Prediction of Adult Mortality Rate using Multiple Linear Regression Model

Jaswinder Singh
MSc in Data Analytics, School of Computing
National College of Ireland
Dublin, Ireland
x19219997(Cohort-'B')

***Abstract*—This report aims to predict the adult mortality rate of a country by employing a multiple linear regression model. The *'Adult Mortality Rate'* which is the dependent variable is predicted by using various independent variables such as *'Human Development Index(HDI)', 'GDP', 'Life Expectancy', 'Health Expenditure', 'HIV Death Rate'* , *'Average Polio Immunity'* and *'Total Population'*. Various assumptions underlying the multiple regression like Multicollinearity, Independence of Residuals(Durbin-Watson Statistic), Constancy of Variance of Residuals, Normal Distribution of Residuals, Cook's distance, etc. were checked for each of the models. The datasets used for the analysis were taken from the United Nations(UN) database. The cleaning of data was done using the pandas library in python and the regression models were deployed in SPSS software. The outputs obtained have been clearly explained.**

***Index Terms*—Adult Mortality Rate, Multiple Linear Regression, Residuals, Variance, Multicollinearity, Durbin- Watson Statistic, Cook's Distance, Normal Distribution,**

I. Introduction

Over the last century, the advancements in medical sciences has enabled humans to live healthier and longer than ever before. Today, we are on the verge of cloning humans much like the other animals. These developments have led to the decline in mortality rates over the world. The mortality rate is defined by the UN as the no. of people that die in a population within a specified period of time. It is measured for different age groups like infants, adults, etc. The amelioration in the Data Science and Statistical techniques has opened new gates of exploration into various domains. The motivation behind this analysis is to help identify various factors influencing the adult mortality rate in a given country. This will help the organizations better understand the different aspects of the improvement in the existing healthcare system. This report makes use of the Multiple Linear Regression technique to estimate the contribution of the different factors that may affect the adult mortality rate of a country.

II. Multiple Regression- A Brief Introduction

The idea of regression is to predict one quantitaive variable by using at least one other quantitative variable. In layman terms, the basic idea of multiple linear regression is to aid us in better explaining the response(also known as dependent) variable by using various predictor(also known as independent) variables. The multiple regression model can be written mathematically as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \epsilon \qquad (1)$$

where X_n is the nth independent variable and ϵ is the error term.

Interpretation of the Coefficients: The coefficients β_n are interpreted as the amount by which the dependent variable Y changes when there is one unit change in the predictor value X_n.

In equation (1), the values of the coefficients β_n are unknown, so we estimate these coefficients and then calculate the estimated Y(i.e $\hat{Y}$) as:

$$\hat{Y} = \hat{\beta_0} + \hat{\beta_1} x_1 + \hat{\beta_2} x_2 + \hat{\beta_3} x_3 + ... + \hat{\beta_n} x_n \qquad (2)$$

These coefficents are estimated using the least square approach i.e β_0, β_1,...,β_n are chosen so that the sum of squared residuals(RSS) is minimum. RSS can be written as:

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2 \\ &= \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip} \right)^2 \end{aligned} \qquad (3)$$

While running a multiple linear regression model, there are several assumptions that need to be checked for the analysis/model to be valid. These assumptions are as follows:

- The relationship between the predictor(independent) variables and the response(dependent) variable should be ***linear***. This assumption can be checked by analysing the scatter plots between the dependent variable and each of the independent variables.
- **Multicollinearity**: The independent variables should not have high values of correlation coefficient, r with other independent variables. In other words, the independent variables must be not be highly correlated with each other.
- The value of residuals should be ***independent***. It can be checked by looking at the Durbin-Watson Statistic. The values very close to 2 indicate no or very little autocorrelation. This assumption can also be interpreted as follows:
  The individual observations should be independent of each other

- **Homoscedasticity** : The variance of resduals(errors) must be constant. This assumption can be checked by looking at the plot between standardized values predicted by our model and the standardized residuals obtained.
- The values of the residuals must be normally distributed. This can be checked by the normal probability plot of the standardized residuals.
- There should be no ***influential data points***. Influential points are the outliers that may be present in our data and can affect our regression estimates. This can be checked by looking at the ***Cook's Distance***.

