Umang Gada

Stat\_641

Regression analysis on Life Expectancy data

Regression Project

REPORT

Note: Received Extension until Monday 27th April 2020. Thank you.

# Section 1: Explanation of Research Topic

## Topic: Life Expectancy data

### Data:

**Dataset:** WHO data for life expectancy in 193 countries for years 2000-2015.

**Goal:**

* The target of this exercise would be to check if one or more factors like adult mortality, infant deaths, alcohol consumption per-capita etc. have impact on the life-expectancy of these countries.
* Life expectancy improvements over the course of 15 years and for which countries it did improve.
* Additional relationships if any.

**Data:** In my analysis, I considered life expectancy as the Response variable, while I considered Year and Country as Categorical variable with 20 other columns as Continuous predictors. Depending on the significance of the predictors in the regression model, I will introduce interactions into the model.

**The following table gives description of each factors.**

|  |  |
| --- | --- |
| Country | Country |
| Year | Year |
| Status | Developed or Developing status |
| Life expectancy | Life Expectancy in age |
| Adult Mortality | Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population) |
| infant deaths | Number of Infant Deaths per 1000 population |
| Alcohol | Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol) |
| percentage expenditure | Expenditure on health as a percentage of Gross Domestic Product per capita(%) |
| Hepatitis B | Hepatitis B (HepB) immunization coverage among 1-year-olds (%) |
| Measles | Measles - number of reported cases per 1000 population |
| BMI | Average Body Mass Index of entire population |
| under-five deaths | Number of under-five deaths per 1000 population |
| Polio | Polio (Pol3) immunization coverage among 1-year-olds (%) |
| Total expenditure | General government expenditure on health as a percentage of total government expenditure (%) |
| Diphtheria | Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%) |
| HIV/AIDS | Deaths per 1 000 live births HIV/AIDS (0-4 years) |
| GDP | Gross Domestic Product per capita (in USD) |
| Population | Population of the country |
| thinness 1-19 years | Prevalence of thinness among children and adolescents for Age 10 to 19 (% ) |
| thinness 5-9 years | Prevalence of thinness among children for Age 5 to 9(%) |
| Income composition of resources | Human Development Index in terms of income composition of resources (index ranging from 0 to 1) |
| Schooling | Number of years of Schooling(years) |

# Section 2: Data Collection/Data Source

**Protocol:**

* In my search for a decent dataset, I visited the public repositories of UCI, Kaggle, Data.world, Lionbridge etc.
* I looked for a data that would have been acquired from a reputed organization such as WHO. We can never confirm the accuracy of the dataset, but this line of thought seemed appropriate in acquiring as accurate data as I could.
* After scouring the repositories for hours, I finally decided upon the life expectancy data. I have linked the source down below.

**Dataset Source:** <https://www.kaggle.com/kumarajarshi/life-expectancy-who>

**Acknowledgement:** The data was collected from WHO and United Nations website by Kumar Rajarshi (a Kaggle user) with the help of Deeksha Russell and Duan Wang.

# Section 3: Method of Analysis

In this section I will use the model

α = 0.05

Assuming our inputs are random, normal and independent.

**Null hypothesis:** There is no linear relationship between response variable (Life Expectancy) and the regressor variables.

**Alternate hypothesis:** There is a linear relationship between response variable (Life Expectancy) and the regressor variables.

We will check the validity of the model based on p-value, VIF for multicollinearity, correlation values between each regressors.

We will check the assumptions for model adequacy based on the residual normal probability plot and residual vs fits plot.

# Section 4: Results

## Data Analysis:

**Summary:**

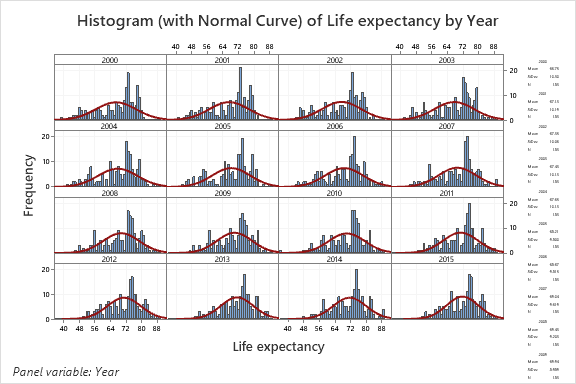
Variable : Life Expectancy

First, we will look at the data grouped by year column.

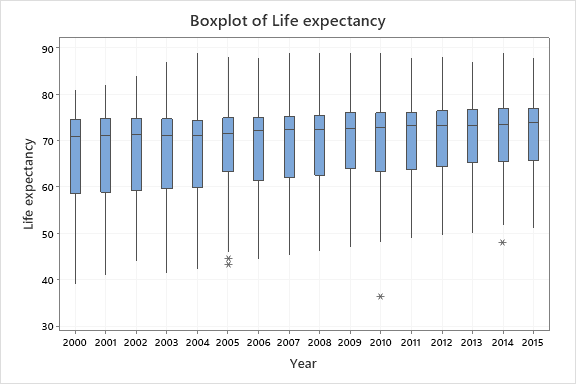
**Statistics**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Year** | **N** | **Mean** | **SE Mean** | **StDev** | **Variance** | **Minimum** | **Median** | **Maximum** | **IQR** |
| 2000 | 183 | 66.750 | 0.761 | 10.296 | 105.998 | 39.000 | 71.000 | 81.100 | 15.900 |
| 2001 | 183 | 67.129 | 0.753 | 10.190 | 103.829 | 41.000 | 71.200 | 82.000 | 16.000 |
| 2002 | 183 | 67.351 | 0.744 | 10.062 | 101.253 | 44.000 | 71.400 | 84.000 | 15.500 |
| 2003 | 183 | 67.433 | 0.749 | 10.128 | 102.570 | 41.500 | 71.100 | 87.000 | 15.000 |
| 2004 | 183 | 67.646 | 0.749 | 10.126 | 102.544 | 42.300 | 71.200 | 89.000 | 14.500 |
| 2005 | 183 | 68.209 | 0.724 | 9.800 | 96.031 | 43.300 | 71.600 | 88.000 | 11.700 |
| 2006 | 183 | 68.668 | 0.726 | 9.815 | 96.338 | 44.300 | 72.100 | 88.000 | 13.500 |
| 2007 | 183 | 69.036 | 0.711 | 9.619 | 92.517 | 45.300 | 72.400 | 89.000 | 13.300 |
| 2008 | 183 | 69.428 | 0.680 | 9.203 | 84.688 | 46.200 | 72.400 | 89.000 | 12.800 |
| 2009 | 183 | 69.938 | 0.664 | 8.989 | 80.804 | 47.100 | 72.600 | 89.000 | 11.900 |
| 2010 | 183 | 70.049 | 0.688 | 9.303 | 86.545 | 36.300 | 72.800 | 89.000 | 12.700 |
| 2011 | 183 | 70.654 | 0.660 | 8.925 | 79.656 | 48.900 | 73.300 | 88.000 | 12.300 |
| 2012 | 183 | 70.917 | 0.633 | 8.562 | 73.310 | 49.700 | 73.200 | 88.000 | 12.100 |
| 2013 | 183 | 71.236 | 0.622 | 8.414 | 70.792 | 49.900 | 73.200 | 87.000 | 11.300 |
| 2014 | 183 | 71.537 | 0.633 | 8.561 | 73.288 | 48.100 | 73.600 | 89.000 | 11.400 |
| 2015 | 183 | 71.617 | 0.601 | 8.124 | 65.995 | 51.000 | 73.900 | 88.000 | 11.300 |

