MSDS 6372: Project 1

Submitted by:

Senthil Kumar

Simerpreet Reddy

Taifur Chowdhury

Date: 10/4/2020

**1 - Introduction**: In this document report, we present the different models created to predict Life Expectancy of a country’s population based on different variables ranging from immunization factors, mortality factors, economic factors, social factors and other health related factors. We also dive deep into some variables which strongly influence Life Expectancy of a country’s population.

**2 - Data Description:** The data set we have captures 22 columns for 193 countries from year 2000-2015 over 2938 rows. The data set has 20 predicting variables. Some of the variables that we will be talking about throughout our analysis are listed below:

Life expectancy – Life Expectancy in age.

Status- Developed or Developing status of a country.

percentage expenditure- Expenditure on health as a percentage of Gross Domestic Product per capita (%).

Total expenditure- General government expenditure on health as a percentage of total government expenditure (%)

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 0-1)

Schooling- Number of years of Schooling (years).

Adult Mortality - Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)

Alcohol- Alcohol, recorded per capita (15+) consumption (in liters of pure alcohol).

Immunizations- Data around Polio, Measles, Hepatitis B and Diphtheria.

HIV/AIDS- Deaths per 1 000 live births HIV/AIDS (0-4 years).

thinness 1-19/ 5-9 years - Prevalence of thinness among children and adolescents for Age 10 to 19 (%) and 5-9 years as well.

BMI: Average Body Mass Index of entire population

infant deaths: Number of Infant Deaths per 1000 population

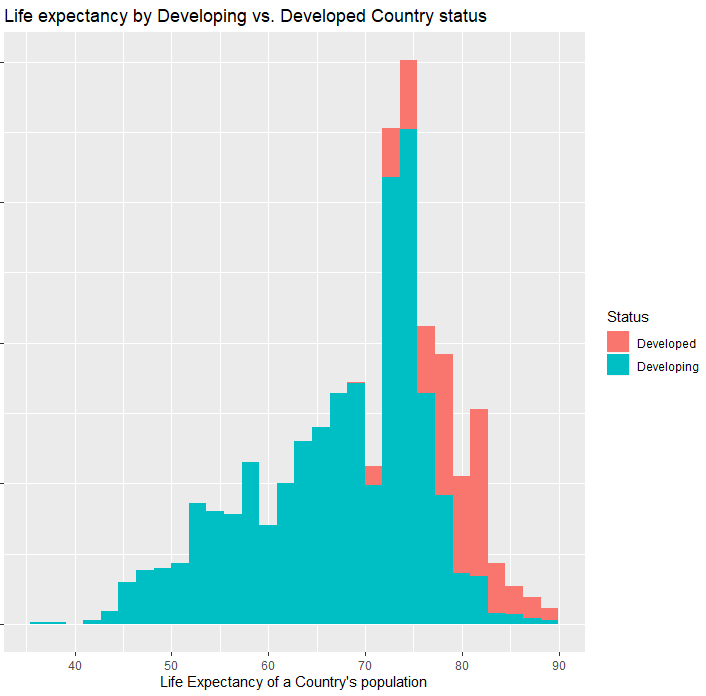
under-five deaths: Number of under-five deaths per 1000 population

**3 - Exploratory Data Analysis:**

**3.1** - Analyzing the categorical variables: Status and Country

We have data for 193 countries, out of which 161 are categorized (column Status) as ‘developing’ and 32 are categorized as ‘developed’. Looking at Life Expectancy variation by ‘Status’, we see that there is a clear distinction between the Life expectancy of country’s population by its ‘Status’.

**Figure 1**



Country is just the country’s name, so we will not be analyzing this column any further.

**3.2 - Correlation:** To study the relationships among different variables, we created scatterplots, please refer to pages 11-13 for scatter plots showing the correlation among all the variables.

Following variables have a significant correlation with Life Expectancy: Schooling(0.7276), Income\_composition\_of\_resources(0.7211) , Adult Mortality (-0.7025), HIV\_AIDS( -0.5922), BMI(0.5420), thinness\_1\_19\_years(-0.4578), thinness\_5\_9\_years(-0.4575) ), GDP(0.4413), percentage\_expenditure(0.4096), Alcohol (0.4027), Diphtheria(0.3413 ) and Polio(0.3273).

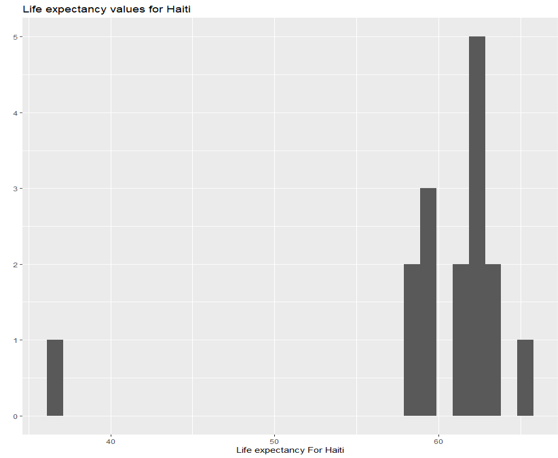
There are also multiple variables that highly correlated to each other:

Variables that are highly correlated to each other: infant\_deaths and under\_five\_deaths at 0.9969; GDP and percentage\_expenditure at 0.9593; thinness\_5\_9\_years and thinness\_1\_19\_years at 0.9279; Schooling and Income composition of resources at 0.7847.

Please refer to page 14 for correlation matrix of all the variables.

**3.3 - Missing Data:** Out of 2938 rows, we see that there are 10 rows missing Life Expectancy. There are several other variables that are missing values over multiple rows, we will handle the missing values for necessary variables as we build the models. We also notice that there are several variables that have value ‘0’ for them. This does not look right, but we will be leaving those values as is for now.

**3.4 - Outliers**: We notice a few countries where the Life expectancy values is either too high or low. For ex. obs. 1128 Haiti (2010) .



1. **- Objective 1**: Build and compare data models to predict life expectancy of a country’s population.

**4.1** - Simple and highly interpretable model:

**Model** **1**: Building model 1 using variables with correlation to Life Expectancy. Running GLM, variable Alcohol proves to be statistically insignificant. Thus, it was removed from the model.

Life\_Expectancy = + + + + + + + +

**Model 2**: Model 1 is a simple and highly interpretable model but looking at the residual plots (please refer to page 24), it can be improved by adding log transforms to Life expectancy, GDP and HIV\_AIDS. After adding log transforms, BMI became statistically insignificant, so it was removed from the model. Please refer to page 25 for Model 2 residual plots.

Log\_Life\_Expectancy = + + + + + + + +

Comparing the PRESS Statistic (CV) and R-sq for both the models.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **PRESS STATISTC (CV)** | **R-sq** | **Adj R-sq** |
| Model 1 | 41012.174 | 0.8218 | 0.8212 |
| Model 2 | 8.48 | 0.8454 | 0.8449 |

Fitting model 2: As Model 2 has significantly lower PRESS Statistic, Model 2 is considered better.

Mean(Log\_Life\_Expectancy|=

3.972903921-0.000230961\*Adult\_Mortality -0.041370168log\_HIV\_AIDS - 0.002062371\* thinness\_1\_19\_years **+** 0.007546849\*log\_GDP **+** 0.000397376\*Diphtheria **+** 0.000213914\*Polio **+** 0.012198913\*Schooling

* 1. - **Interpretation**: According to Model 2, Life\_Expectancy of a country’s population is related to the following variables in below described ways:

Though, it is not practical, but if all the factors considered on this model are 0, Life Expectancy of a country’s population will e3.972903921 = 53.13.

Adult\_Mortality: Keeping all other factors constant, it is estimated that, a one unit increase in Adult\_Mortality is associated with a ( = 0.999769066) 0.023% decrease in median Life\_Expectancy. 95% confidence interval for this decrease is between ( = 0.999793) 0.021% and ( = 0.999744679) 0.026%.

HIV\_AIDS: Keeping all other factors constant, it is estimated that, a doubling of HIV\_AIDS is associated with a =0.959473898) 4.053% decrease in median Life\_Expectancy. 95% confidence interval for this decrease is between ( = 0.961453) 3.855% and ( = 0.957498707) 4.25%.

thinness\_1\_19\_years: Keeping all other factors constant, it is estimated that, a one unit increase in thinness\_1\_19\_years is associated with a ( = 0.997939754) 0.206% decrease in median Life\_Expectancy. 95% confidence interval for this decrease is between ( = 0.99853507) 0.146% and ( = 0.997344794) 0.266%.

GDP: Keeping all other factors constant, it is estimated that, keeping all other factors constant, a doubling of GDP is associated with a =1.007575398) 0.758% increase in median Life\_Expectancy. 95% confidence interval for this decrease is between ( = 1.005974879) 0.597% and ( = 1.009178464) 0.918%.

Diphtheria: Keeping all other factors constant, it is estimated that, Keeping all other factors constant, a one unit increase in Diphtheria is associated with a ( = 1.000397455) 0.04 % increase in median Life\_Expectancy. 95% confidence interval for this decrease is between ( = 1.000258204) 0.026% and ( = 1.000536725) 0.054%.