III. Data Sources and Data Cleaning

The datasets for the response and the predictor variables used for the multiple linear regression analysis were extracted from the United Nations(UN) database and were cleaned, joined and pre processed using the Numpy and Pandas libraries in python. The datastets for the independent variables were first imported into separate dataframes and cleaned using pandas library. These were then joined together by the common column *Country*. The predictor variable datasets which did not contain the data for a particular country were removed from the data frame. The datasets contained the information for various years. It was filtered to only include the rows for the chosen year i.e 2010. The null values were then removed from the merged dataframe and exported into a csv file for analysis in SPSS. The following table enlists the dependent and independent variables. Here the dependent variable is **Adult Mortality Rate**.

Variable Name	Data Type	Description
Adult Mortality Rate	Float	Probability that those who have reached age 15 will die before reaching age 60
Human Development Index(HDI)	Float	Indicator of the overall development of a country.
Gross Domestic Product(GDP)	Float	Monetary measure of the market value of all final goods and services produced in a specific time period.
Life Expectancy	Float	Measure of the average time a person is expected to live.
Health Expenditure	Float	Sum of outlays by government entities to purchase health care services and goods.
HIV Death Rate	Float	No. of death due to HIV per 100,000 population
Average Polio Immunization	Float	Average percentage of population(both males and female) who have been received the Polio Immunization.
Total Population	Integer	Total Population of a Country

Fig. 1. Description of response and predictor variables

IV. Model Assumptions and Analysis

A. Assumptions

The following variables and metrics were analysed for the assumptions discussed in the previous section.

Assumption/Model Parameters	Metric Used
Linear Relationship between the response and the predictor variables	Scatter plot between each of the independent variables and the dependent variable
Multicollinearity	Correlation Coefficient (r), Variance Inflation Factor (VIF) [for independent variables]
Independence of Residuals	Durbin Watson (DW) Statistic
Homoscedasticity	Plot between 'Predicted Standardized Residuals' and 'Obtained standardized Residuals'
Normal distribution of residuals	Normal distribution plot for standardized residuals
Influential data points and outliers	Cook's Distance
Goodness of Model	Adjusted R squared value
Significance of a coefficient	p-value of coefficients, ANOVA table(F-statistic)

Fig. 2. Assumptions for Multiple Linear Regression

B. Model Analysis

1) **Model - 1**
   - *Dependent Variable:*
     a) Adult Mortality Rate
   - *Independent Variables:*
     a) Human Development Index(HDI)
     b) Gross Domestic Product(GDP)
     c) Life Expectancy
     d) Health Expenditure
     e) HIV Death rate
     f) Average Polio Immunization
     g) Total Population
     (*Since the total population is a very large number, we take the logarithm to scale it with the other variables*)
   - *Model Statistics:*
     – **Model Equation**:
     Adult Mortality Rate = 915.328 - 173.522 (**HDI**) - 0.001 (**GDP**) - 12.445 (**Life Expectancy**) - 1.700 (**Health Expenditure**) - 0.168 (**HIV Death Rate**) - 0.032 (**Average Polio Immunization**) - 0.463 (**Log(Total population)**)
     – **Adjusted** R^2 = 0.954
     – **Pearson Correlation Values**:
     HDI : Life Expectancy = 0.886
     Adult Mortality Rate : Life Expectancy = -0.945
     All other correlation values are less than 0.8
     – **p-Value(ANOVA)** < 0.05
     – **Coefficient p-values**:
       a) Average Polio Immunization = 0.87
       b) Log(Total Population) = 0.853

c) All the other coefficients' p values are < 0.05

For this model, the adjusted R-Squared value is optimum but the variables *HDI* and *Life Expectancy* show high correlation. This is expected since life expectancy is one of the indicators in the Human Development Index(HDI). *Life Expectancy* also shows very high correlation value with the *Adult Mortality Rate*. The F statistic value is also significant as can be seen from the ANOVA table. All the coefficients except for the *Average Polio Immunization* and *Log(Total Population)* have the p-value less than 0.05. This means that the Null Hypothesis(H_0 can be rejected and hence these coefficients are significant. In the next model we take the log of the GDP, since it is a very large number and remove the variable *Life Expectancy* due to its high correlation with variable *HDI*. We use the variable *log(GDP)* instead of GDP since *GDP* is a large number.

