Regression Analysis
on
Life Expectancy

By Aastha Sumra, Prottush Das, Kartikeya Sinha & Navya Garg
Bsc Statistics(H)

Introduction

Our data is based on life expectancy (in age) and the factors affecting it. Life expectancy is an estimate of the average period of time that a person may expect to live. It is a hypothetical measure. We have taken in consideration the following factors for testing their relation with life expectancy: 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. The variables in our data represent:

1. Adult mortality rates of both sexes are taken and can be defined as the probability of dying between 15 and 60 years per 1000 population.
2. Infant death rate measures human infant deaths in a group younger than one year of age per 1000 population.
3. Alcohol consumption is recorded per capita (15+) (in litres of pure alcohol).
4. Percentage expenditure is the expenditure on health as a percentage of Gross Domestic Product per capita(%).
5. Hepatitis B refers to the immunization coverage among 1-year-olds (%).
6. Measles depicts the number of reported cases per 1000 population.
7. BMI includes the average Body Mass Index of the entire population.
8. Under-five deaths show the number of under-five deaths per 1000 population.
9. Polio refers to the immunization coverage among 1-year-olds (%).
10. Total expenditure is the general government expenditure on health as a percentage of total government expenditure (%).
11. Diphtheria mentions the diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%).
12. HIV/AIDS gives the deaths per 1000 live births due to HIV/AIDS for 0–4-year-olds.
13. GDP is Gross Domestic Product per capita (in USD).
14. Population depicts the population of the country.
15. Thinness 1-19 years provides with the prevalence of thinness among children and adolescents for ages 10 to 19 years (%).
16. Thinness 5-9 years provides the prevalence of thinness among children for ages 5 to 9 years (%).
17. Human Development Index in terms of income composition of resources (index ranging from 0 to 1) is given at next.
18. Schooling gives the number of years of schooling.

(The data relative to the above-mentioned factors is given for various countries of different developmental levels for different years.)
Data Source: https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

Data Cleaning and Selection

We proceeded to select the data for the year 2015 only to remove the time factor. Since 97%, 99% and 99% entries were missing for alcohol, total expenditure, and percentage expenditure columns, respectively, so, we did not take these factors into consideration. After this, we dropped the rows containing at least one missing entry.

Then, we divided our data into two parts – **training data and test data**.

Training data - 115 observations, Test Data – 15 observations.
We have attached an excel file of the train data and test data after cleaning.
This division was done by exporting the data from Excel to R Studio. Using sample function, we generated a random sample of size 115 from a sequence of numbers from 1 to 130.
Then, we selected the data for countries corresponding to the sample number generated in our train data and used the rest of the data as test data.
Fitting of the model is done on the train data, and Model's performance is evaluated using the test data.

Data Analysis Part 1: Simple Linear Regression Model

SRLM (Simple Regression Linear Model) is a statistical method used to model the relationship between a dependent variable and a single independent variable. It is often used to make predictions and understand the relationship between variables.
On performing SLRM using excel between Life expectancy and each of the other 15 independent variables, only two came out to be non-significant namely **Measles and Population** *(Appendix, Simple linear regression model).*

Data Analysis Part 2: Multiple Linear Regression Model

Step 1 : Multicollinearity

We have calculated the correlation between the independent variables using the correlation function of data analysis. (*Appendix, Multicollinearity)*
We have computed correlation coefficients of various variables with each other using correlation function of Data Analysis in excel. On computing, correlation coefficient between the following variables was very high, so we retained one of the variables out of them and removed the other to obtain the required data. The description to the very high correlation between variables is as given below:

a) Infant death rate has very high correlation with Measles and under-five deaths i.e., 0.82563 and 0.99406, respectively, so we retained Measles and under-five deaths and removed infant deaths.
b) Diphtheria has very high correlation of 0.87081 with Hepatitis B, so we retained Hepatitis B and removed Diphtheria.
c) Thinness 5-9 years has very high correlation of 0.98095 with thinness 1-19 years, so we retained thinness 1-19 years and removed thinness 5-9 years.
d) Schooling has very high correlation of 0.91814 with Income composition of resources, so we retained Income composition of resources and removed schooling.