Polio: Keeping all other factors constant, it is estimated that, Keeping all other factors constant, a one unit increase in Polio is associated with a ( = 1.000213937) 0.021 % increase in median Life\_Expectancy. 95% confidence interval for this decrease is between ( = 1.000074265) 0.007% and ( = 1.000353628) 0.035%.

Schooling: Keeping all other factors constant, it is estimated that, Keeping all other factors constant, a one unit increase in Schooling is associated with a ( = 1.012273623) 1.227 % increase in median Life\_Expectancy. 95% confidence interval for this decrease is between ( = 1.011179187) 1.18% and ( = 1.013369244) 1.337%.

* 1. - **Complex model (Model3)**: After running feature selection techniques and adding interaction with the variable ‘Status’, below is the model that we have created:

Log\_Life\_Expectancy(for Countries with Status=’Developing’) = + + + + +

Log\_Life\_Expectancy(for Countries with Status=’Developed’) = + + + + + +

**Comparing models**

Comparing the PRESS Statistic (CV) and R-sq for all the different models.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **PRESS STATISTC (CV)** | **R-sq** | **Adj R-sq** |
| Model 1 | 41012.174 | 0.8218 | 0.8212 |
| Model 2 | 8.48 | 0.8454 | 0.8449 |
| Model 3 (with interaction) | 7.78 | 0.8502 | 0.8497 |

Model 3 is the preferred model for the following reasons:

1. From Figure 1, we can see that the ‘Status’ of a country strongly affects the Life expectancy of its population. Model 3 takes this into account.
2. Model 3 has better CV statistics, R-sq and Adjust R-sq.
3. Model 3 has fewer variables with log transforms, so it is easier to interpret.

**Fitting model 3:**

Mean(Log\_Life\_Expectancy{for Countries with Status=’Developing’}) = 3.972351146 -0.000211583 + 0.104705191 + +

Mean(Log\_Life\_Expectancy{for Countries with Status=’Developed’}) = 3.972351146 + + 0.104705191 + + +

**Outliers:**

* Outlier Analysis using all the variables in Model3: We see obs. 1128 (Haiti-2010) has a very high cook’s D. Please see refer to Page 34 – Cook’s D : Graph 1
* Running model3 after removing obs. 1128 did not impact the model. So we can keep obs. 1128 in the data set. Please refer to Page 34 – Cook’s D Graph 2.

**5 - Objective 2:**

For our nonparametric regression, we have chosen regression tree method to build our model using life expectancy as the response variable. Below were the findings from the model:

1. We notice that HIV/AIDS, Income composition of resources, adult mortality, and BMI are the variables that regression tree method used to build the model.
2. Model had a Residual mean deviance which is the sum of square error of 12.96.
3. These variables are also important compared to the rest of the variables in terms of calculating predictive life expectancy values. (Please refer to Page 31 – Regression analysis chart)

Cross-validation:

In our cross-validation plot (page 32), we can notice that there is not much of a deviation between our predicted life expectancy values and test data set’s life expectancy values. The plot also indicates that our prediction data for age range of 60-80 is very close to the actual data. The error variance of 55-75 age group is less compare to age groups of 55 or less and age groups of 80 or more. The mean squared error for this regression is 2.12

**6- Conclusion:**

By comparing model 3 to regression tree model, we can conclude that model 3 is the better model since it has a lower ASE score of 0.00333 (compared to 2.12 from regression tree). We notice that HIV/AID, Income composition of resources, and adult mortality are important explanatory variables which significantly predict life expectancy for both the models. But Model 3 utilizes other important variables as well, such as Status, GDP, Schooling, Diptheria immunization and more. Life expectancy of a country’s population depends on numerous factors such as Immunization, mortality factors, and socioeconomic factors, and Model 3 considers all these in some ways. It has the lowest ASE as well. Thus, we believe Model 3 is a better model among all these.

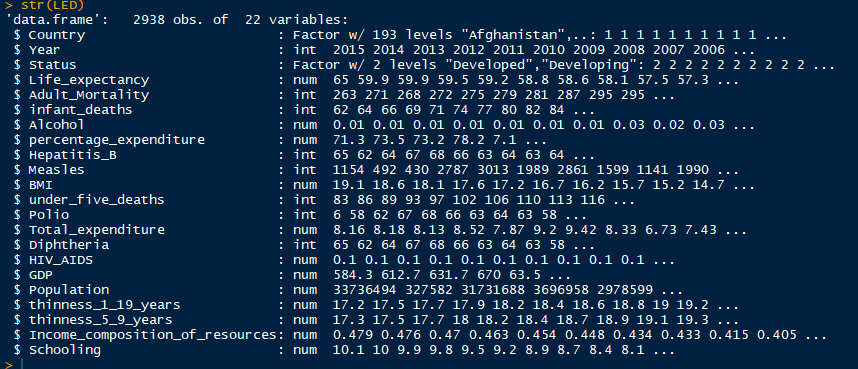
**7 – If we had more time, we could have focused more on the following:**

1. **Data Cleanse** –
   1. - We used SAS for objective 1 where SAS automatically handled(skipped) null/problematic records. If we had more time, we could have replaced missing values by more appropriate values and could have used the complete data set. Most of the time SAS skipped around 600 problematic rows. Thus, considering around 80% of the total data for modelling. Cleanse might have helped us consider more data.
   2. - For objective 2, we used R, so we had to explicitly handle missing values, we replaced the missing values by mean values for the important variables. As mean is not always the best solution, we could have spent some more time finding more appropriate replacement for the missing values.
2. Analyze other non-parametric models: For objective 2, we researched the following techniques: KNN, regression tree, Spline. We got stuck at predicting continuous variables using KNN, if we had more time we could have made that work. We could also use a better understanding on regression trees.

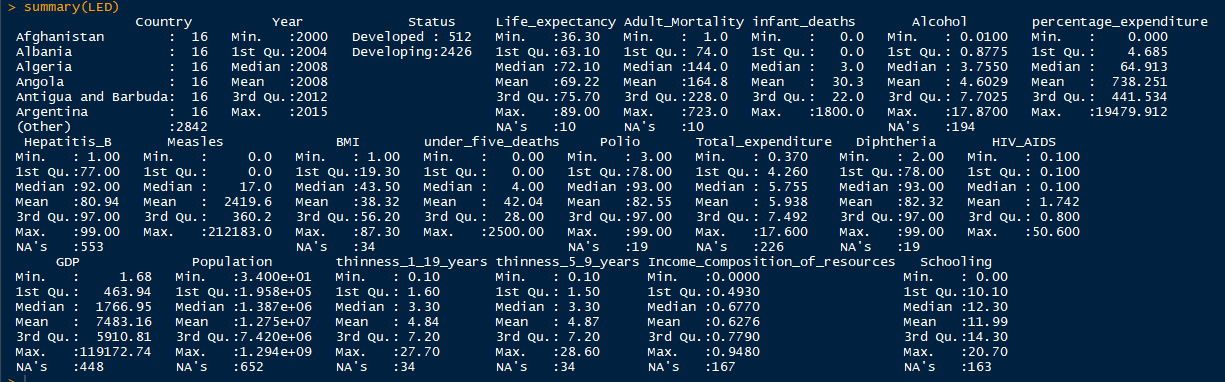
**Appendix:**

LED <-read.csv(file.choose())

Reading the structure of the data



Reading the summary of the data



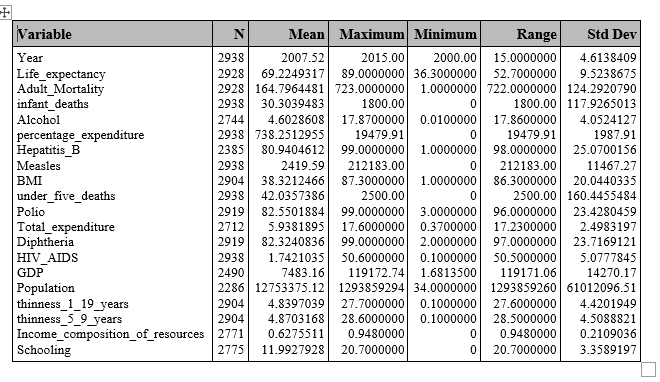
Some similar analysis in SAS

**proc** **means** data=LED n mean max min range std;

output out=meansout mean=mean std=std;