2) **Model - 2**
- *Dependent Variable:*
  a) Adult Mortality Rate
- *Independent Variables:*
  a) Human Development Index(HDI)
  b) Log(GDP)
  c) Health Expenditure
  d) HIV Death rate
  e) Average Polio Immunization
  f) Log(Total Population)

- *Model Statistics:*
  - **Model Equation**:
    Adult Mortality Rate = 526.298 - 277.479 (**HDI**) - 12.274 (**Log(GDP)** - 1.924 (**Health Expenditure**) - 0.429 (**HIV Death Rate**) - 0.884 (**Average Polio Immunization**) - 4.869 (**Log(Total population)**)
  - **Adjusted** R^2 = 0.826
  - **Pearson Correlation Values**:
    HDI : Log(GDP) = 0.910
    All other correlation values are less than 0.8
  - **p-Value(ANOVA)** < 0.05
  - **Coefficient p-values**:
    a) Log(GDP) = 0.466
    b) Log(Total Population) = 0.318
    c) All the other coefficients' p values are < 0.05

For this model, the adjusted R-Squared value has reduced substantially as compared to Model-1. *HDI* shows very high correlation with *Log(GDP)*. Therefore it has to be removed. We can also remove the variable *Log(Total Population)* since it's coefficient is not significant.

3) **Model - 3** (Final Model)
- *Dependent Variable:*
  a) Adult Mortality Rate
- *Independent Variables:*
  a) Health Expenditure
  b) HIV Death rate
  c) Average Polio Immunization
  d) Log(GDP)

- *Model Statistics:*
  - **Model Equation**:
    Adult Mortality Rate = 572.467 - 71.021 (**Log(GDP)** - 2.349 (**Health Expenditure**) - 0.475(**HIV Death Rate**) - 1.440 (**Average Polio Immunization**)
  - **Adjusted** R^2 = 0.803
  - **Pearson Correlation Values**:
    All correlation values are less than 0.8
  - **p-Value(ANOVA)** < 0.05
  - **Coefficient p-values**:
    a) All coefficients' p values are < 0.05

For this model, although the adjusted R-squared value is not optimum, but it completely satisfies all the other assumptions put forward in section earlier(listed in Fig.2). Hence this is our final model. The normal P-P plots and the residuals plots were also checked for the model. Both the plots seem to satisfy the assumptions of *Normal distribution of residuals* and *Homoscedasticity*. The cook's distance was calculated for all the data points and the maximum value was found out to be 0.03.

V. Plots and Outputs

The following plots and outputs were obtained in SPss after running a multiple linear regression analysis.