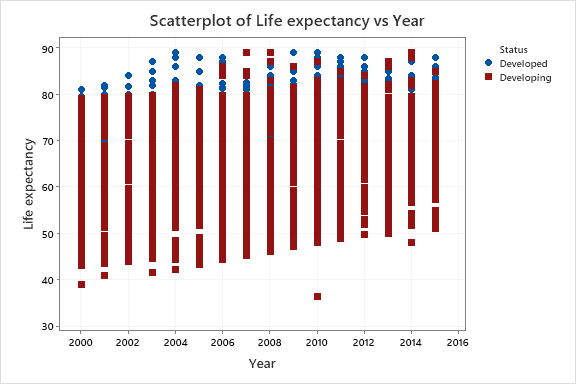
* From the above summary statistics, we can see that Median life expectancy is generally higher than the mean life expectancy of the population across 183 countries over the course of 15 years.
* This means that the data is pretty skewed. **We do not have a normal distribution.**



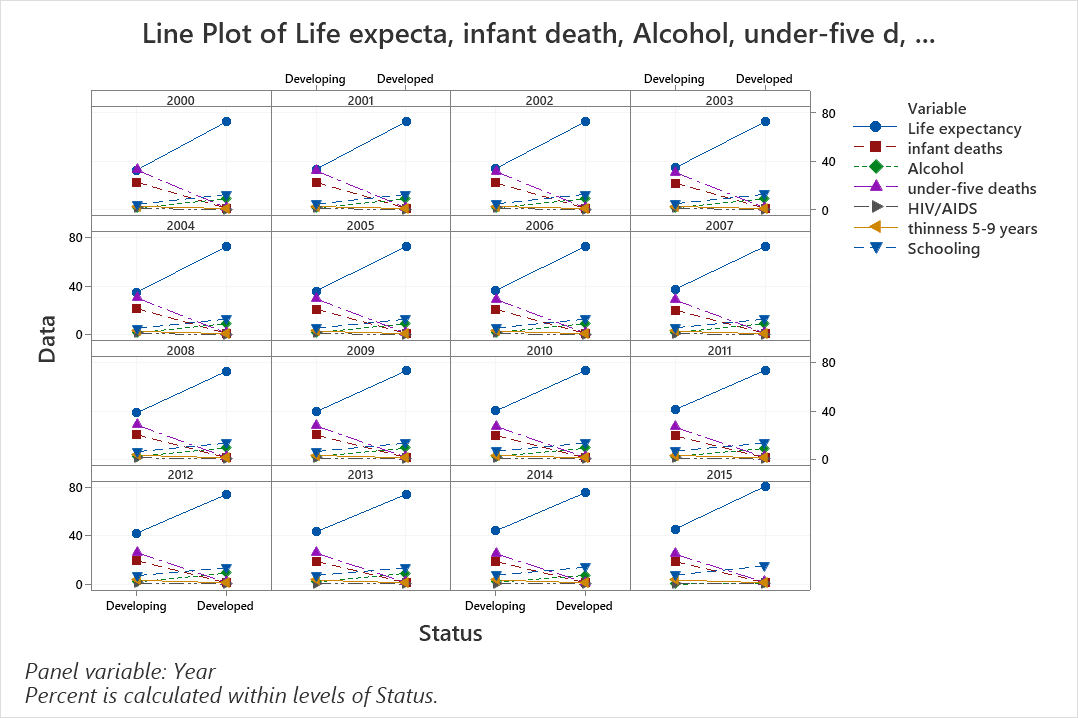
* The suspicions from above seem to be true, the distribution for each year for all the countries is negatively skewed.
* Another interesting thing is that the data pretty much remains the same over the course of 15 years.



* The boxplot confirms the skew with all the years having pretty similar variability.
* We do see outliers for years 2005, 2010 and 2014.
* Although the range pretty much remains the same, it seems like the IQR decreases slightly as we go forward in years.



* A scatter plot for 183 countries was not really clear but with status we can see that developed countries tend to have higher life expectancy.
* In the later years, developing countries have increased the life expectancy overall.



* The above line plot shows different factors vs country status.
* We see that Developed countries have higher life expectancy, higher schooling and generally higher alcohol consumption per capita.
* Developing countries on the other hand have higher infant and under-five deaths.
* Whereas the rest of the selected factors are pretty evenly divided.

# **Statistical Analysis:**

We’ll first put the data in a regression model and get an initial result.