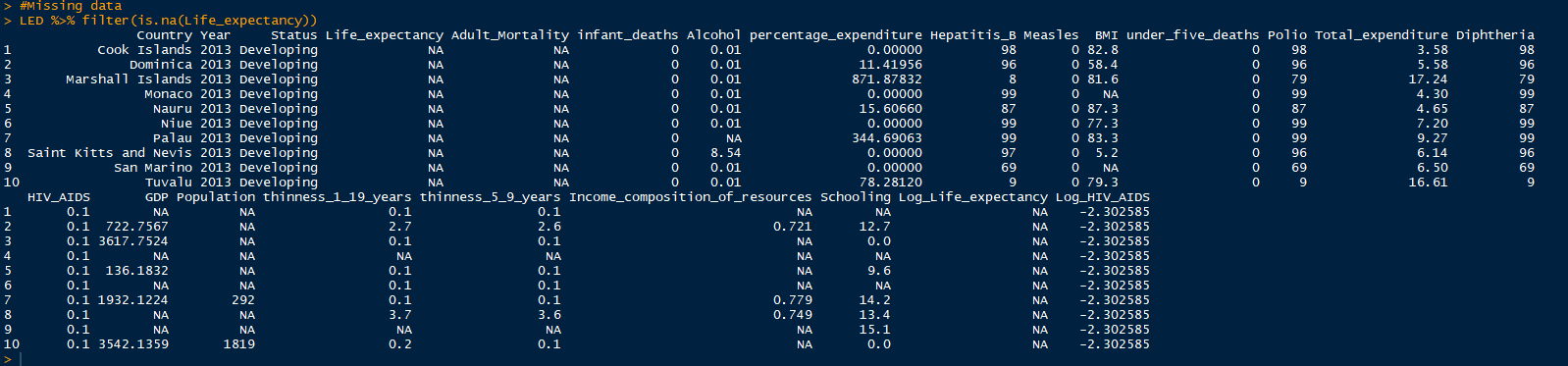
title 'Summary of LifeExpectancyData';

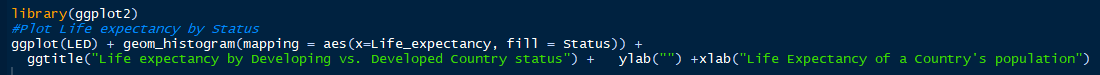
**run**;

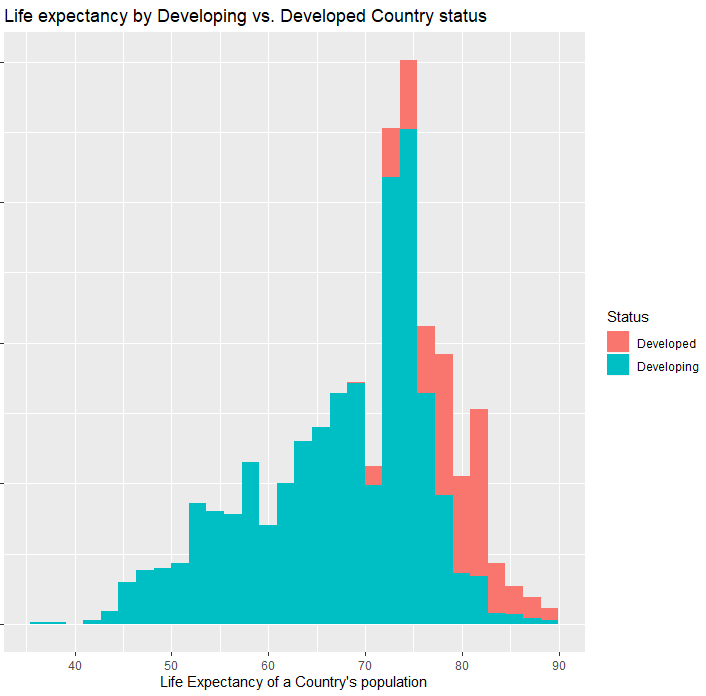


#Missing data

LED %>% filter(is.na(Life\_expectancy))

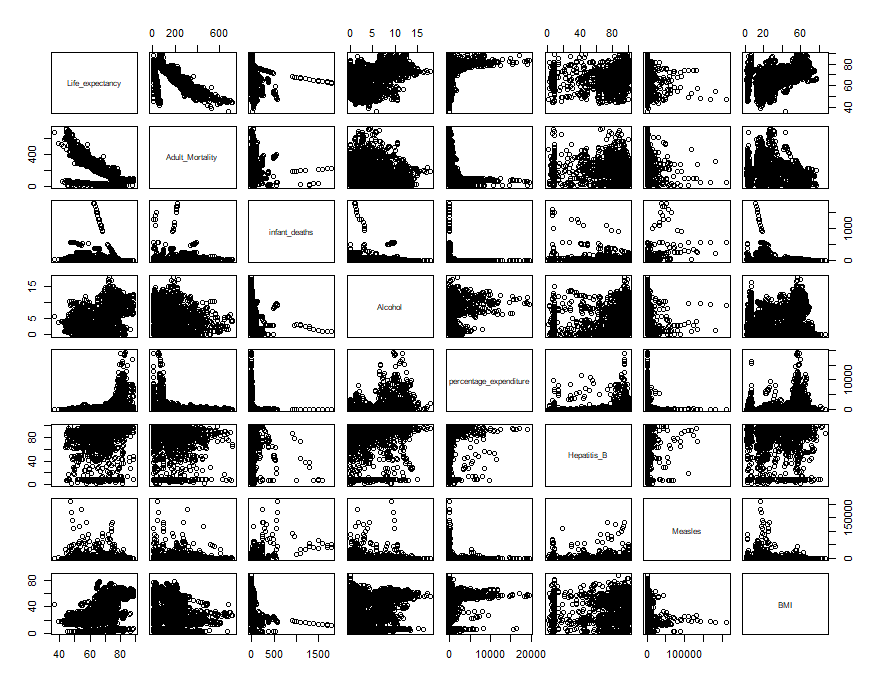




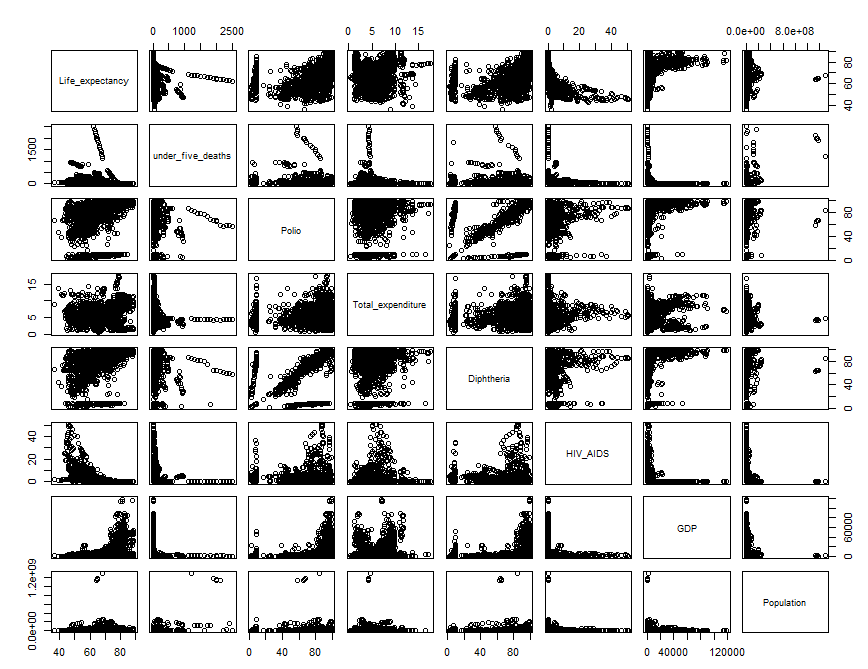


Scatter Plots:

plot(LED[,c(4,5,6,7,8,9,10,11)])



plot(LED[,c(4,12,13,14,15,16,17,18)])



plot(LED[,c(4,19,20,21,22)])