1) **Model-1**

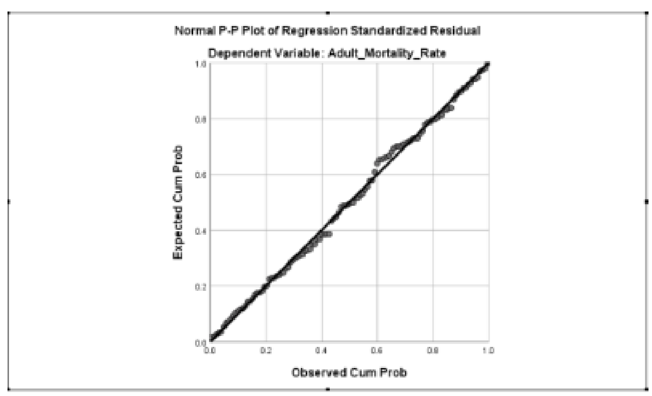

Fig. 3. Model-1: Normal PP plot

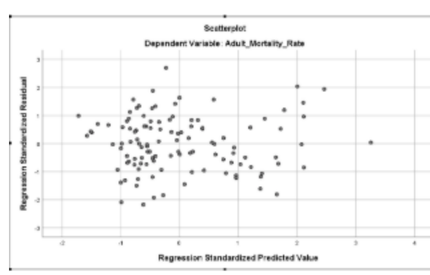

Fig. 4. Model-1: Scatter Plot

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics: R Square Change	F Change	df1	df2	Sig. F Change	Durbin-Watson
1	.978[a]	.957	.954	19.32722	.957	343.247	7	108	.000	1.960

a. Predictors: (Constant), Log_Total_Population, HIV_death_rate, Health_Expenditure, GDP, Average_Polio_Immunization, HDI, Life_Expectancy

b. Dependent Variable: Adult_Mortality_Rate

Fig. 5. Model-1: Model Summary

Correlations

		Adult_Mortality_Rate	HDI	GDP	Life_Expectancy	Health_Expenditure	HIV_death_rate	Average_Polio_Immunization	Log_Total_Population
Pearson Correlation	Adult_Mortality_Rate	1.000	-.770	-.474	-.945	-.069	.745	-.412	.078
	HDI	-.770	1.000	.602	.886	.084	-.400	.429	-.113
	GDP	-.474	.602	1.000	.499	.001	-.184	.062	-.090
	Life_Expectancy	-.945	.886	.499	1.000	.186	-.600	.477	-.104
	Health_Expenditure	-.069	.084	.001	.186	1.000	.106	.040	-.142
	HIV_death_rate	.745	-.400	-.184	-.600	.106	1.000	-.183	.042
	Average_Polio_Immunization	-.412	.429	.062	.477	.040	-.183	1.000	-.216
	Log_Total_Population	.078	-.113	-.090	-.104	-.142	.042	-.216	1.000

Fig. 6. Model-1: Pearson Correlation Matrix

Coefficients[a]

Model		Unstandardized Coefficients B	Std. Error	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance	VIF
1	(Constant)	915.328	35.586		25.722	.000		
	HDI	173.522	32.799	.272	5.290	.000	.151	6.638
	GDP	-.001	.000	-.084	-3.211	.002	.575	1.738
	Life_Expectancy	-12.445	.733	-1.033	-16.982	.000	.108	9.296
	Health_Expenditure	1.700	.500	.078	3.403	.001	.764	1.308
	HIV_death_rate	.168	.023	.211	7.179	.000	.459	2.179
	Average_Polio_Immunization	.032	.197	.004	.163	.870	.651	1.536
	Log_Total_Population	-.463	2.482	-.004	-.186	.853	.915	1.093

a. Dependent Variable: Adult_Mortality_Rate

Fig. 7. Model-1: Coefficients and their p- value

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	897517.822	7	128216.832	343.247	.000[b]
	Residual	40342.476	108	373.541		
	Total	937860.298	115			

a. Dependent Variable: Adult_Mortality_Rate

b. Predictors: (Constant), Log_Total_Population, HIV_death_rate, Health_Expenditure, GDP, Average_Polio_Immunization, HDI, Life_Expectancy