So, in all we removed four variables and retained 11 variables for the required data. We have removed infant deaths, Diphtheria, thinness 5-9 years and schooling.

Step 2: Model fitting and Hypothesis Testing

We fit our data using excel to Multiple Regression Linear Model. Next, we perform hypothesis testing on multiple linear regression (MLR) to determine the statistical significance of the independent variables in the model. Hypothesis testing allows us to assess the statistical significance of each independent variable by testing the null hypothesis that its coefficient (β) equals zero against the alternative hypothesis that it does not equal zero. We observed that 4 variables came out to be significant namely: **Adult Mortality, HIV/Aids, Hepatitis B and Income Composition of Resources** since their p value is less than 0.05*(Appendix, Model fitting and hypothesis testing, Table 1)*.

The model equation we fitted is :

$$\hat{y} = 47.76311 - 0.02063 * x_1 + 0.034 * x_2 - 0.49539 * x_3 + 35.43532 * x_4$$

Where x1= Adult Mortality

x2 = Hepatitis B

x3 = HIV/AIDS

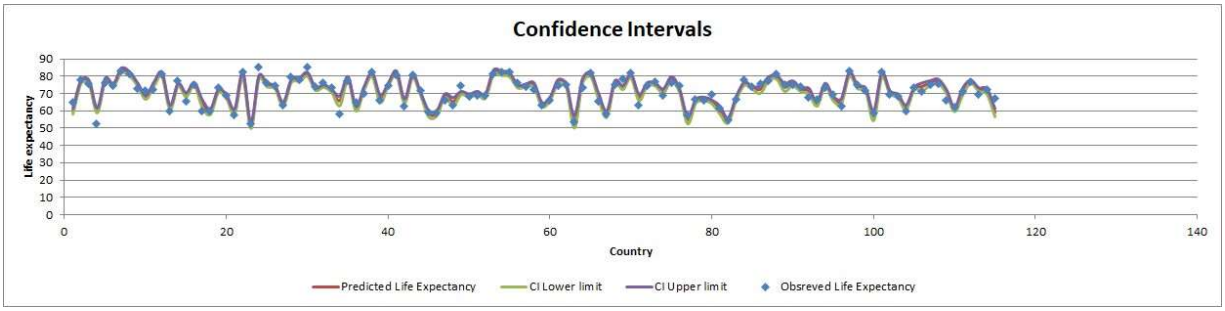
x4 = Income Composition of Resources

Also, MSE of our chosen multiple linear regression model is minimum as compared to all simple linear regression models. Hence, fitting the multiple linear regression model is a better choice.

Step 3: Confidence Intervals

Confidence Intervals on Mean Response at a point x_0

Confidence Intervals were computed for fitted values of life expectancy in 115 countries taken in the model fitting *(Appendix, Confidence Intervals, Table 1)*. Since 95% Confidence Intervals were computed, there is a 95% probability that our fitted value of life expectancy for a given set of factors will lie in the constructed interval.



Confidence Intervals on individual regression coefficients

95% Confidence Intervals for individual regression coefficients were computed as follows:-

	Lower Limit	Upper Limit	Estimated Value
Adult mortality	-0.029350686	-0.011908944	-0.020629815
Hepatitis B	0.008414742	0.059587545	0.034001143
HIV/AIDS	-1.030043965	0.039256427	-0.495393769
income composition of resources	30.70250398	40.16813915	35.43532156

(Appendix, Confidence Intervals, Table 2)