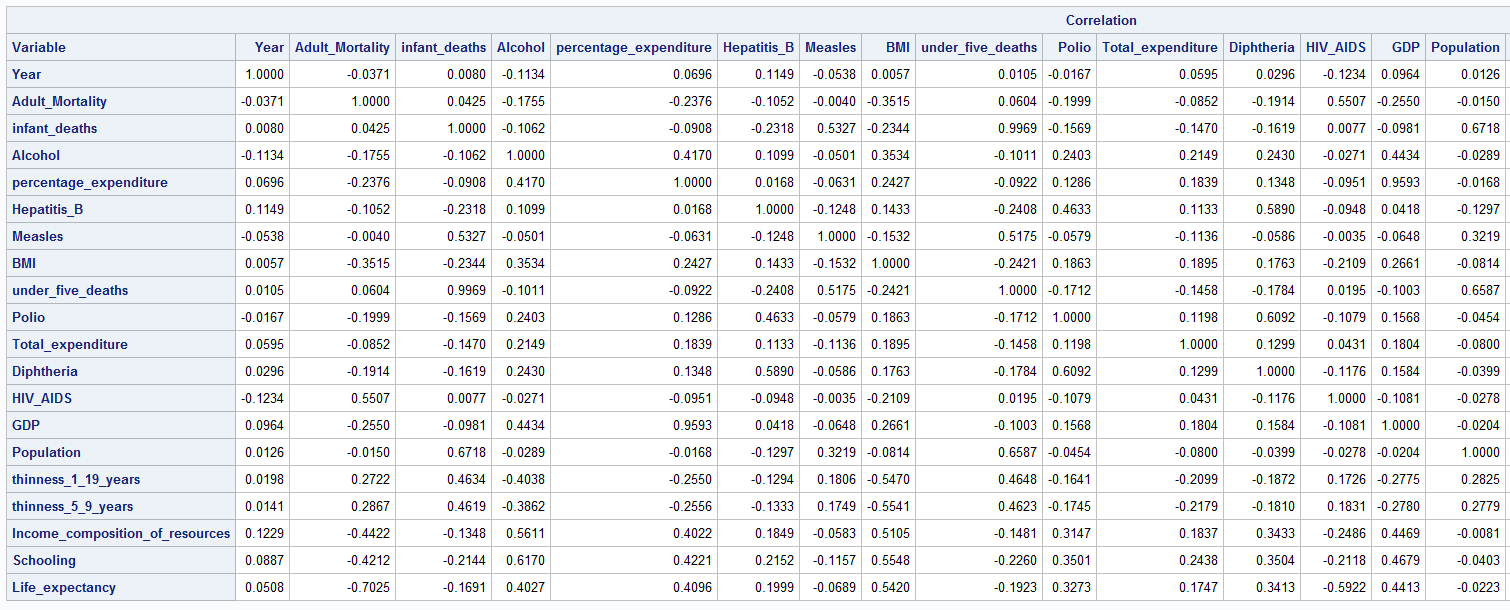


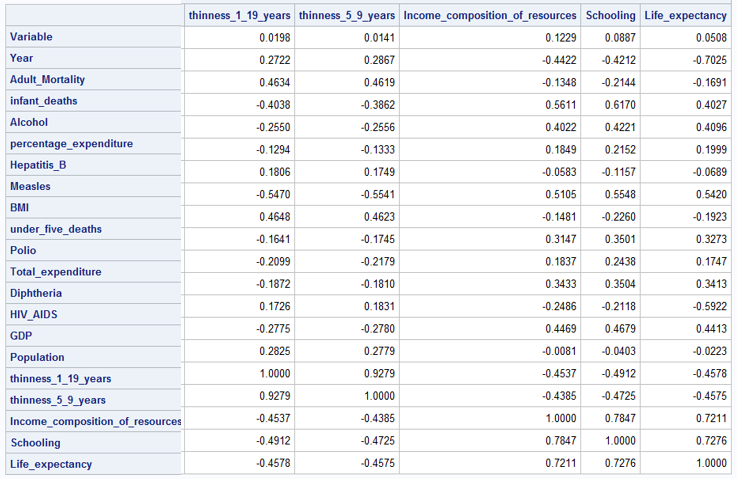
**Correlation:**

**proc** **reg** data=LED corr plots(label) = (rstudentleverage cooksd);

model Life\_expectancy= Year Adult\_Mortality infant\_deaths Alcohol percentage\_expenditure Hepatitis\_B Measles BMI under\_five\_deaths Polio Total\_expenditure Diphtheria HIV\_AIDS GDP Population thinness\_1\_19\_years thinness\_5\_9\_years Income\_composition\_of\_resources Schooling / VIF;

**run**; **quit**;





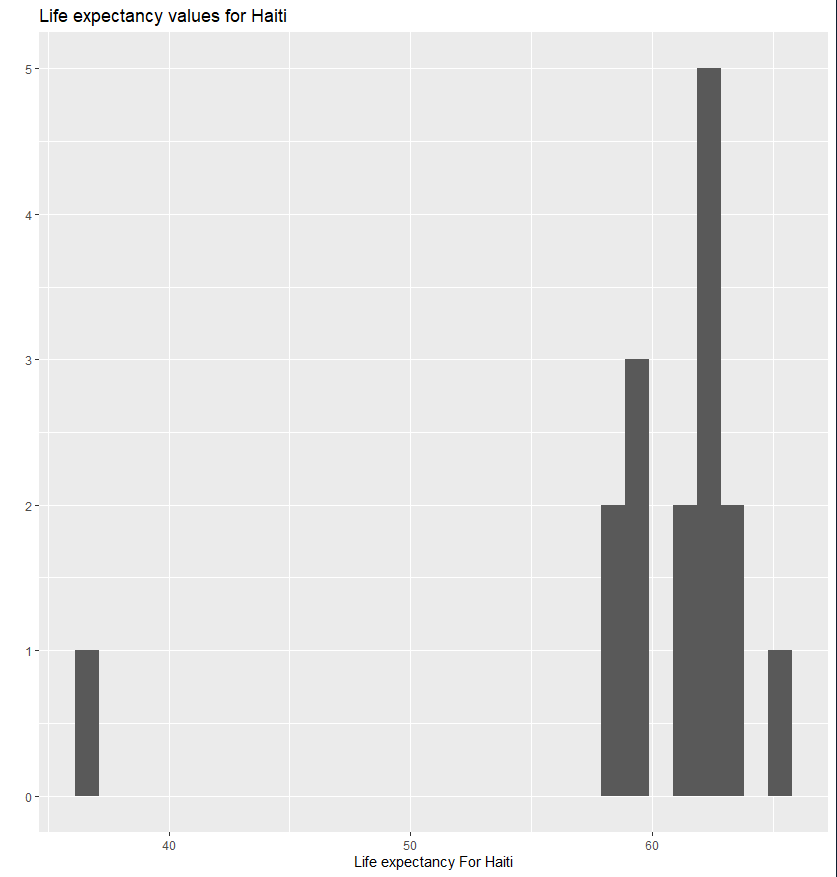
Outliers:

#Missing data

LED\_Haiti <-LED %>% filter((Country == 'Haiti'))

ggplot(LED\_Haiti) + geom\_histogram(mapping = aes(x= Life\_expectancy, fill = Year)) +

ggtitle("Life expectancy values for Haiti") + ylab("") +xlab("Life expectancy For Haiti")

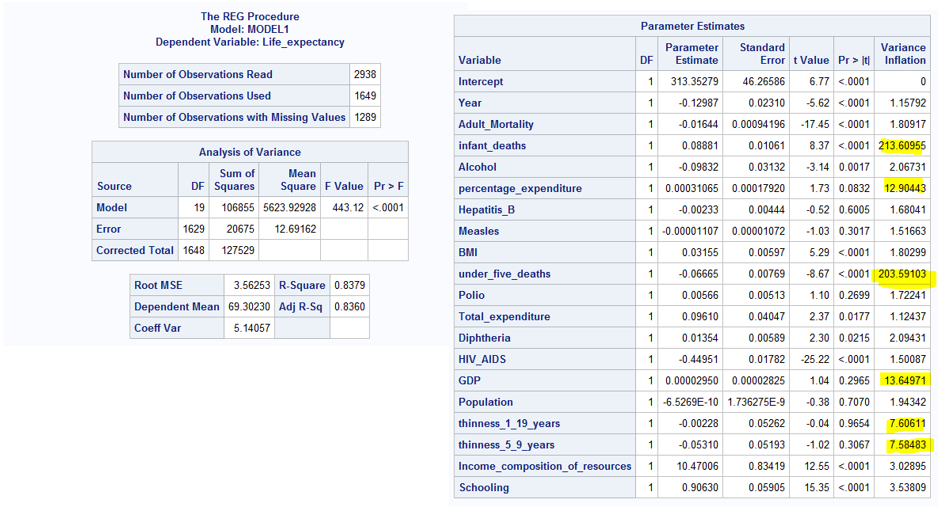


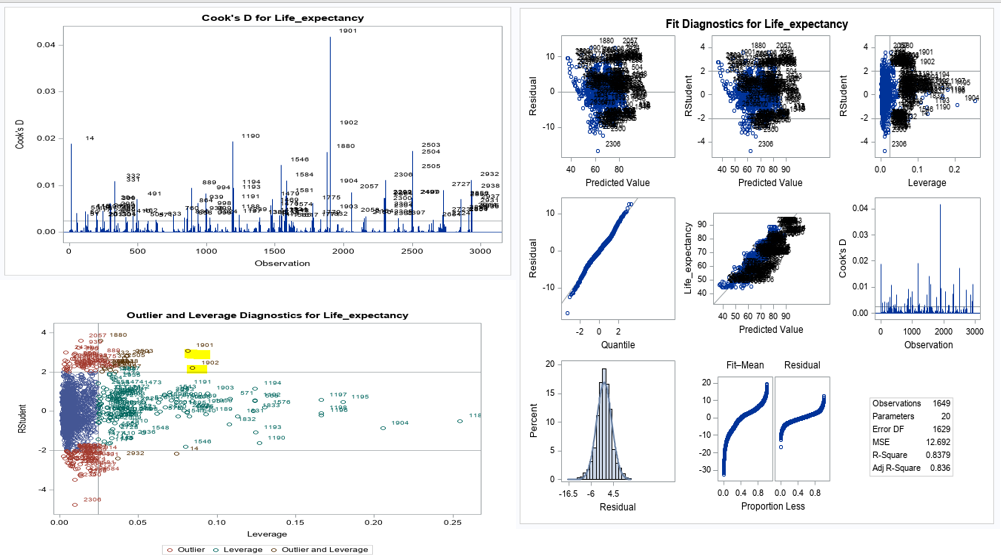
Models:

**proc** **reg** data=LED corr plots(label) = (rstudentleverage cooksd);

model Life\_expectancy= Year Adult\_Mortality infant\_deaths Alcohol percentage\_expenditure Hepatitis\_B Measles BMI under\_five\_deaths Polio Total\_expenditure Diphtheria HIV\_AIDS GDP Population thinness\_1\_19\_years thinness\_5\_9\_years Income\_composition\_of\_resources Schooling / VIF;

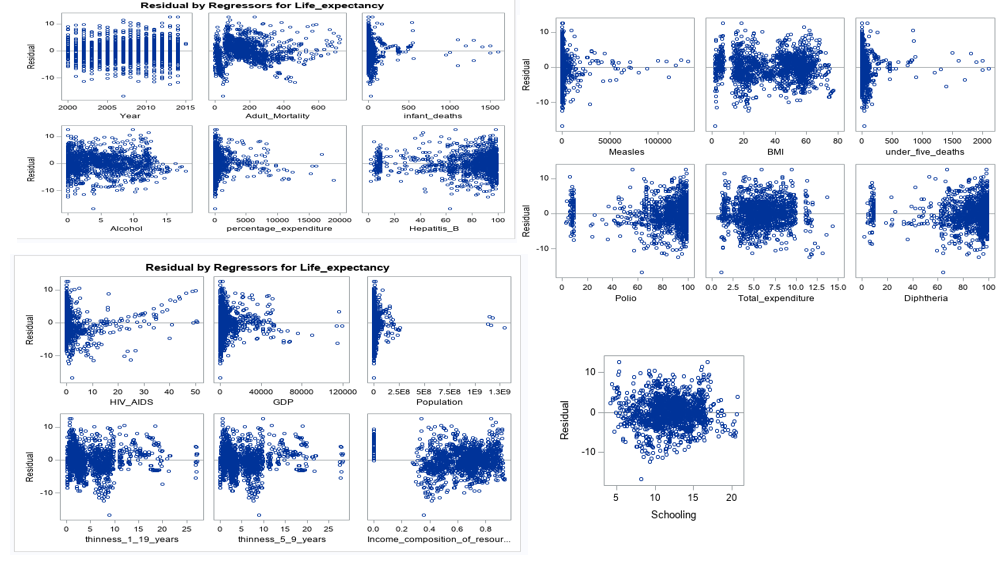
**run**; **quit**;





Graphical user interface, chart, scatter chart

Description automatically generated



\*Important observations;

\*1) infant\_deaths and under\_five\_deaths are highly correlated at 0.9969. Take infant\_deaths out;

\*2) GDP and percentage\_expenditure are highly correlated at 0.9593. Take GDP out ;

\*3) thinness\_5\_9\_years and thinness\_1\_19\_years 0.9279. Take thinness\_5\_9\_years out;

\*4) What do we do about under\_five\_deaths and Diphtheria? From VIFs, it looks like there's a correlation among these, but looking at the correlation factors above, we see it is -0.1784. So, we can keep them both in the model;

\* Take infant\_deaths,percentage\_expenditure and GDP out and check VIFs;

**proc** **reg** data=LED corr plots(label) = (rstudentleverage cooksd);

model Life\_expectancy= Year Adult\_Mortality Alcohol percentage\_expenditure Hepatitis\_B Measles BMI under\_five\_deaths Polio Total\_expenditure Diphtheria HIV\_AIDS Population thinness\_1\_19\_years Income\_composition\_of\_resources Schooling / VIF;

**run**; **quit**;

\* Looking at the residual plots, it feels necessary to do some transformations.;

**data** LED;

SET LED;

log\_Life\_expectancy = log(Life\_expectancy);

log\_percentage\_expenditure = log(percentage\_expenditure);

log\_Adult\_Mortality = log(Adult\_Mortality);

log\_Hepatitis\_B = log(Hepatitis\_B);

log\_Measles = log(Measles);

log\_BMI = log(BMI);

log\_under\_five\_deaths = log(under\_five\_deaths);

log\_Polio = log(Polio);

log\_Total\_expenditure = log(Total\_expenditure);

log\_Diphtheria = log(Diphtheria);

log\_HIV\_AIDS = log(HIV\_AIDS);

log\_Population = log(Population);

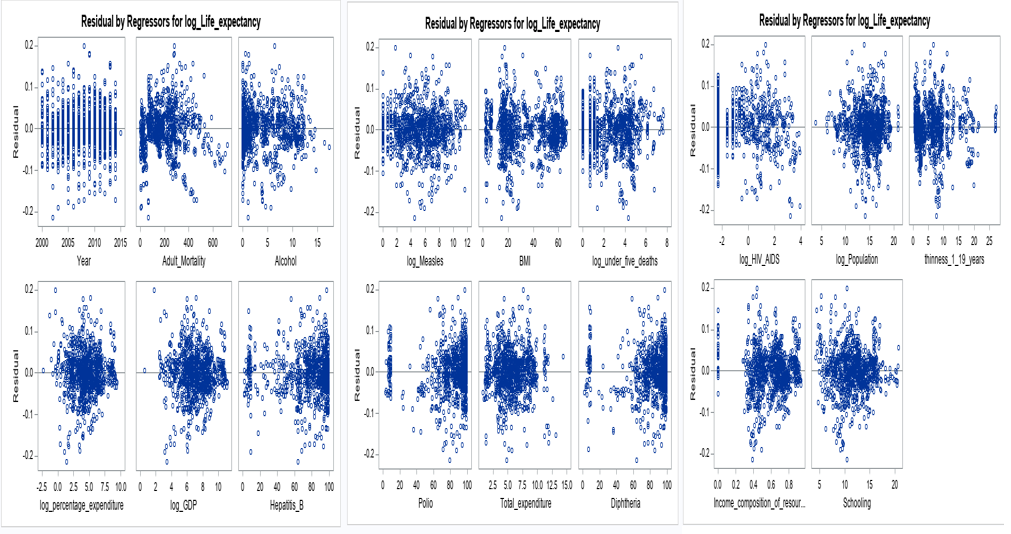
log\_GDP = log(GDP);

**run**;

**proc** **reg** data=LED corr plots(label) = (rstudentleverage cooksd);

model log\_Life\_expectancy= Year Adult\_Mortality Alcohol log\_percentage\_expenditure log\_GDP Hepatitis\_B log\_Measles BMI log\_under\_five\_deaths Polio Total\_expenditure Diphtheria log\_HIV\_AIDS log\_Population thinness\_1\_19\_years Income\_composition\_of\_resources Schooling / VIF;

**run**; **quit**;



Model 1: Using variables with high correlation to Life\_Expectancy

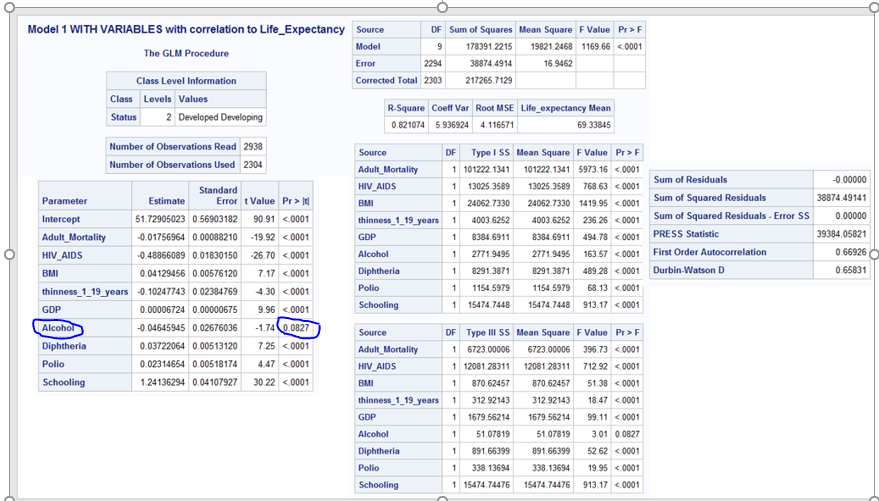
**PROC** **GLM** DATA= LED plots = all;

title "Model 1 WITH VARIABLES with correlation to Life\_Expectancy ";

class Status;