Fig. 8. Model-1: ANOVA Table

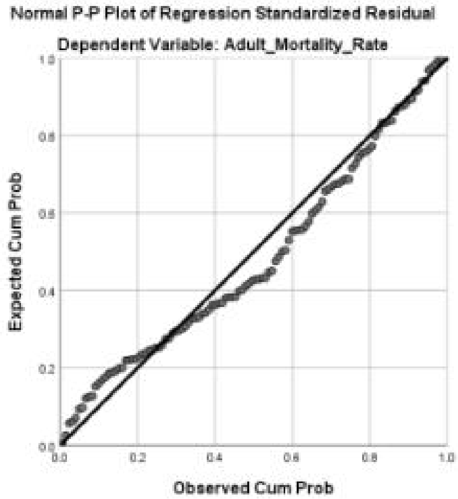

Fig. 9. Model-2: Normal PP Plot

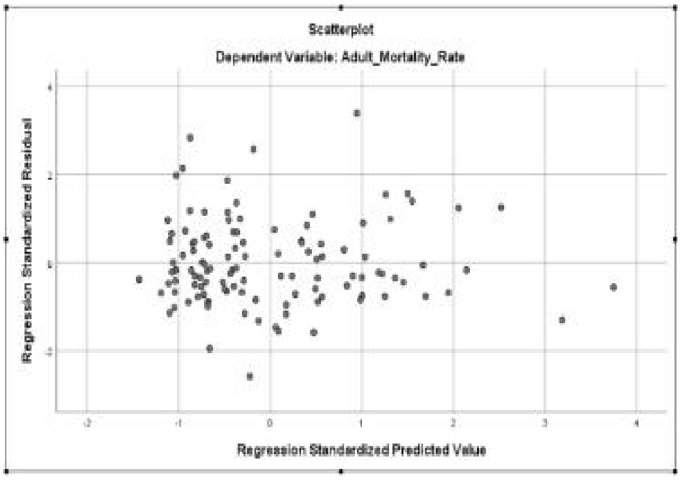

Fig. 10. Model-2: Scatter Plot

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	526.298	61.892		8.503	.000		
	Log_GDP	-12.474	17.058	-.075	-.731	.466	.142	7.027
	HDI	-277.479	72.687	-.435	-3.817	.000	.116	8.605
	Health_Expenditure	-1.924	.881	-.088	-2.184	.031	.932	1.073
	HIV_death_rate	.429	.035	.540	12.304	.000	.784	1.276
	Average_Polio_Immunization	-.884	.371	-.111	-2.381	.019	.693	1.444
	Log_Total_Population	-4.869	4.856	-.041	-1.003	.318	.906	1.104

a. Dependent Variable: Adult_Mortality_Rate

Fig. 12. Model-2 Coefficients and their p- values

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics R Square Change	F Change	df1	df2	Sig. F Change	Durbin-Watson
1	.914[a]	.836	.826	37.61983	.836	92.280	6	109	.000	1.831

a. Predictors: (Constant), Log_Total_Population, HIV_death_rate, Health_Expenditure, Average_Polio_Immunization, Log_GDP, HDI

b. Dependent Variable: Adult_Mortality_Rate

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	783597.847	6	130599.641	92.280	.000[b]
	Residual	154262.451	109	1415.252		
	Total	937860.298	115			

a. Dependent Variable: Adult_Mortality_Rate

b. Predictors: (Constant), Log_Total_Population, HIV_death_rate, Health_Expenditure, Average_Polio_Immunization, Log_GDP, HDI

Fig. 11. Model-2: ANOVA table and Model Summary

Correlations

		Adult_Mortality_Rate	Log_GDP	HDI	Health_Expenditure	HIV_death_rate	Average_Polio_Immunization	Log_Total_Population
Pearson Correlation	Adult_Mortality_Rate	1.000	-.659	-.770	-.069	.745	-.412	.078
	Log_GDP	-.659	1.000	.910	.041	-.295	.274	-.139
	HDI	-.770	.910	1.000	.084	-.400	.429	-.113
	Health_Expenditure	-.069	.041	.084	1.000	.106	.040	-.142
	HIV_death_rate	.745	-.295	-.400	.106	1.000	-.183	.042
	Average_Polio_Immunization	-.412	.274	.429	.040	-.183	1.000	-.216
	Log_Total_Population	.078	-.139	-.113	-.142	.042	-.216	1.000

