2;
id country;
model life_expectancy = Hepatitis_B total_expenditure HIV_AIDS
log_thinness_5_9_years income_composition_of_resources schooling school_income;
title1 'Testing Interaction Term Schooling*income_composition_of_resources';
run;
```

Final Model SAS Code

```
/* Final Model */
proc reg data=train2 noprint;
model life_expectancy = Hepatitis_B total_expenditure HIV_AIDS
log_thinness_5_9_years income_composition_of_resources;
store regModel2;
output out=testing residual=residual predicted=predicted;
run;

/* Add log_thinness_5_9_years to test data set */
data test; set test;
log_thinness_5_9_years = log(thinness_5_9_years);
run;

/* Fit a model with NO variables */
proc reg data=train2 noprint;
model life_expectancy = ;
store regModel3;
run;
```

Alternative to OLS (Robust Regression) SAS Code

```
/* TESTING ALTERNATIVE TO OLS */

/* Weighted Least Squares */
/* DB Look for relationship between SD of resid and X */
/* data train2_out; set train2_out; */
/*  abs_resid = abs(residual); */
/* DB Get estimate of SD of resid based on X */
/* proc reg data=train2_out noprint; */
/*  model abs_resid = alcohol percentage_expenditure Hepatitis_B Measles */
/*  BMI polio total_expenditure diphtheria HIV_AIDS GDP population thinness_1_19_years *
/*  thinness_5_9_years income_composition_of_resources schooling; */
/*  output out=out1 p=estSD; */
/* run; */
/* DB Define weight */
/* data out1; set out1; */
/*  useWeight = 1/estSD**2; */
/* run; */
/* DB Fit WLS model */
/* proc reg data=out1; */
/*  model life_expectancy = alcohol percentage_expenditure Hepatitis_B Measles */
/*  BMI polio total_expenditure diphtheria HIV_AIDS GDP population thinness_1_19_years *
/*  thinness_5_9_years income_composition_of_resources schooling; */
/*  weight useWeight; */
/*  store WLSModel; */
/*  title1 'WLS model fit'; */
/* run; */

/* Robust regression  */
proc robustreg data=combined method=M (wf=bisquare);
model robusty = alcohol percentage_expenditure Hepatitis_B Measles
BMI polio total_expenditure diphtheria HIV_AIDS GDP population thinness_1_19_years
thinness_5_9_years income_composition_of_resources schooling;
output out=robustout predicted=predicted;
title1 'Robust (M) regression on training data';
run;
```

Accuracy SAS Code

```
/* ACCURACY SECTION ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++  */
/* Calculate MSPR of Initial Model - Step 1 */
proc plm restore=regModel;
score data=test out=newTest predicted;
run;
/* Calculate MSPR of Final Model - Step 1 */
proc plm restore=regModel2;
score data=test out=newTest2 predicted;
run;
/* Calculate MSPR of NO VAR Model - Step 1 */
proc plm restore=regModel3;
score data=test out=newTest3 predicted;
run;
/* Calculate MSPR of WLS Model - Step 1 */
/* proc plm restore=WLSModel; */
/*  score data=test out=newTest4 predicted; */
/* run; */


/* Calculate MSPR of Initial Model - Step 2 */
data newTest; set newTest;
MSE=(life_expectancy - Predicted)**2;
run;
/* Calculate MSPR of Final Model - Step 2 */
data newTest2; set newTest2;
MSE=(life_expectancy - Predicted)**2;
run;
/* Calculate MSPR of NO VAR Model - Step 2 */
data newTest3; set newTest3;
MSE=(life_expectancy - Predicted)**2;
run;
/* Calculate MSPR of WLS Model - Step 2 */
/* data newTest4; set newTest4; */
/*  MSE=(life_expectancy - Predicted)**2; */
/* run; */
/* Calculate MSPR of Robust Model - Step 2 */
data newTest5; set robustout; where Selected=1;
MSE=(life_expectancy - Predicted)**2;
run;

/* Calculate MSPR of Initial Model - Step 3 */
proc means data=newTest;
var MSE;
```

```
title1 'Initial Model';
run;
/* Calculate MSPR of Final Model - Step 3 */
proc means data=newTest2;
var MSE;
title1 'Final Model';
run;
/* Calculate MSPR of NO VAR Model - Step 3 */
proc means data=newTest3;
var MSE;
title1 'NO VAR Model';
run;
/* Calculate MSPR of WLS Model - Step 3 */
/* proc means data=newTest4; */
/*  var MSE; */
/*  title1 'WLS Model'; */
/* run; */
/* Calculate MSPR of Robust Model - Step 3 */
proc means data=newTest5;
var MSE;
title1 'Robust Model';
run;
```

Additional Corrections SAS Code

```
/* ADDITIONS AFTER CORRECTIONS */

/* Add log-thinness_5_9_years */
data combined; set combined;
log_thinness_5_9_years = log(thinness_5_9_years);
run;

/* Robust regression on Final Model variables */
proc robustreg data=combined method=M (wf=bisquare);
model robusty = Hepatitis_B total_expenditure HIV_AIDS
log_thinness_5_9_years income_composition_of_resources;
output out=robustout2 predicted=predicted;
title1 'Robust (M) regression on training data with Final Model variables';
run;

/* Calculate MSPR of Robust Model with Final Model variables */
data newTest6; set robustout2; where Selected=1;
MSE=(life_expectancy - Predicted)**2;
run;
```

```
/* Calculate MSPR of Robust Model with Final Model variables */
proc means data=newTest6;
var MSE;
title1 'Robust Model with subset of variables';
run;

/* Look at high MSE */
proc print data=newTest;
where MSE > 90;
var country;
title1 'High MSE 1';
run;

/* Look at high MSE */
proc print data=newTest2;
where MSE > 90;
var country;
title1 'High MSE 2';
run;

/* Look at high MSE */
proc print data=newTest3;
where MSE > 90;
var country;
title1 'High MSE 3';
run;

/* Look at high MSE */
proc print data=newTest5;
where MSE > 90;
var country;
title1 'High MSE 5';
run;

/* Look at high MSE */
proc print data=newTest6;
where MSE > 90;
var country;
title1 'High MSE 6';
run;
```