**Coefficients**

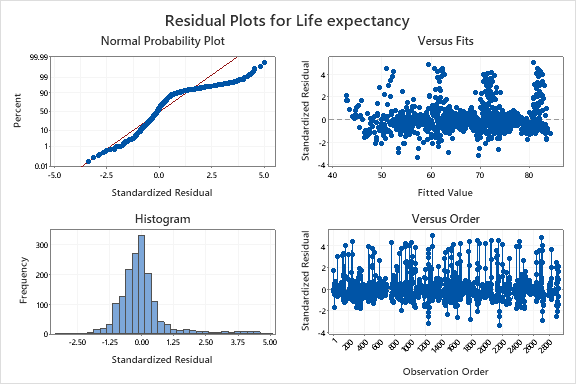
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Term** | **Coef** | **SE Coef** | **95% CI** | **T-Value** | **P-Value** | **VIF** |
| Constant | -391.0 | 34.3 | (-458.2, -323.8) | -11.41 | 0.000 |  |
| Year | 0.2219 | 0.0173 | (0.1880, 0.2559) | 12.84 | 0.000 | 2.97 |
| Adult\_Mortality | -0.000646 | 0.000541 | (-0.001707, 0.000415) | -1.19 | 0.233 | 2.73 |
| infant\_deaths | 0.0497 | 0.0155 | (0.0193, 0.0801) | 3.20 | 0.001 | **2087.19** |
| Alcohol | -0.0652 | 0.0303 | (-0.1247, -0.0057) | -2.15 | 0.032 | 8.88 |
| percentage\_expenditure | -0.000076 | 0.000122 | (-0.000316, 0.000163) | -0.62 | 0.532 | 27.53 |
| Hepatitis\_B | 0.00318 | 0.00242 | (-0.00156, 0.00792) | 1.32 | 0.189 | 2.28 |
| Measles | -0.000007 | 0.000006 | (-0.000019, 0.000006) | -1.01 | 0.311 | 2.53 |
| BMI | -0.00147 | 0.00340 | (-0.00815, 0.00520) | -0.43 | 0.666 | 2.69 |
| under\_five deaths | -0.0372 | 0.0109 | (-0.0587, -0.0158) | -3.40 | 0.001 | **1892.23** |
| Polio | -0.00060 | 0.00260 | (-0.00570, 0.00450) | -0.23 | 0.817 | 2.02 |
| Total\_expenditure | -0.0221 | 0.0264 | (-0.0738, 0.0297) | -0.84 | 0.403 | 2.19 |
| Diphtheria | 0.00090 | 0.00299 | (-0.00497, 0.00677) | 0.30 | 0.764 | 2.48 |
| HIV\_AIDS | -0.3029 | 0.0158 | (-0.3340, -0.2719) | -19.16 | 0.000 | 5.41 |
| GDP | 0.000014 | 0.000018 | (-0.000022, 0.000050) | 0.77 | 0.442 | 26.67 |
| Population | -0.000000 | 0.000000 | (-0.000000, 0.000000) | -0.32 | 0.749 | 2.55 |
| thinness\_10\_19\_years | 0.0115 | 0.0327 | (-0.0527, 0.0757) | 0.35 | 0.726 | 13.47 |
| thinness\_5\_9\_years | 0.0664 | 0.0312 | (0.0051, 0.1276) | 2.12 | 0.034 | 12.57 |
| Income\_composition\_of\_resources | 0.985 | 0.594 | (-0.181, 2.151) | 1.66 | 0.098 | 7.04 |
| Schooling | 0.2849 | 0.0780 | (0.1319, 0.4379) | 3.65 | 0.000 | 28.27 |
| Country |  |  |  |  |  |  |
| Albania | 16.548 | 0.857 | (14.867, 18.230) | 19.30 | 0.000 | 4.20 |
| Algeria | 14.313 | 0.850 | (12.646, 15.980) | 16.84 | 0.000 | 2.85 |
| Angola | -6.457 | 0.817 | (-8.060, -4.854) | -7.90 | 0.000 | 1.92 |
| Argentina | 15.44 | 1.07 | (13.35, 17.53) | 14.47 | 0.000 | 5.30 |
| Armenia | 14.873 | 0.836 | (13.233, 16.513) | 17.79 | 0.000 | 3.75 |
| Australia | 21.36 | 1.30 | (18.81, 23.92) | 16.41 | 0.000 | 8.49 |
| Austria | 22.37 | 1.07 | (20.28, 24.46) | 21.00 | 0.000 | 6.09 |
| Azerbaijan | 12.402 | 0.821 | (10.791, 14.012) | 15.11 | 0.000 | 3.14 |
| Bangladesh | 10.942 | 0.703 | (9.563, 12.321) | 15.57 | 0.000 | 2.12 |
| Belarus | 11.16 | 1.00 | (9.20, 13.12) | 11.15 | 0.000 | 5.37 |
| Belgium | 21.21 | 1.11 | (19.03, 23.40) | 19.04 | 0.000 | 6.66 |
| Belize | 10.681 | 0.844 | (9.024, 12.337) | 12.65 | 0.000 | 3.83 |
| Benin | 0.296 | 0.692 | (-1.063, 1.654) | 0.43 | 0.669 | 2.23 |
| Bhutan | 7.148 | 0.673 | (5.829, 8.467) | 10.63 | 0.000 | 2.43 |
| Bosnia and Herzegovina | 17.127 | 0.890 | (15.381, 18.873) | 19.24 | 0.000 | 3.13 |
| Botswana | 2.045 | 0.811 | (0.455, 3.635) | 2.52 | 0.012 | 3.52 |
| Brazil | 13.852 | 0.921 | (12.045, 15.659) | 15.04 | 0.000 | 4.55 |
| Bulgaria | 14.201 | 0.944 | (12.349, 16.053) | 15.04 | 0.000 | 4.78 |
| Burkina Faso | 0.639 | 0.802 | (-0.934, 2.211) | 0.80 | 0.426 | 2.08 |
| Burundi | -0.982 | 0.732 | (-2.418, 0.454) | -1.34 | 0.180 | 2.11 |
| Cabo Verde | 13.562 | 0.779 | (12.034, 15.089) | 17.42 | 0.000 | 2.82 |
| Cambodia | 7.090 | 0.753 | (5.613, 8.568) | 9.41 | 0.000 | 1.83 |
| Cameroon | -1.240 | 0.773 | (-2.756, 0.275) | -1.61 | 0.109 | 2.14 |
| Canada | 22.72 | 1.10 | (20.56, 24.87) | 20.69 | 0.000 | 5.18 |
| Central African Republic | -4.885 | 0.877 | (-6.605, -3.164) | -5.57 | 0.000 | 1.66 |
| Chad | -3.978 | 0.843 | (-5.632, -2.324) | -4.72 | 0.000 | 1.79 |
| Chile | 20.09 | 1.04 | (18.04, 22.14) | 19.25 | 0.000 | 3.52 |
| China | 14.49 | 1.21 | (12.13, 16.85) | 12.02 | 0.000 | 7.79 |
| Colombia | 14.723 | 0.839 | (13.078, 16.368) | 17.56 | 0.000 | 3.77 |
| Comoros | 3.433 | 0.728 | (2.004, 4.862) | 4.71 | 0.000 | 2.28 |
| Costa Rica | 19.911 | 0.871 | (18.203, 21.620) | 22.86 | 0.000 | 4.07 |
| Croatia | 17.26 | 1.06 | (15.17, 19.34) | 16.25 | 0.000 | 3.24 |
| Cyprus | 20.667 | 0.977 | (18.749, 22.584) | 21.15 | 0.000 | 5.12 |
| Djibouti | 6.034 | 0.854 | (4.359, 7.709) | 7.07 | 0.000 | 2.09 |
| Dominican Republic | 14.082 | 0.848 | (12.418, 15.745) | 16.61 | 0.000 | 3.86 |
| Ecuador | 15.965 | 0.873 | (14.254, 17.677) | 18.30 | 0.000 | 4.09 |
| El Salvador | 13.082 | 0.849 | (11.416, 14.749) | 15.40 | 0.000 | 3.87 |
| Equatorial Guinea | -0.46 | 1.76 | (-3.91, 2.99) | -0.26 | 0.795 | 1.12 |
| Eritrea | 4.850 | 0.790 | (3.300, 6.399) | 6.14 | 0.000 | 2.24 |
| Estonia | 14.94 | 1.07 | (12.85, 17.04) | 13.99 | 0.000 | 4.50 |
| Ethiopia | 6.226 | 0.856 | (4.547, 7.906) | 7.27 | 0.000 | 2.11 |
| Fiji | 9.190 | 0.889 | (7.445, 10.934) | 10.33 | 0.000 | 4.24 |
| France | 23.22 | 1.10 | (21.06, 25.38) | 21.11 | 0.000 | 6.49 |
| Gabon | 6.017 | 0.857 | (4.335, 7.699) | 7.02 | 0.000 | 2.64 |
| Georgia | 14.950 | 0.849 | (13.285, 16.614) | 17.61 | 0.000 | 3.86 |
| Germany | 21.81 | 1.12 | (19.60, 24.01) | 19.42 | 0.000 | 6.77 |
| Ghana | 3.810 | 0.707 | (2.424, 5.196) | 5.39 | 0.000 | 2.32 |
| Greece | 21.92 | 1.05 | (19.86, 23.99) | 20.83 | 0.000 | 5.94 |
| Guatemala | 14.727 | 0.846 | (13.067, 16.386) | 17.41 | 0.000 | 2.57 |
| Guinea | -0.033 | 0.781 | (-1.565, 1.499) | -0.04 | 0.966 | 1.75 |
| Guinea-Bissau | 0.168 | 0.866 | (-1.530, 1.866) | 0.19 | 0.846 | 1.62 |
| Guyana | 7.789 | 0.780 | (6.260, 9.318) | 9.99 | 0.000 | 3.05 |
| Haiti | 4.20 | 1.31 | (1.63, 6.78) | 3.20 | 0.001 | 1.25 |