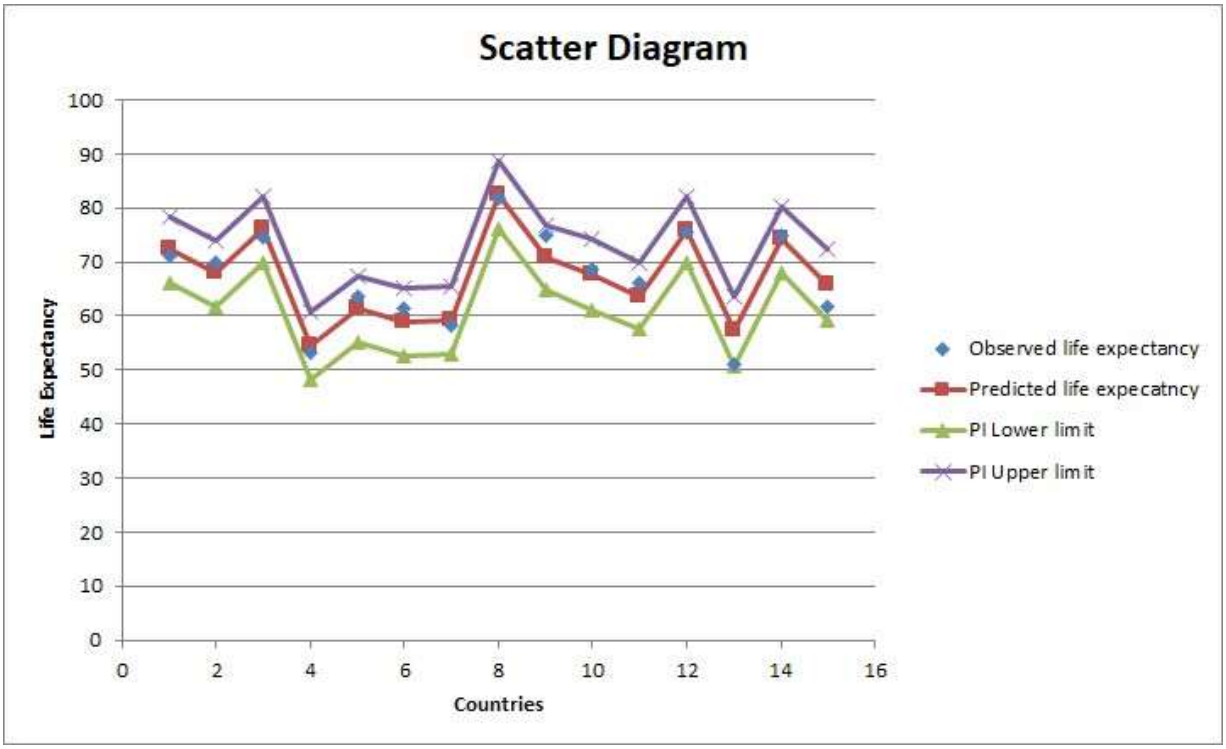
Step 4: Prediction Intervals

After fitting a suitable model to the data, we predicted the life expectancies for the 15 countries we included in our test data. We also calculated the 95% prediction intervals for our test data by which we mean that there is a 95% chance that the predicted life expectancy is going to lie in that interval.

Sno	Countries	Given life expectancy	Lower limit	Upper limit	Predicted life expectancy
1	Belize	71	66.16212822	78.37238547	72.26725684
2	Bhutan	69.8	61.80263378	74.0605028	67.93156829
3	Bulgaria	74.5	69.97434198	82.18598713	76.08016456
4	Chad	53.1	48.29794246	60.81678018	54.55736132
5	Djibouti	63.5	55.11131041	67.41207672	61.26169357
6	Liberia	61.4	52.63782269	65.10832532	58.87307401
7	Mali	58.2	53.05092471	65.40753788	59.2292313
8	Netherlands	81.9	76.27578723	88.6053711	82.44057917
9	Nicaragua	74.8	64.69310848	76.91456085	70.80383466
10	Philippines	68.5	61.1385437	74.11199345	67.62526858
11	Rwanda	66.1	57.49940105	69.86801653	63.68370879
12	Serbia	75.6	69.77581405	81.97587278	75.87584342
13	Sierra Leone	51	50.71112575	63.67291458	57.19202017
14	Thailand	74.9	68.07323416	80.2822403	74.17773723
15	Zambia	61.8	59.08857544	72.44735632	65.76796588

(Appendix, Prediction Intervals)

We have observed life expectancies of these countries from the original data, so we plotted a graph for observed life expectancy, predicted life expectancy and prediction intervals against different countries.