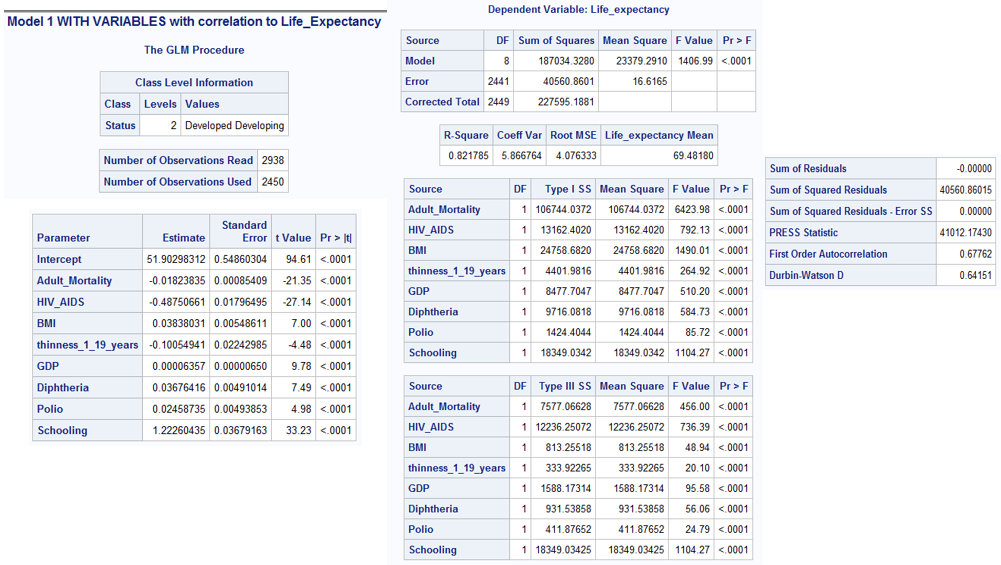
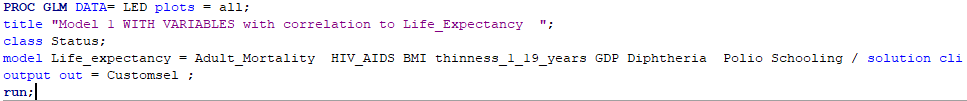
model Life\_expectancy = Adult\_Mortality HIV\_AIDS BMI thinness\_1\_19\_years GDP Alcohol Diphtheria Polio Schooling / solution ;

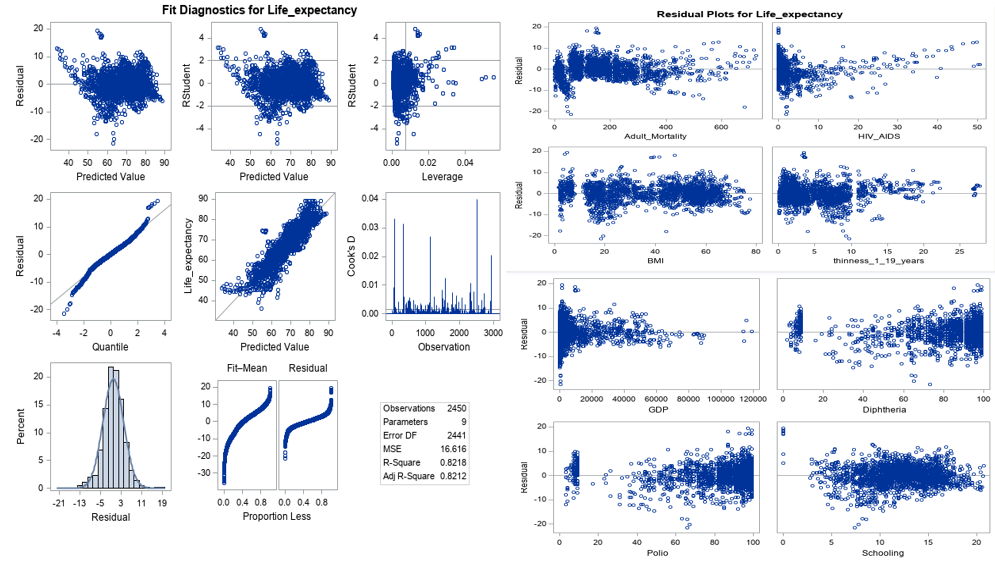
output out = Customsel ;



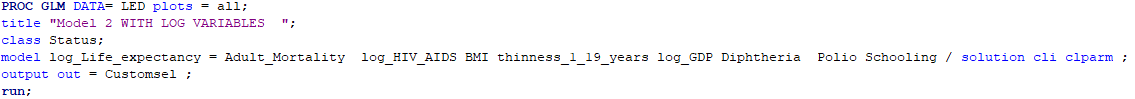
Variable Alcohol being statistically insignificant, removing it from the model.

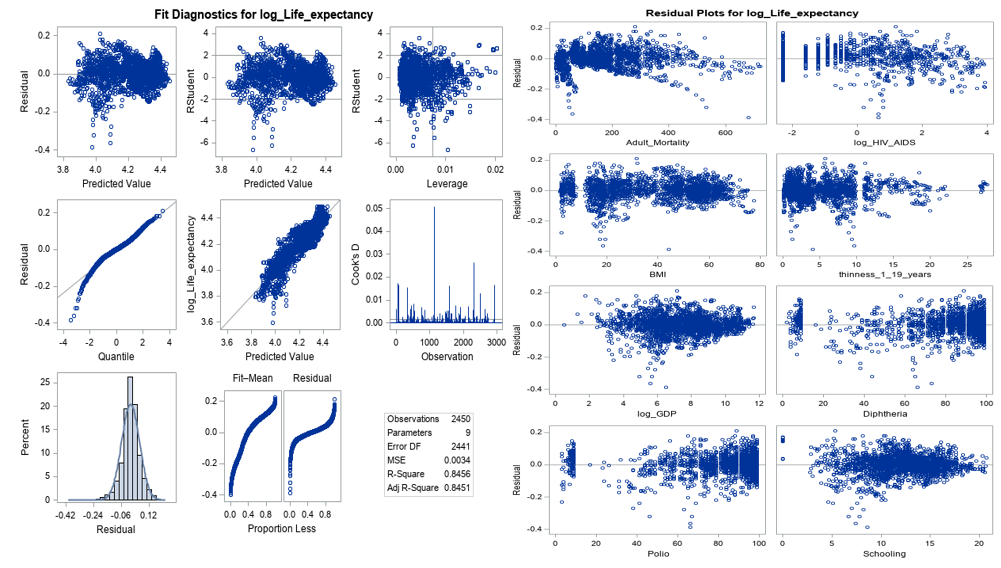
**Revised Model1**





**Model2**: After taking log on Life\_Expectancy, HIV\_AIDS and GDP.

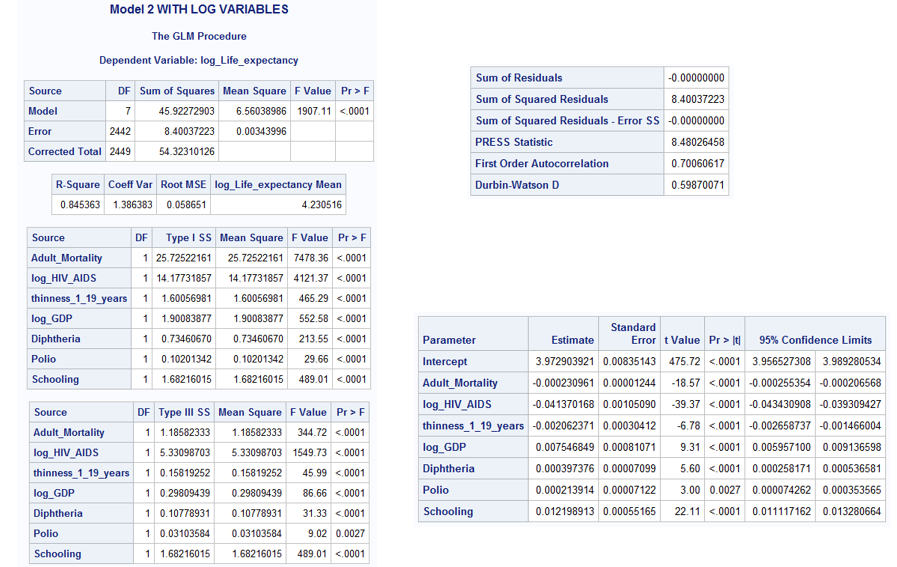


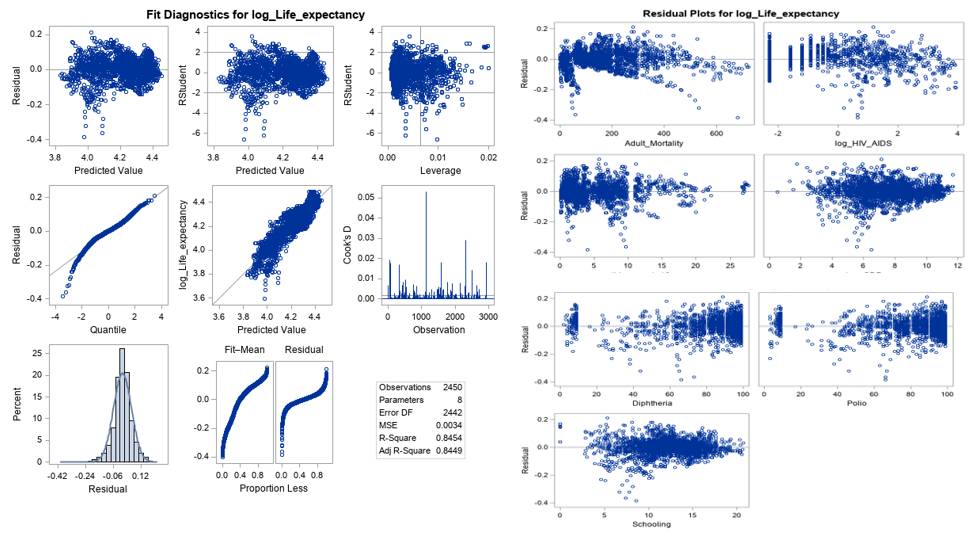




**Revised Model 2:** After taking log transform, BMI becomes statistically insignificant. Revising the model to remove BMI as a model parameter.