| Honduras | 15.028 | 0.797 | (13.466, 16.591) | 18.87 | 0.000 | 3.40 |
| India | 5.83 | 2.86 | (0.23, 11.43) | 2.04 | 0.041 | 32.15 |
| Indonesia | 8.361 | 0.827 | (6.739, 9.984) | 10.11 | 0.000 | 3.67 |
| Iraq | 11.494 | 0.776 | (9.973, 13.016) | 14.82 | 0.000 | 2.37 |
| Ireland | 22.32 | 1.32 | (19.72, 24.91) | 16.88 | 0.000 | 3.14 |
| Israel | 21.37 | 1.03 | (19.36, 23.38) | 20.84 | 0.000 | 5.64 |
| Italy | 22.78 | 1.06 | (20.69, 24.86) | 21.43 | 0.000 | 6.06 |
| Jamaica | 15.858 | 0.878 | (14.137, 17.580) | 18.07 | 0.000 | 3.31 |
| Jordan | 13.708 | 0.857 | (12.028, 15.388) | 16.00 | 0.000 | 3.94 |
| Kazakhstan | 7.418 | 0.916 | (5.622, 9.214) | 8.10 | 0.000 | 4.50 |
| Kenya | 2.183 | 0.712 | (0.786, 3.579) | 3.07 | 0.002 | 2.36 |
| Kiribati | 7.059 | 0.855 | (5.381, 8.736) | 8.25 | 0.000 | 3.92 |
| Latvia | 14.40 | 1.00 | (12.44, 16.36) | 14.40 | 0.000 | 5.36 |
| Lebanon | 15.107 | 0.846 | (13.447, 16.766) | 17.86 | 0.000 | 3.84 |
| Lesotho | -3.110 | 0.821 | (-4.720, -1.499) | -3.79 | 0.000 | 2.90 |
| Liberia | 2.649 | 0.823 | (1.035, 4.262) | 3.22 | 0.001 | 1.70 |
| Lithuania | 13.49 | 1.05 | (11.43, 15.55) | 12.86 | 0.000 | 5.90 |
| Luxembourg | 22.08 | 1.05 | (20.01, 24.14) | 20.98 | 0.000 | 5.94 |
| Madagascar | 5.265 | 0.696 | (3.900, 6.629) | 7.57 | 0.000 | 2.25 |
| Malawi | -2.739 | 0.770 | (-4.249, -1.228) | -3.56 | 0.000 | 2.76 |
| Malaysia | 14.168 | 0.787 | (12.623, 15.712) | 17.99 | 0.000 | 3.33 |
| Maldives | 15.811 | 0.717 | (14.404, 17.217) | 22.05 | 0.000 | 2.76 |
| Mali | -0.832 | 0.741 | (-2.285, 0.621) | -1.12 | 0.262 | 2.36 |
| Malta | 21.692 | 0.996 | (19.739, 23.645) | 21.79 | 0.000 | 4.26 |
| Mauritania | 5.299 | 0.745 | (3.838, 6.760) | 7.11 | 0.000 | 1.99 |
| Mauritius | 13.137 | 0.820 | (11.527, 14.746) | 16.01 | 0.000 | 3.61 |
| Mexico | 16.930 | 0.860 | (15.242, 18.618) | 19.68 | 0.000 | 3.97 |
| Mongolia | 7.005 | 0.844 | (5.351, 8.660) | 8.30 | 0.000 | 3.82 |
| Montenegro | 15.240 | 0.986 | (13.305, 17.174) | 15.45 | 0.000 | 3.14 |
| Morocco | 13.499 | 0.730 | (12.067, 14.931) | 18.49 | 0.000 | 2.86 |
| Mozambique | -0.025 | 0.760 | (-1.516, 1.465) | -0.03 | 0.973 | 2.89 |
| Myanmar | 5.900 | 0.675 | (4.576, 7.224) | 8.74 | 0.000 | 1.96 |
| Namibia | 6.448 | 0.940 | (4.604, 8.293) | 6.86 | 0.000 | 1.59 |
| Nepal | 7.566 | 0.682 | (6.229, 8.903) | 11.10 | 0.000 | 2.00 |
| Netherlands | 19.70 | 1.42 | (16.91, 22.48) | 13.89 | 0.000 | 2.90 |
| Nicaragua | 15.321 | 0.808 | (13.736, 16.905) | 18.97 | 0.000 | 3.50 |
| Niger | 5.201 | 0.978 | (3.284, 7.119) | 5.32 | 0.000 | 2.06 |
| Nigeria | 1.05 | 1.66 | (-2.21, 4.30) | 0.63 | 0.528 | 9.87 |
| Pakistan | 5.28 | 1.09 | (3.14, 7.42) | 4.84 | 0.000 | 5.12 |
| Panama | 17.840 | 0.885 | (16.105, 19.575) | 20.17 | 0.000 | 3.92 |
| Papua New Guinea | 4.750 | 0.768 | (3.242, 6.257) | 6.18 | 0.000 | 3.17 |
| Paraguay | 14.148 | 0.863 | (12.456, 15.841) | 16.40 | 0.000 | 3.20 |
| Peru | 14.833 | 0.908 | (13.053, 16.614) | 16.34 | 0.000 | 3.54 |
| Philippines | 8.791 | 0.820 | (7.183, 10.399) | 10.72 | 0.000 | 3.61 |
| Poland | 16.35 | 1.00 | (14.38, 18.31) | 16.33 | 0.000 | 5.38 |
| Portugal | 20.88 | 1.07 | (18.78, 22.98) | 19.51 | 0.000 | 6.15 |
| Romania | 15.129 | 0.923 | (13.319, 16.939) | 16.40 | 0.000 | 4.57 |
| Russian Federation | 8.742 | 0.952 | (6.875, 10.609) | 9.19 | 0.000 | 4.54 |
| Rwanda | 3.884 | 0.738 | (2.436, 5.332) | 5.26 | 0.000 | 2.54 |
| Samoa | 15.182 | 0.885 | (13.446, 16.917) | 17.16 | 0.000 | 4.20 |
| Sao Tome and Principe | 7.341 | 0.758 | (5.855, 8.827) | 9.69 | 0.000 | 2.47 |
| Senegal | 6.145 | 0.750 | (4.674, 7.616) | 8.19 | 0.000 | 2.22 |
| Serbia | 15.096 | 0.952 | (13.229, 16.964) | 15.86 | 0.000 | 3.57 |
| Seychelles | 13.332 | 0.828 | (11.708, 14.957) | 16.10 | 0.000 | 3.68 |
| Sierra Leone | -9.472 | 0.779 | (-11.000, -7.943) | -12.16 | 0.000 | 1.74 |
| Solomon Islands | 10.278 | 0.784 | (8.740, 11.817) | 13.11 | 0.000 | 3.30 |
| South Africa | 4.428 | 0.820 | (2.820, 6.036) | 5.40 | 0.000 | 3.61 |
| Spain | 22.71 | 1.08 | (20.58, 24.84) | 20.95 | 0.000 | 6.30 |
| Sri Lanka | 13.033 | 0.796 | (11.473, 14.594) | 16.38 | 0.000 | 2.72 |
| Suriname | 12.146 | 0.874 | (10.431, 13.861) | 13.89 | 0.000 | 2.74 |
| Swaziland | 3.352 | 0.905 | (1.577, 5.127) | 3.70 | 0.000 | 4.39 |
| Sweden | 21.44 | 1.27 | (18.95, 23.93) | 16.89 | 0.000 | 2.32 |
| Syrian Arab Republic | 15.470 | 0.857 | (13.788, 17.152) | 18.04 | 0.000 | 2.11 |
| Tajikistan | 8.291 | 0.779 | (6.764, 9.818) | 10.65 | 0.000 | 2.82 |
| Thailand | 13.962 | 0.788 | (12.415, 15.508) | 17.71 | 0.000 | 3.34 |
| Timor-Leste | 6.763 | 0.864 | (5.067, 8.458) | 7.82 | 0.000 | 1.88 |
| Togo | -0.407 | 0.844 | (-2.062, 1.248) | -0.48 | 0.630 | 1.79 |
| Tonga | 13.430 | 0.944 | (11.579, 15.281) | 14.23 | 0.000 | 4.78 |
| Trinidad and Tobago | 12.549 | 0.826 | (10.929, 14.170) | 15.19 | 0.000 | 3.18 |
| Tunisia | 14.653 | 0.859 | (12.969, 16.337) | 17.07 | 0.000 | 3.95 |
| Turkey | 14.674 | 0.811 | (13.083, 16.265) | 18.10 | 0.000 | 3.53 |
| Turkmenistan | 6.567 | 0.804 | (4.989, 8.144) | 8.16 | 0.000 | 3.01 |
| Uganda | 1.544 | 0.780 | (0.013, 3.075) | 1.98 | 0.048 | 2.84 |
| Ukraine | 11.166 | 0.933 | (9.335, 12.996) | 11.97 | 0.000 | 4.67 |
| Uruguay | 16.787 | 0.979 | (14.867, 18.708) | 17.14 | 0.000 | 5.14 |
| Uzbekistan | 9.162 | 0.813 | (7.567, 10.756) | 11.27 | 0.000 | 3.07 |
| Vanuatu | 13.574 | 0.798 | (12.010, 15.139) | 17.02 | 0.000 | 3.41 |
| Zambia | 1.286 | 0.802 | (-0.287, 2.858) | 1.60 | 0.109 | 2.30 |
| Zimbabwe | -0.686 | 0.796 | (-2.249, 0.876) | -0.86 | 0.389 | 3.40 |