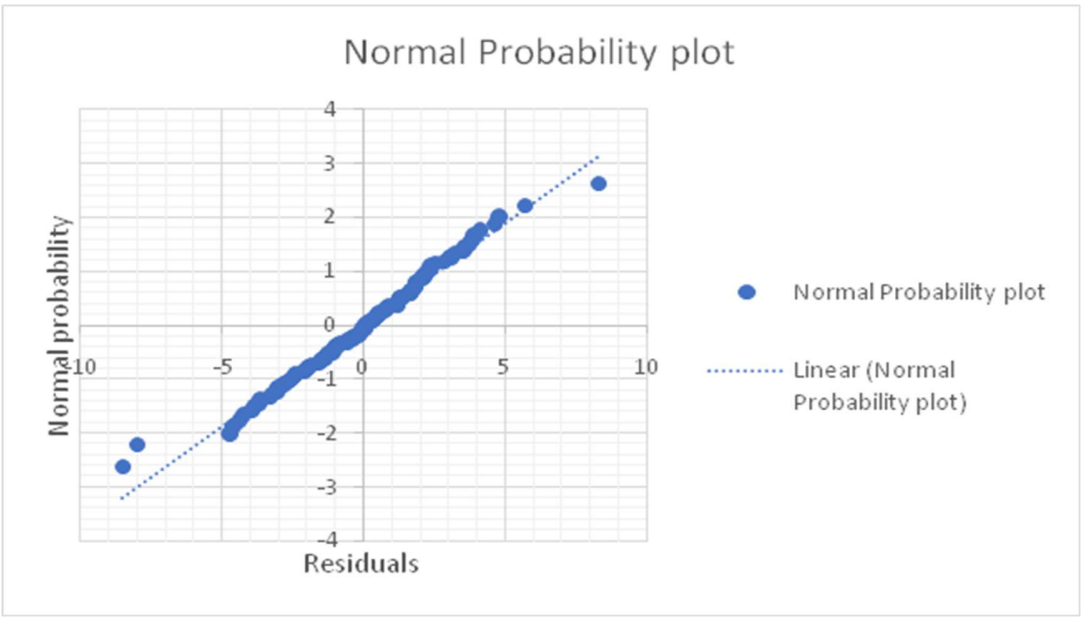


We can see from the graph that the life expectancy is lying within these prediction limits only, so we can conclude that our fitted model is suitable for future predictions of life expectancy.

Step 5: Residual Analysis

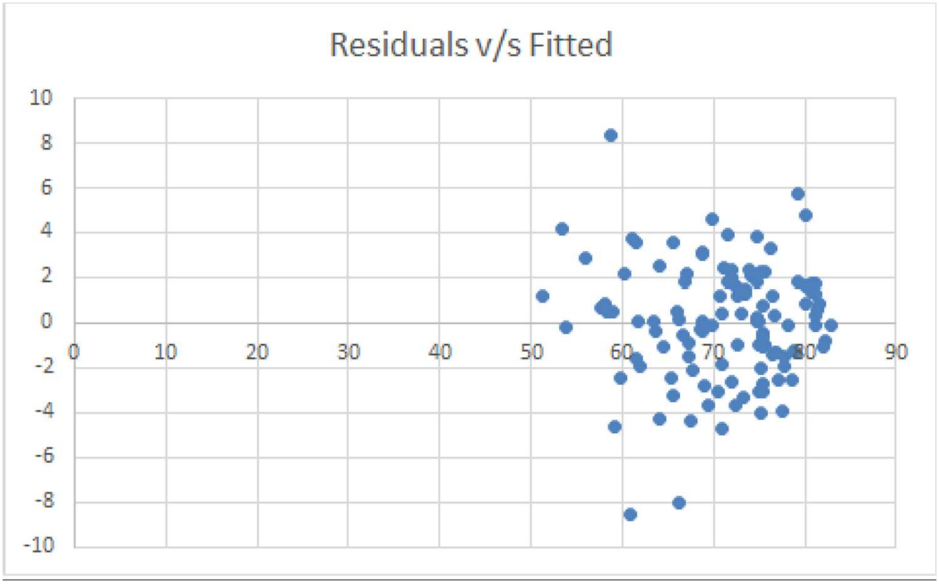
The residuals are given in Table no. 1 under Residual Analysis.