**Finding a complex model: Run models using different feature selection techniques (primarily CV)**

\* Model assesment using ASE;

\*Run different models using partition fraction(test = .7) and ASEPlots ;

**PROC** **GLMSELECT** DATA= LED plots(stepaxis = number) = (criterionpanel ASEPlot) seed = **1**;

title "**Forward** selection using CV Press as stop criteria with ASE";

partition fraction(test = **.5**);

class Status;

model log\_Life\_expectancy= Year Adult\_Mortality Alcohol log\_percentage\_expenditure Hepatitis\_B log\_Measles BMI log\_under\_five\_deaths Polio Total\_expenditure Diphtheria log\_HIV\_AIDS log\_Population thinness\_1\_19\_years Income\_composition\_of\_resources Schooling / selection = Forward(stop=CV);

output out = Forwardsel ;

**run**;



**PROC** **GLMSELECT** DATA= LED plots(stepaxis = number) = (criterionpanel ASEPlot) seed = **1**;

title "**Backward** selection using CV Press as stop criteria with ASE";

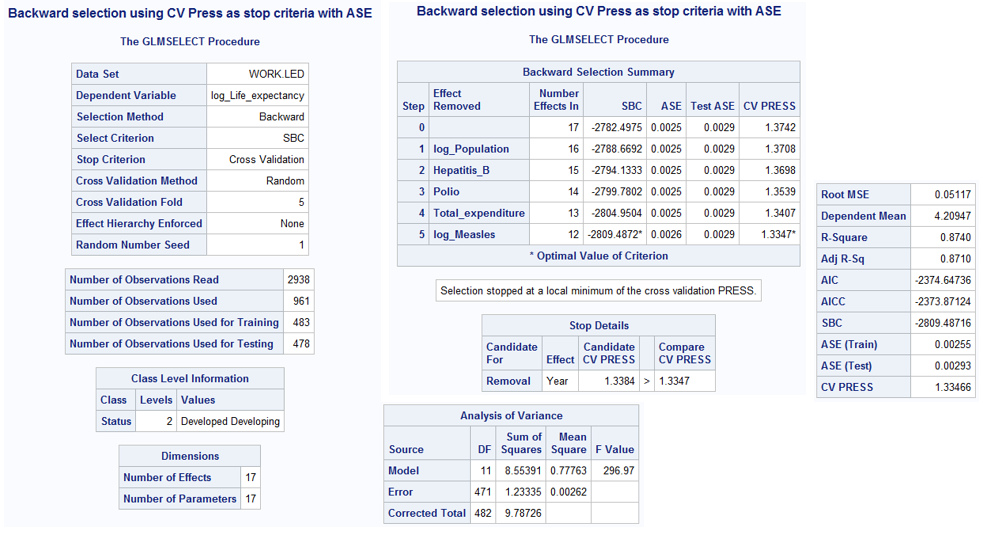
partition fraction(test = **.5**);

class Status;

model log\_Life\_expectancy= Year Adult\_Mortality Alcohol log\_percentage\_expenditure Hepatitis\_B log\_Measles BMI log\_under\_five\_deaths Polio Total\_expenditure Diphtheria log\_HIV\_AIDS log\_Population thinness\_1\_19\_years Income\_composition\_of\_resources Schooling / selection = backward(stop=CV);

output out = Backwardsel ;

**run**;



**PROC** **GLMSELECT** DATA= LED plots(stepaxis = number) = (criterionpanel ASEPlot) seed = **1**;

title "**Stepwise** selection using CV Press as stop criteria with ASE";

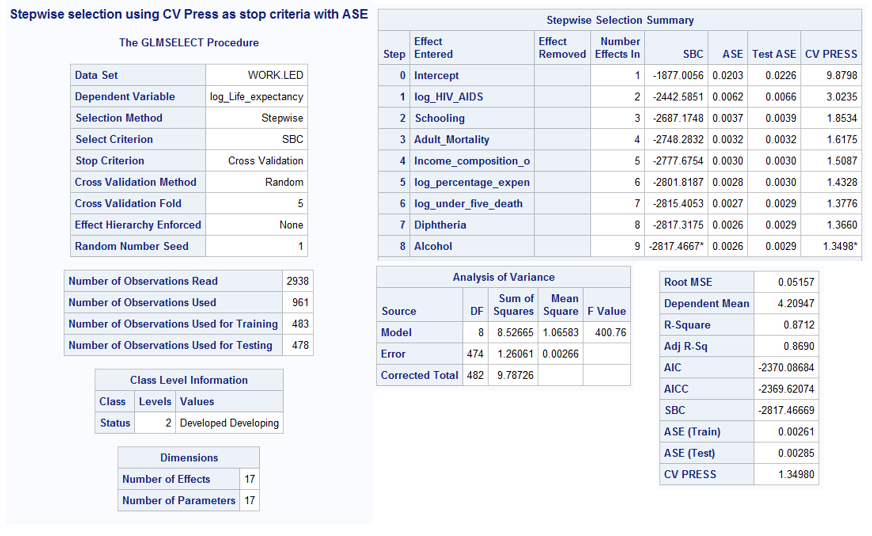
partition fraction(test = **.5**);

class Status;

model log\_Life\_expectancy= Year Adult\_Mortality Alcohol log\_percentage\_expenditure Hepatitis\_B log\_Measles BMI log\_under\_five\_deaths Polio Total\_expenditure Diphtheria log\_HIV\_AIDS log\_Population thinness\_1\_19\_years Income\_composition\_of\_resources Schooling / selection = Stepwise(stop=CV);

output out = Stepwisesel ;

**run**;



**PROC** **GLMSELECT** DATA= LED plots(stepaxis = number) = (criterionpanel ASEPlot) seed = **1**;

title "**LASSO** selection using CV Press as stop criteria with ASE";

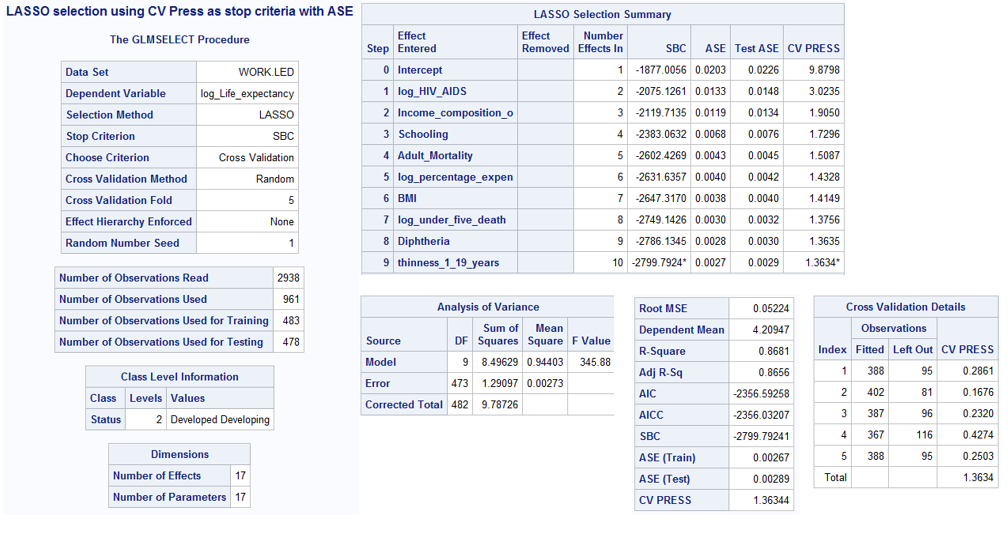
partition fraction(test = **.5**);

class Status;

model log\_Life\_expectancy= Year Adult\_Mortality Alcohol log\_percentage\_expenditure Hepatitis\_B log\_Measles BMI log\_under\_five\_deaths Polio Total\_expenditure Diphtheria log\_HIV\_AIDS log\_Population thinness\_1\_19\_years Income\_composition\_of\_resources Schooling / selection = lasso( choose = cv) CVDETAILS;

output out = LASSOsel ;

**run**;



**PROC** **GLMSELECT** DATA= LED plots(stepaxis = number) = (criterionpanel ASEPlot) seed = **1**;

title "**LARS** selection using CV Press as stop criteria with ASE";