* From the VIF above, we see high multicollinearity in our factors. With the infant deaths and under-five deaths having the highest VIF of above 1000!
* Common sense dictates that if either infant deaths or under-five deaths increases, the life expectancy should increase or decrease yet we see one with a positive regression coefficient (infant deaths) and another with negative regression coefficient (under-five deaths). This is the result of multicollinearity.
* Also, infant deaths could be considered as a subset for under-five deaths.

**Analysis of Variance**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Adj SS** | **Adj MS** | **F-Value** | **P-Value** |
| Regression | 165 | 123430 | 748.061 | 270.63 | 0.000 |
| Adult Mortality | 1 | 4 | 3.837 | 1.39 | 0.239 |
| infant deaths | 1 | 27 | 26.541 | 9.60 | 0.002 |
| Alcohol | 1 | 14 | 14.213 | 5.14 | 0.024 |
| percentage expenditure | 1 | 1 | 1.386 | 0.50 | 0.479 |
| Hepatitis B | 1 | 5 | 4.773 | 1.73 | 0.189 |
| Measles | 1 | 2 | 2.159 | 0.78 | 0.377 |
| BMI | 1 | 1 | 0.564 | 0.20 | 0.652 |
| under-five deaths | 1 | 30 | 29.943 | 10.83 | 0.001 |
| Polio | 1 | 0 | 0.109 | 0.04 | 0.843 |
| Total expenditure | 1 | 2 | 1.513 | 0.55 | 0.460 |
| Diphtheria | 1 | 0 | 0.068 | 0.02 | 0.875 |
| HIV/AIDS | 1 | 999 | 998.686 | 361.30 | 0.000 |
| GDP | 1 | 2 | 1.999 | 0.72 | 0.395 |
| Population | 1 | 0 | 0.083 | 0.03 | 0.862 |
| thinness  10-19 years | 1 | 0 | 0.184 | 0.07 | 0.796 |
| thinness 5-9 years | 1 | 13 | 13.283 | 4.81 | 0.029 |
| Income composition of resources | 1 | 6 | 6.072 | 2.20 | 0.139 |
| Schooling | 1 | 36 | 35.703 | 12.92 | 0.000 |
| Year | 15 | 505 | 33.660 | 12.18 | 0.000 |
| Country | 132 | 16517 | 125.125 | 45.27 | 0.000 |
| Error | 1483 | 4099 | 2.764 |  |  |
| Total | 1648 | 127529 |  |  |  |

* The above ANOVA table shows a significant regression, since the p-value (very small) < 0.05. We can say that the life expectancy is related to one or more factors.
* Not all factors are significant. The significant factors include infant deaths, Alcohol consumption. Under-five deaths, deaths per 100 live births with HIV/AIDS, thinness for age 5-9 years
* population, Schooling, Year and Country.
* From the initial analysis, it looks like economic factors do not have any effect on life expectancy.
* From the VIF values, it doesn’t look like these p-values are very reliable. We can also check for model adequacy below.



* From the NPP above, we can see that the residuals are not normally distributed.
* The versus fits plot also have these tower patterns which means that the variance is not constant either.

Even though we have a high R-square, the model is not adequate.

**Initial Conclusion:** The model with all the factors in it is not a good model, we are yet to add interactions but first we need to mitigate some of the multicollinearity and adjust for model adequacy. Let’s look at a correlation table to get a clearer picture of which factors are highly correlated.