Normal Probability plot



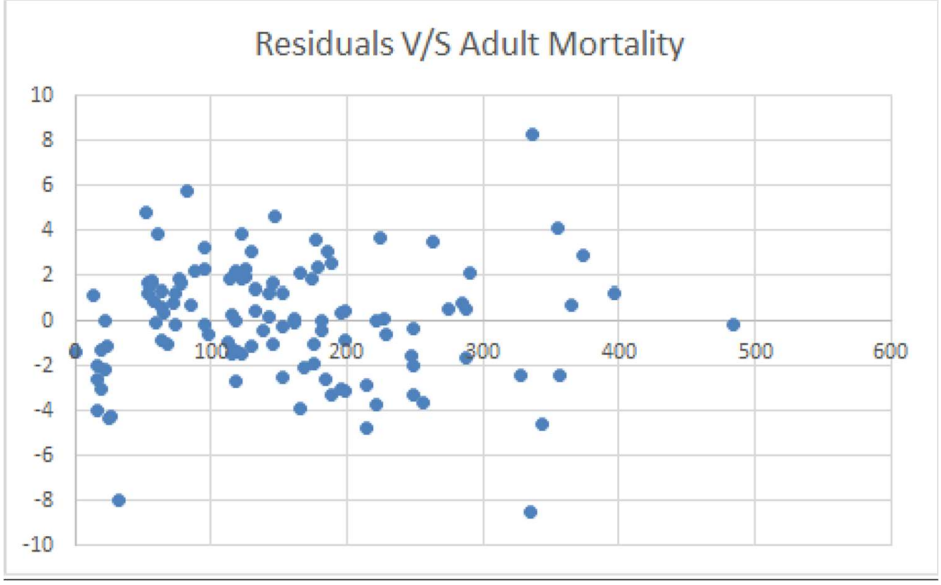
Since, the points are concentrated on the diagonal. Therefore, residuals are normally distributed.
3 outliers are present at the points (-8.50310146031102 , -2.62379), (-7.9896762516054 , -2.22491) and (8.31614297927339 , 2.62379).