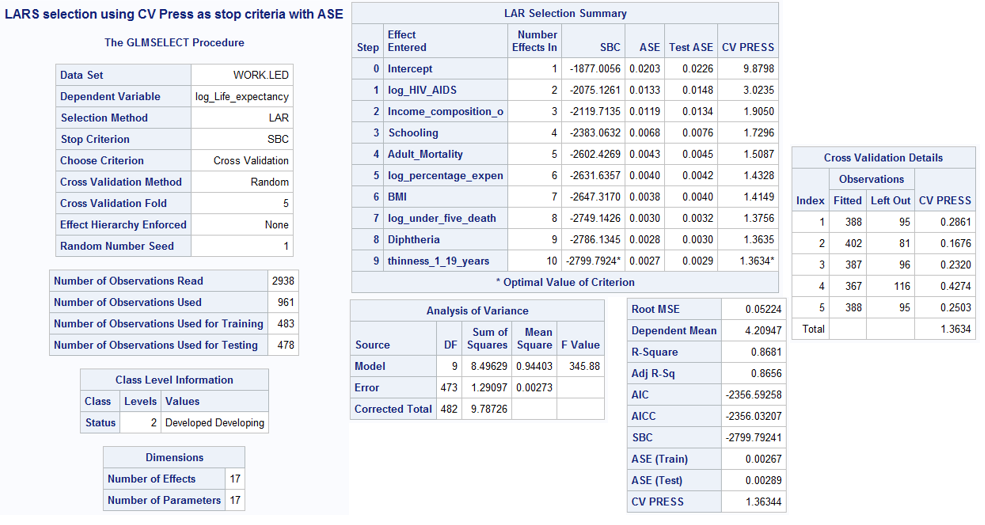
partition fraction(test = **.5**);

class Status;

model log\_Life\_expectancy= Year Adult\_Mortality Alcohol log\_percentage\_expenditure Hepatitis\_B log\_Measles BMI log\_under\_five\_deaths Polio Total\_expenditure Diphtheria log\_HIV\_AIDS log\_Population thinness\_1\_19\_years Income\_composition\_of\_resources Schooling / selection = lars( choose = cv) CVDETAILS;

output out = LARSsel ;

**run**;



Both, Forward and Backward, selection methods have the least CV PRESS and ASE -Train and Test. Looking at the variables selected by each of the models, I will go ahead with the creating a custom model using the variables selected by ‘Forward selection’.

Custom Model:

**PROC** **GLM** DATA= LED plots = all;

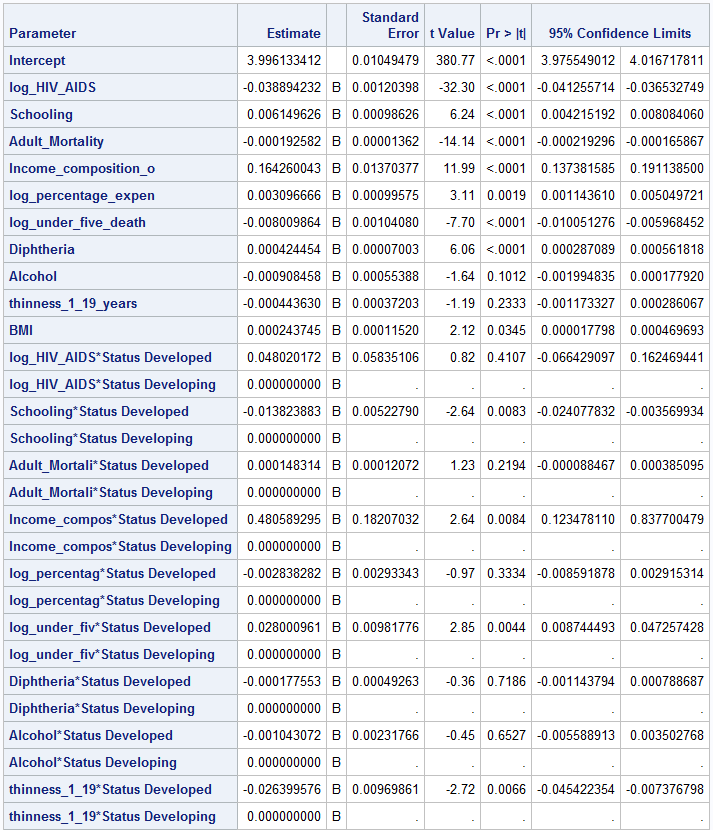
title "Custom Model with variables selected from Forward model with STATUS interaction ";

class Status;

model log\_Life\_expectancy = log\_HIV\_AIDS Schooling Adult\_Mortality Income\_composition\_of\_resources log\_percentage\_expenditure log\_under\_five\_deaths Diphtheria Alcohol thinness\_1\_19\_years BMI log\_HIV\_AIDS\*Status Schooling\*Status Adult\_Mortality\*Status Income\_composition\_of\_resources\*Status log\_percentage\_expenditure\*Status log\_under\_five\_deaths\*Status Diphtheria\*Status Alcohol\*Status thinness\_1\_19\_years\*Status BMI\*Status / solution clparm ;

output out = Customsel ;

**run**;





Removing all the statistically insignificant variables from the model:

**Revised Customer Model**

**PROC** **GLM** DATA= LED plots = all;

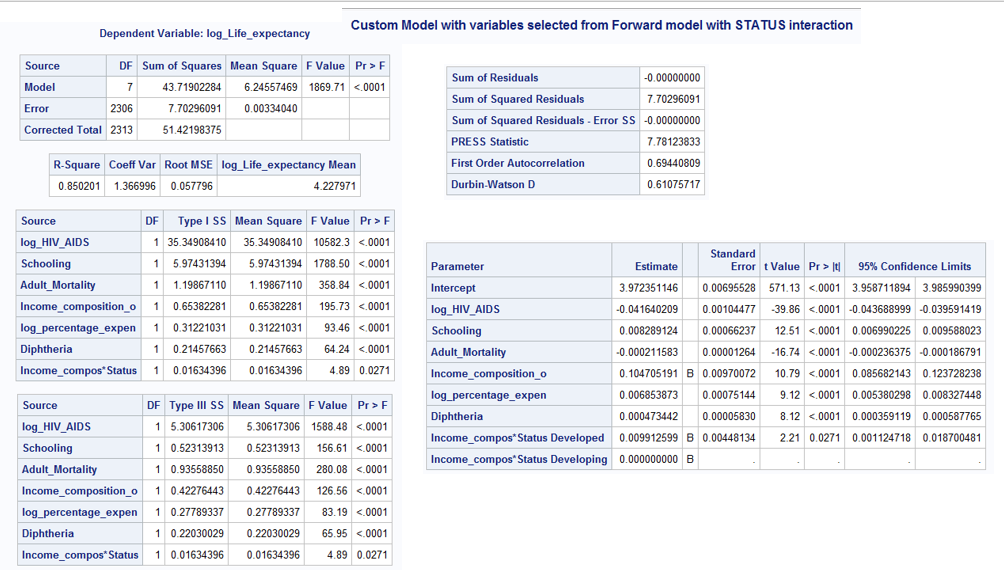
title "Custom Model with variables selected from Forward model with STATUS interaction ";

class Status;

model log\_Life\_expectancy = log\_HIV\_AIDS Schooling Adult\_Mortality Income\_composition\_of\_resources log\_percentage\_expenditure Diphtheria Income\_composition\_of\_resources\*Status / solution cli clparm ;

output out = Customsel ;

**run**;



Outlier Analysis Graph 1 and Graph 2:

Outlier Analysis using all the variables in Model3: We see obs. 1128 has a very high cook’s D.

Cook’s D - Graph 1:

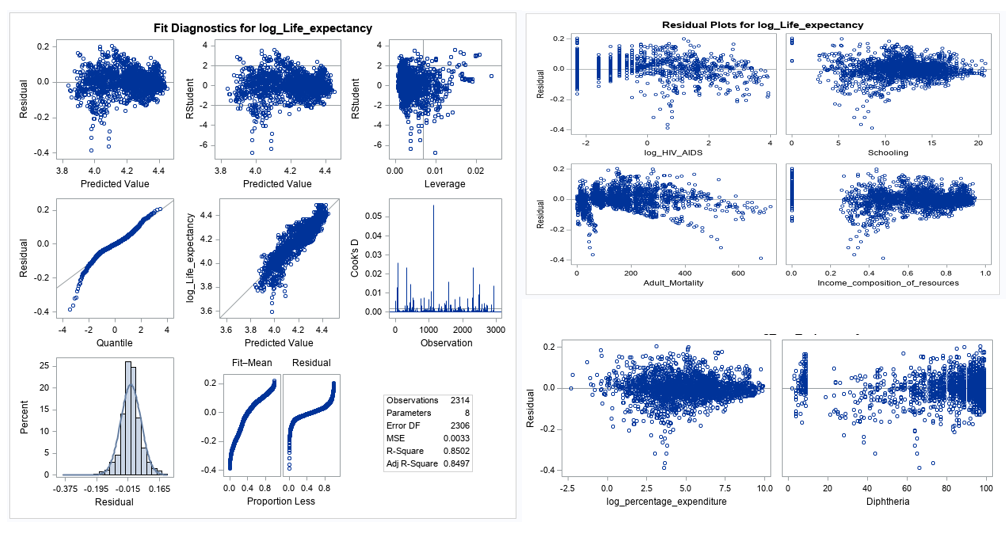
Chart, scatter chart

Description automatically generated

Cooks D Graph 2:

Graphical user interface

Description automatically generated



**Non-Parametric methods**

Regression Analysis Chart:

Text, letter

Description automatically generated

Cross Validation plot:

Chart

Description automatically generated