Fig. 13. Model-2 Pearson Correlation Matrix

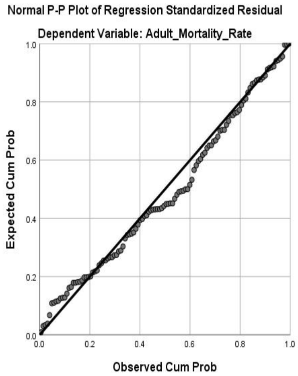

Fig. 14. Model-3: Normal PP plot

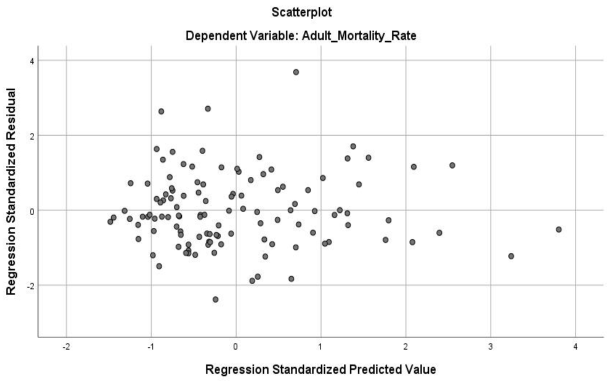

Fig.15. Model-3: Scatter Plot

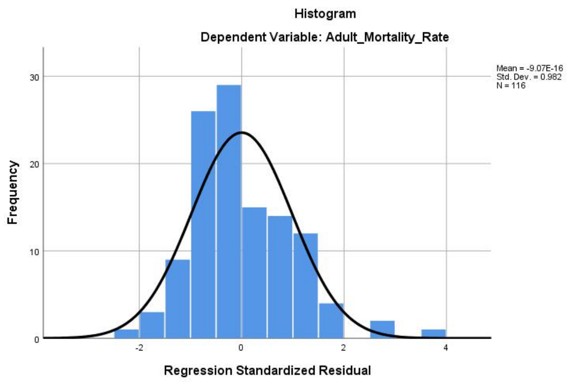

Fig. 16. Model-3: Normal Distribution for Residuals plot

Coefficient Correlations[a]

Model			Log_GDP	Health_Expenditure	Average_Polio_Immunization	HIV_death_rate
1	Correlations	Log_GDP	1.000	-.064	-.230	.264
		Health_Expenditure	-.064	1.000	-.044	-.128
		Average_Polio_Immunization	-.230	-.044	1.000	.116
		HIV_death_rate	.264	-.128	.116	1.000
	Covariances	Log_GDP	54.779	-.434	-.588	.068
		Health_Expenditure	-.434	.838	-.014	-.004
		Average_Polio_Immunization	-.588	-.014	.119	.001
		HIV_death_rate	.068	-.004	.001	.001

a. Dependent Variable: Adult_Mortality_Rate

Fig.17. Model-3: Coefficients Correlations

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	572.467	36.837		15.540	.000		
	Health_Expenditure	-2.349	.916	-.107	-2.565	.012	.981	1.019
	HIV_death_rate	.475	.035	.597	13.575	.000	.887	1.127
	Average_Polio_Immunization	-1.440	.345	-.181	-4.176	.000	.912	1.097
	Log_GDP	-71.021	7.401	-.429	-9.596	.000	.860	1.163

a. Dependent Variable: Adult_Mortality_Rate

Fig.18. Model-3: Coefficients and their p- values

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics R Square Change	F Change	df1	df2	Sig. F Change	Durbin-Watson
1	.900[a]	.810	.803	40.11829	.810	117.928	4	111	.000	1.785

a. Predictors: (Constant), Log_GDP, Health_Expenditure, Average_Polio_Immunization, HIV_death_rate

b. Dependent Variable: Adult_Mortality_Rate

Fig.19. Model-3: Model Summary

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	759208.373	4	189802.093	117.928	.000[b]
	Residual	178651.925	111	1609.477		
	Total	937860.298	115			

a. Dependent Variable: Adult_Mortality_Rate

b. Predictors: (Constant), Log_GDP, Health_Expenditure, Average_Polio_Immunization, HIV_death_rate