Residuals v/s fitted values

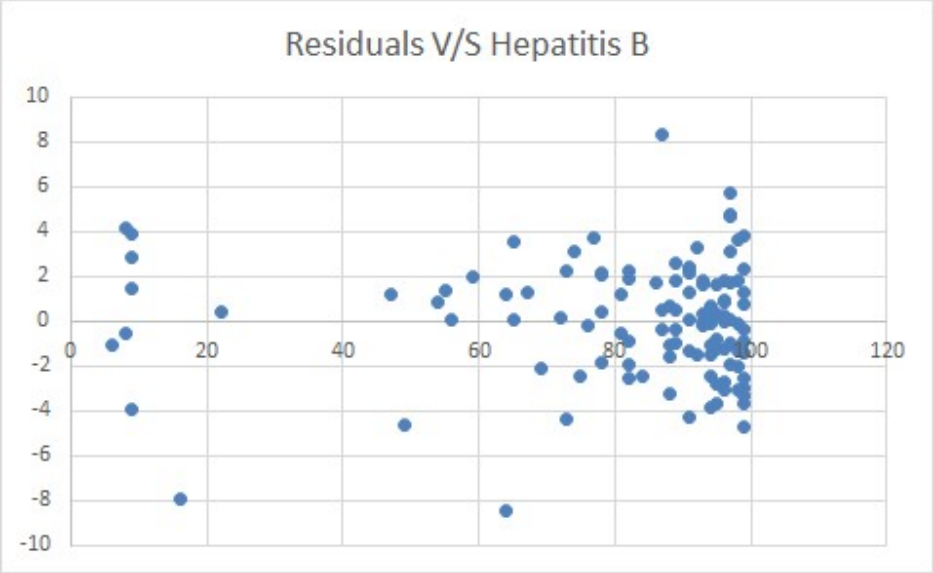


Data points lie in a band and are evenly distributed above and below 0. Hence, the residuals have a constant 0 mean and a constant variance.

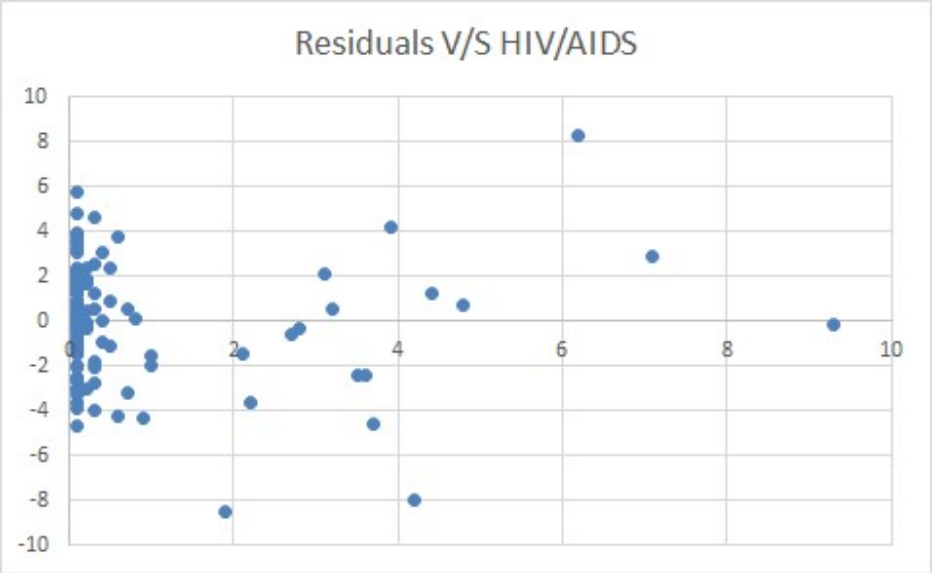
Residuals v/s regressors



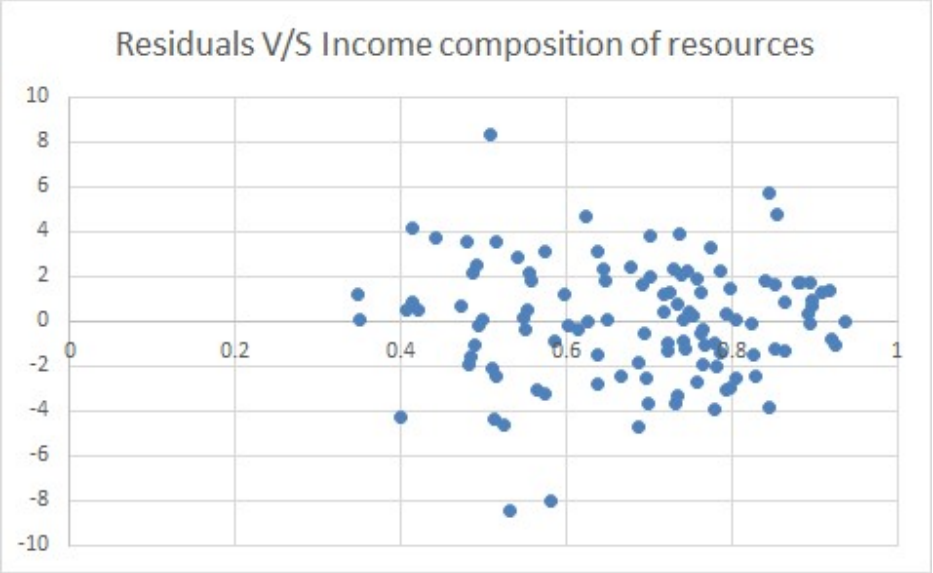
There is a linear relationship between Adult mortality and residuals. However, we can observe some clustering of points. Therefore, there is some variability left unexplained.