**Correlations**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Year** | **Life**  **exp** | **Adult Mor** | **Infant** | **Alc.** | **Per.**  **Expen.** | **Hep. B** | **Measles** | **BMI** | **under5** | **Polio** | **Total\_**  **Expen.** |
| Life  Expectancy | 0.170 |  |  |  |  |  |  |  |  |  |  |  |
| Adult  Mortality | -0.079 | -0.696 |  |  |  |  |  |  |  |  |  |  |
| Infant deaths | -0.036 | -0.197 | 0.079 |  |  |  |  |  |  |  |  |  |
| Alcohol | -0.049 | 0.405 | -0.196 | -0.117 |  |  |  |  |  |  |  |  |
| Percentage  expenditure | 0.033 | 0.382 | -0.243 | -0.086 | 0.341 |  |  |  |  |  |  |  |
| Hepatitis\_B | 0.106 | 0.257 | -0.162 | -0.225 | 0.086 | 0.017 |  |  |  |  |  |  |
| Measles | -0.082 | -0.158 | 0.031 | 0.501 | -0.053 | -0.057 | -0.121 |  |  |  |  |  |
| BMI | 0.105 | 0.568 | -0.387 | -0.227 | 0.338 | 0.231 | 0.155 | -0.176 |  |  |  |  |
| Under five  Deaths | -0.042 | -0.223 | 0.094 | 0.987 | -0.113 | -0.088 | -0.234 | 0.508 | -0.238 |  |  |  |
| Polio | 0.094 | 0.466 | -0.275 | -0.171 | 0.222 | 0.148 | 0.484 | -0.136 | 0.288 | -0.189 |  |  |
| Total  expenditure | 0.087 | 0.218 | -0.115 | -0.129 | 0.303 | 0.176 | 0.071 | -0.106 | 0.239 | -0.130 | 0.144 |  |
| Diphtheria | 0.135 | 0.479 | -0.275 | -0.176 | 0.222 | 0.144 | 0.610 | -0.142 | 0.286 | -0.196 | 0.672 | 0.160 |
| HIV\_AIDS | -0.139 | -0.557 | 0.524 | 0.025 | -0.050 | -0.098 | -0.113 | 0.031 | -0.244 | 0.038 | -0.160 | -0.0003 |
| GDP | 0.103 | 0.461 | -0.296 | -0.109 | 0.354 | 0.899 | 0.084 | -0.077 | 0.304 | -0.112 | 0.212 | 0.141 |
| Population | 0.017 | -0.022 | -0.014 | 0.557 | -0.035 | -0.026 | -0.124 | 0.266 | -0.072 | 0.544 | -0.039 | -0.080 |
| Thinness  10\_19\_years | -0.045 | -0.477 | 0.303 | 0.466 | -0.432 | -0.252 | -0.122 | 0.225 | -0.531 | 0.468 | -0.223 | -0.276 |
| Thinness  5\_9\_years | -0.048 | -0.472 | 0.308 | 0.471 | -0.421 | -0.254 | -0.127 | 0.221 | -0.538 | 0.472 | -0.224 | -0.283 |
| Income composition  of resources | 0.243 | 0.725 | -0.458 | -0.145 | 0.450 | 0.382 | 0.199 | -0.129 | 0.509 | -0.163 | 0.381 | 0.166 |
| Schooling | 0.213 | 0.752 | -0.455 | -0.195 | 0.548 | 0.391 | 0.223 | -0.138 | 0.558 | -0.211 | 0.416 | 0.263 |
|  | **Diph.** | **HIV**  **AIDS** | **GDP** | **Pop.** | **Thin 10\_19** | **Thin**  **5\_9** | **Income** | |
| Life\_expectancy |  |  |  |  |  |  |  | |
| Adult\_Mortality |  |  |  |  |  |  |  | |
| infant\_deaths |  |  |  |  |  |  |  | |
| Alcohol |  |  |  |  |  |  |  | |
| percentage\_expenditure |  |  |  |  |  |  |  | |
| Hepatitis\_B |  |  |  |  |  |  |  | |
| Measles |  |  |  |  |  |  |  | |
| BMI |  |  |  |  |  |  |  | |
| under\_five\_deaths |  |  |  |  |  |  |  | |
| Polio |  |  |  |  |  |  |  | |
| Total\_expenditure |  |  |  |  |  |  |  | |
| Diphtheria |  |  |  |  |  |  |  | |
| HIV\_AIDS | -0.165 |  |  |  |  |  |  | |
| GDP | 0.201 | -0.137 |  |  |  |  |  | |
| Population | -0.029 | -0.028 | -0.028 |  |  |  |  | |
| thinness\_10\_19\_years | -0.230 | 0.204 | -0.287 | 0.254 |  |  |  | |
| thinness\_5\_9\_years | -0.224 | 0.207 | -0.292 | 0.251 | 0.939 |  |  | |
| Income\_composition\_of\_resources | 0.401 | -0.249 | 0.461 | -0.009 | -0.422 | -0.411 |  | |
| Schooling | 0.424 | -0.222 | 0.450 | -0.032 | -0.477 | -0.466 | 0.800 | |

From the above correlations table for significant factors we can see that

* There are some really strong positive correlations between my predictors.
* Infant deaths and Under-five deaths have a correlation very close to 1. It makes sense that infant deaths would be included in under-five deaths.
* There is a high correlation between percent expenditure and GDP as percent expenditure is a percentage per GDP of capital.
* Thinness prevalent in children aged 5-9 and 10-19 too are highly correlated. As this is a 15-years data, it seems possible that the same children might have been observed as they moved from one age-group to another.
* Finally, there seems to be a high correlation between Schooling and Income composition of resources (Human Development Index based on income composition of resources.) It’s likely that Schooling influences the Human development index calculation up to an extent.

This time I will run this without the categorical variable: Country. This is because we have a lot of missing data and since each country will have a different statistic, filling these up with overall medians would not make much sense and would probably mess the data even more.

Instead, we will remove these observations and refit model again.

Let’s check the correlation table again.