Fig.20. Model-3: ANOVA Table

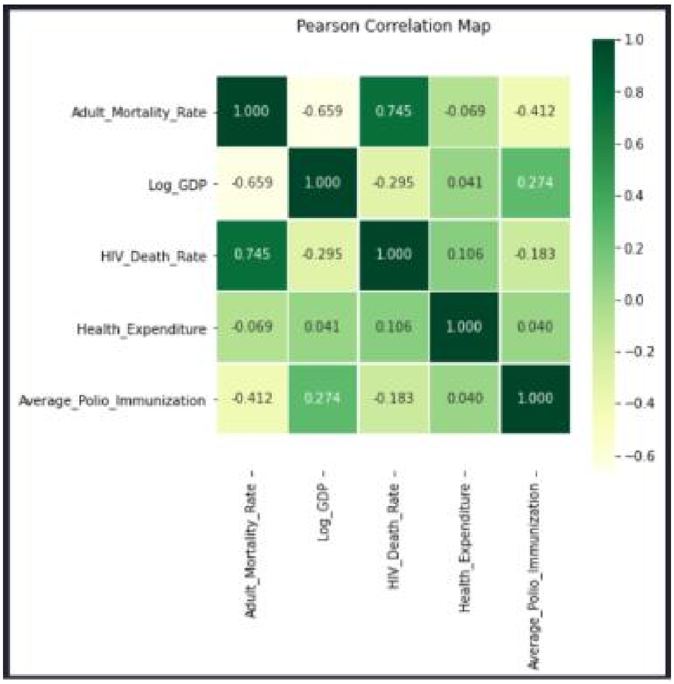

Fig.21. Model-3 :Pearson Correlation Heatmap(created using python)

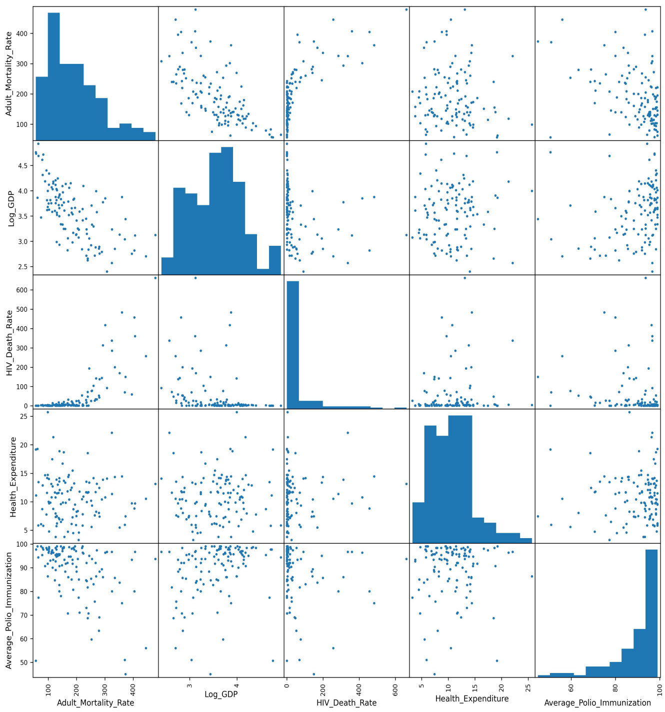

Fig.22. Model-3: Scatter plots of all independent variables with with dependent variable

CONCLUSIONS

After running various regression models in SPSS, the model satisfying all the assumptions and an optimum adjusted R-Squared value was selected to be the final model for the analysis. The final model equation is:
Adult Mortality Rate = 572.467 - 71.021 (**Log(GDP)** - 2.349 (**Health Expenditure**) - 0.475(**HIV Death Rate**) - 1.440 (**Average Polio Immunization**)

This model satisfies all the assumption put forward in section IV. The pearson correlation heatmap and the scatter plots for all the independent variables(shown in previous section) were also created using various visualization libraries in python.