The data points show a non-random pattern and are not evenly distributed in a band. Therefore, there is no linear relationship between hepatitis and residuals.



The data points show a non-random pattern and are not evenly distributed in a band. Therefore, there is no linear relationship between HIV/AIDS and residuals.



We can see an even distribution of data points in a band. Therefore, there is a linear relationship between Income composition of resources and residuals.

Step 6 : Lack of fit test

Lack of fit test requires repeated observations for at least one setting of X to estimate the pure error for these repeated observations. As we don't have repeated observations at the same setting for any of the regressor variables(X), so we can't perform the lack of fit test on our model.

Step 7: Model Building

Model building is the process of developing a probabilistic model that best describes the relationship between the dependent and independent variables. We used SPSS to fit a multiple linear regression model using stepwise regression method.

We can see from the model building analysis *(Appendix, Model building, Table 2 and Table 3)* that Adult Mortality, HIV/Aids, Hepatitis B and Income Composition of Resources are the variables which best explains the response variable in the sense that

1) The coefficient of determination(R^2) is maximum for the model which contains the variables stated above.
2) The Mean Square Error (MSE) is also minimum for this model.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.892[a]	.796	.794	3.53552
2	.932[b]	.868	.866	2.85760
3	.939[c]	.882	.879	2.70784
4	.942[d]	.887	.883	2.66690

a. Predictors: (Constant), Incomw

b. Predictors: (Constant), Incomw, Adult_mortality

c. Predictors: (Constant), Incomw, Adult_mortality, Hepatitis_B

d. Predictors: (Constant), Incomw, Adult_mortality, Hepatitis_B, HIV_AIDS

Conclusion

We conclude from the above computations and calculations that the final model equation only includes four variables: adult mortality, Hepatitis B, HIV/AIDS, and income composition of resources.
Final Model Equation
The model equation we fitted is :

$$\hat{y} = 47.76311 - 0.02063 * x_1 + 0.034 * x_2 - 0.49539 * x_3 + 35.43532 * x_4$$

Where x1= Adult Mortality
x2 = Hepatitis B
x3 = HIV/AIDS
x4 = Income Composition of Resources

Firstly, the selected train data model of 115 countries is fitted, and then the computed factors are used to predict the values of the test data model of 15 countries. We conclude that the predicted values were approximately equal to the original data values.
From Residual Analysis, we concluded from the plots that clusters are obtained, which clearly point at the non-linear nature of the data model and also suggested the need of a different curvature for a better model. Outliers are also obtained, mainly for three countries: Angola, Cyprus, and Tajikistan in the drawn plots. If we remove the above outliers, we may get a better model.
Next, model fitting was done on the data of both developing countries and developed countries. The mean life expectancy of developed countries was 80.247 years and that of developing countries was 69.654 years which clearly indicates that the mean life expectancy was greater for developed countries than developing countries, which might be due to availability of effective healthcare facilities, higher standards of living, more resources in determinants of health, etc.

Future Analysis

- Since we had to remove the time factor from our data, time series analysis can be done if we choose to take the years into the account as well.
- We had to remove some of the variables due to lack of enough entries under them. How those variables effect the life expectancy can be done if a larger data is presented to us.
- As evident from graphs plotted under residual analysis, there is some variation left unexplained in the data. So, there might be a different curve (quadratic, cubic, polynomial etc.) that might be an even better fit for our data, in terms of lesser MSE.
- There are some outliers, namely, Angola, Tajikistan and Cyprus. Therefore, to analyse due to what conditions they are behaving as outliers, further study into factors affecting life expectancy should be undertaken.