**Correlations**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Year** | **Life expectancy** | | | **Adult Mortality** | | | **infant deaths** | | **Alcohol** | | | **percentage expenditure** | |
| Life expectancy | 0.172 |  | | |  | | |  | |  | | |  | |
| Adult Mortality | -0.071 | -0.678 | | |  | | |  | |  | | |  | |
| infant deaths | -0.040 | -0.173 | | | 0.053 | | |  | |  | | |  | |
| Alcohol | -0.053 | 0.432 | | | -0.219 | | | -0.117 | |  | | |  | |
| percentage expenditure | 0.030 | 0.411 | | | -0.255 | | | -0.090 | | 0.413 | | |  | |
| Hepatitis B | 0.107 | 0.212 | | | -0.106 | | | -0.222 | | 0.110 | | | 0.016 | |
| Measles | -0.099 | -0.137 | | | -0.000 | | | 0.519 | | -0.039 | | | -0.061 | |
| BMI | 0.095 | 0.593 | | | -0.379 | | | -0.229 | | 0.399 | | | 0.259 | |
| under-five deaths | -0.045 | -0.199 | | | 0.068 | | | 0.997 | | -0.114 | | | -0.092 | |
| Polio | 0.089 | 0.442 | | | -0.236 | | | -0.160 | | 0.254 | | | 0.162 | |
| Total expenditure | 0.073 | 0.236 | | | -0.128 | | | -0.151 | | 0.255 | | | 0.239 | |
| Diphtheria | 0.137 | 0.464 | | | -0.235 | | | -0.163 | | 0.259 | | | 0.162 | |
| HIV/AIDS | -0.149 | -0.566 | | | 0.534 | | | 0.008 | | -0.054 | | | -0.110 | |
| GDP | 0.090 | 0.466 | | | -0.286 | | | -0.104 | | 0.454 | | | 0.938 | |
| Population | 0.018 | -0.023 | | | -0.013 | | | 0.557 | | -0.036 | | | -0.026 | |
| thinness  10-19 years | -0.051 | -0.486 | | | 0.303 | | | 0.487 | | -0.424 | | | -0.272 | |
| thinness 5-9 years | -0.055 | -0.479 | | | 0.309 | | | 0.492 | | -0.414 | | | -0.274 | |
| Income composition of resources | 0.231 | 0.759 | | | -0.453 | | | -0.151 | | 0.564 | | | 0.408 | |
| Schooling | 0.206 | 0.769 | | | -0.446 | | | -0.203 | | 0.616 | | | 0.416 | |
|  | **Hepatitis B** | | | **Measles** | | **BMI** | | | **under-five deaths** | | **Polio** | | **Total expenditure** | |
| Life expectancy |  | | |  | |  | | |  | |  | |  | |
| Adult Mortality |  | | |  | |  | | |  | |  | |  | |
| infant deaths |  | | |  | |  | | |  | |  | |  | |
| Alcohol |  | | |  | |  | | |  | |  | |  | |
| percentage expenditure |  | | |  | |  | | |  | |  | |  | |
| Hepatitis B |  | | |  | |  | | |  | |  | |  | |
| Measles | -0.113 | | |  | |  | | |  | |  | |  | |
| BMI | 0.143 | | | -0.172 | |  | | |  | |  | |  | |
| under-five deaths | -0.232 | | | 0.526 | | -0.239 | | |  | |  | |  | |
| Polio | 0.465 | | | -0.114 | | 0.279 | | | -0.178 | |  | |  | |
| Total expenditure | 0.117 | | | -0.111 | | 0.247 | | | -0.151 | | 0.178 | |  | |
| Diphtheria | 0.610 | | | -0.118 | | 0.281 | | | -0.184 | | 0.681 | | 0.188 | |
| HIV/AIDS | -0.097 | | | 0.022 | | -0.243 | | | 0.020 | | -0.148 | | 0.006 | |
| GDP | 0.046 | | | -0.071 | | 0.308 | | | -0.108 | | 0.198 | | 0.237 | |
| Population | -0.124 | | | 0.266 | | -0.072 | | | 0.544 | | -0.039 | | -0.080 | |
| thinness  10-19 years | -0.122 | | | 0.237 | | -0.559 | | | 0.488 | | -0.223 | | -0.254 | |
| thinness 5-9 years | -0.129 | | | 0.232 | | -0.566 | | | 0.492 | | -0.225 | | -0.269 | |
| Income composition of resources | 0.192 | | | -0.138 | | 0.547 | | | -0.171 | | 0.404 | | 0.236 | |
| Schooling | 0.224 | | | -0.145 | | 0.588 | | | -0.220 | | 0.435 | | 0.299 | |
|  | **Diphtheria** | | **HIV/AIDS** | | | | **GDP** | **Population** | | | | **thinness 10-19 years** | | **thinness 5-9 years** | **Income composition of resources** |
| Life expectancy |  | |  | | | |  |  | | | |  | |  |  |
| Adult Mortality |  | |  | | | |  |  | | | |  | |  |  |
| infant deaths |  | |  | | | |  |  | | | |  | |  |  |
| Alcohol |  | |  | | | |  |  | | | |  | |  |  |
| percentage expenditure |  | |  | | | |  |  | | | |  | |  |  |
| Hepatitis B |  | |  | | | |  |  | | | |  | |  |  |
| Measles |  | |  | | | |  |  | | | |  | |  |  |
| BMI |  | |  | | | |  |  | | | |  | |  |  |
| under-five deaths |  | |  | | | |  |  | | | |  | |  |  |
| Polio |  | |  | | | |  |  | | | |  | |  |  |
| Total expenditure |  | |  | | | |  |  | | | |  | |  |  |
| Diphtheria |  | |  | | | |  |  | | | |  | |  |  |
| HIV/AIDS | -0.158 | |  | | | |  |  | | | |  | |  |  |
| GDP | 0.195 | | -0.130 | | | |  |  | | | |  | |  |  |
| Population | -0.030 | | -0.028 | | | | -0.028 |  | | | |  | |  |  |
| thinness  10-19 years | -0.238 | | 0.204 | | | | -0.307 | 0.254 | | | |  | |  |  |
| thinness 5-9 years | -0.233 | | 0.206 | | | | -0.309 | 0.252 | | | | 0.932 | |  |  |
| Income composition of resources | 0.438 | | -0.258 | | | | 0.474 | -0.010 | | | | -0.465 | | -0.450 |  |
| Schooling | 0.452 | | -0.223 | | | | 0.485 | -0.034 | | | | -0.495 | | -0.482 | 0.831 |

We will remove one of the factors that are highly collinear as highlighted above.

Factors to be removed: Infant deaths, percent\_expenditure, thinness 10\_19 years and Income composition of resources (Human Development Index)

**Regression Equation**

|  |  |  |
| --- | --- | --- |
| Life expectancy | = | 269.1 - 0.019130 Adult Mortality - 0.0788 Alcohol + 0.04197 BMI + 0.0714 Total expenditure + 0.02667 Diphtheria - 0.4590 HIV/AIDS + 0.000086 GDP - 0.1076 Year - 0.00539 Hepatitis B + 0.000017 Measles - 0.002364 under-five deaths + 0.00808 Polio + 0.000000 Population - 0.0507 thinness 5-9 years + 1.3142 Schooling |

From the above model,

We can say that

* For each unit increase in Life expectancy: Adult mortality decreases by 0.019 units, Alcohol consumption decreases by 0.0788 units, HIV/AIDS deaths/1000 births decreases by 0.459 units, Hepatitis B immunization coverage decreases by 0.005 units, under-five deaths decreases by 0.002 units and thinness for children aged 5-9 years decreases by 0.05 units.
* This makes sense as decrease in mortality, alcohol consumption and/or thinness should ideally increase life expectancy.
* For each unit increase in Life expectancy: BMI increases by 0.04 units, Total expenditure of government on health increases by 0.0714 units, Diphtheria (DTP3) immunization coverage increases by 0.026 units, GDP increases by 0.000086 units, number of reported Measles cases increases by 0.000017 units, Polio immunization coverage increases by 0.00808 units, Population increases by a small factor, and Schooling years increases by 1.3142
* This makes sense too as higher Life expectancy should allow government to focus on economy, education and immunizations.

**Coefficients**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Term** | **Coef** | **SE Coef** | **95% CI** | **T-Value** | **P-Value** | **VIF** |
| Constant | 269.1 | 48.9 | (173.1, 365.0) | 5.50 | 0.000 |  |
| Adult Mortality | -0.019130 | 0.000993 | (-0.021077, -0.017183) | -19.27 | 0.000 | 1.75 |
| Alcohol | -0.0788 | 0.0325 | (-0.1424, -0.0151) | -2.43 | 0.015 | 1.93 |
| BMI | 0.04197 | 0.00635 | (0.02952, 0.05442) | 6.61 | 0.000 | 1.78 |
| Total expenditure | 0.0714 | 0.0433 | (-0.0134, 0.1563) | 1.65 | 0.099 | 1.12 |
| Diphtheria | 0.02667 | 0.00624 | (0.01443, 0.03892) | 4.27 | 0.000 | 2.05 |
| HIV/AIDS | -0.4590 | 0.0191 | (-0.4965, -0.4215) | -24.03 | 0.000 | 1.50 |
| GDP | 0.000086 | 0.000010 | (0.000067, 0.000105) | 8.83 | 0.000 | 1.40 |
| Year | -0.1076 | 0.0244 | (-0.1554, -0.0597) | -4.41 | 0.000 | 1.12 |
| Hepatitis B | -0.00539 | 0.00475 | (-0.01472, 0.00393) | -1.13 | 0.257 | 1.67 |
| Measles | 0.000017 | 0.000011 | (-0.000005, 0.000038) | 1.51 | 0.132 | 1.41 |
| under-five deaths | -0.002364 | 0.000949 | (-0.004226, -0.000502) | -2.49 | 0.013 | 2.70 |
| Polio | 0.00808 | 0.00548 | (-0.00267, 0.01883) | 1.47 | 0.141 | 1.71 |
| Population | 0.000000 | 0.000000 | (-0.000000, 0.000000) | 1.69 | 0.092 | 1.83 |
| thinness 5-9 years | -0.0507 | 0.0282 | (-0.1061, 0.0046) | -1.80 | 0.072 | 1.95 |
| Schooling | 1.3142 | 0.0548 | (1.2068, 1.4216) | 24.00 | 0.000 | 2.65 |

From the above table, we can see that the VIF values are pretty reasonable now.

**Model Summary**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **S** | **R-sq** | **R-sq(adj)** | **PRESS** | **R-sq(pred)** | **AICc** | **BIC** |
| 3.82073 | 81.31% | 81.14% | 24402.3 | 80.87% | 9118.73 | 9210.29 |

From the above model summary table, we can see that the R-squared is a bit lower from when we had country in the model, but the PRESS value is smaller than before and the predicted R-squared is also pretty good.

**Analysis of Variance**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Seq SS** | **Contribution** | **Adj SS** | **Adj MS** | **F-Value** | **P-Value** |
| Regression | 15 | 103691 | 81.31% | 103691 | 6912.72 | 473.54 | 0.000 |
| Adult Mortality | 1 | 62941 | 49.35% | 5422 | 5421.80 | 371.41 | 0.000 |
| Alcohol | 1 | 10272 | 8.05% | 86 | 85.97 | 5.89 | 0.015 |
| BMI | 1 | 7199 | 5.64% | 638 | 638.11 | 43.71 | 0.000 |
| Total expenditure | 1 | 127 | 0.10% | 40 | 39.80 | 2.73 | 0.099 |
| Diphtheria | 1 | 2442 | 1.92% | 267 | 266.51 | 18.26 | 0.000 |
| HIV/AIDS | 1 | 8515 | 6.68% | 8430 | 8430.14 | 577.49 | 0.000 |
| GDP | 1 | 2742 | 2.15% | 1138 | 1137.78 | 77.94 | 0.000 |
| Year | 1 | 11 | 0.01% | 283 | 283.34 | 19.41 | 0.000 |
| Hepatitis B | 1 | 2 | 0.00% | 19 | 18.78 | 1.29 | 0.257 |
| Measles | 1 | 7 | 0.01% | 33 | 33.12 | 2.27 | 0.132 |
| under-five deaths | 1 | 413 | 0.32% | 91 | 90.55 | 6.20 | 0.013 |
| Polio | 1 | 259 | 0.20% | 32 | 31.72 | 2.17 | 0.141 |
| Population | 1 | 228 | 0.18% | 41 | 41.45 | 2.84 | 0.092 |
| thinness 5-9 years | 1 | 125 | 0.10% | 47 | 47.19 | 3.23 | 0.072 |
| Schooling | 1 | 8408 | 6.59% | 8408 | 8408.22 | 575.99 | 0.000 |
| Error | 1633 | 23838 | 18.69% | 23838 | 14.60 |  |  |
| Total | 1648 | 127529 | 100.00% |  |  |  |  |

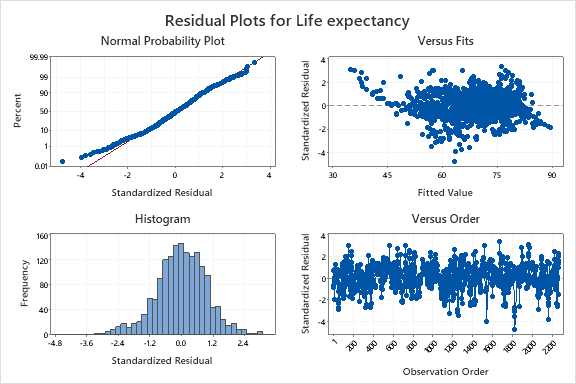
Looking at the p-values, we can conclude that Adult Mortality, Alcohol consumption, BMI, Diphtheria immunization, HIV/AIDS deaths/1000 births, GDP, Year, under-five deaths and Schooling are all significant.

We still need to check for interactions. Now that we have reduced the number of significant factors to 9 from 19, we can hope to perform much less 2-way interactions.

**We also found 102 Large Residuals (Outliers) and 62 Unusual X (Leverage points).**

*A table can be found in the Minitab file at the end of the report.*

We will not remove these values as each values represent a national data for 1 year.



From the above residual plot,

* the residuals mostly follow the regression line, we can assume normally distributed residuals.
* The versus fits plot looks fine, we can assume constant variance.
* We can assume that the model is adequate.

Let’s add interactions to the model.

Also, we will standardize the predictors to balance out the structural multicollinearity.

**Continuous predictor standardization**

Subtract the mean

|  |  |
| --- | --- |
| **Predictor** | **Mean** |
| Year | 2008 |
| Adult Mortality | 168 |
| Alcohol | 5 |
| Hepatitis B | 79 |
| BMI | 38 |
| Measles | 2224 |
| under-five deaths | 44 |
| Polio | 84 |
| Total expenditure | 6 |
| Diphtheria | 84 |
| HIV/AIDS | 2 |
| GDP | 5566 |
| Population | 14653626 |
| thinness 5-9 years | 5 |
| Schooling | 12 |

**Regression Equation in Uncoded Units**

|  |  |  |
| --- | --- | --- |
| Life expectancy | = | 464 + 1.099 Adult Mortality - 30.9 Alcohol - 7.17 BMI + 0.0922 Total expenditure + 2.84 Diphtheria + 28.8 HIV/AIDS + 0.00009 GDP - 0.212 Year - 0.00084 Hepatitis B + 0.000026 Measles - 0.659 under-five deaths + 0.00512 Polio - 0.000000 Population - 0.0676 thinness 5-9 years - 17.1 Schooling - 0.000953 Adult Mortality\*Alcohol - 0.000030 Adult Mortality\*BMI + 0.000012 Adult Mortality\*under-five deaths - 0.000079 Adult Mortality\*Diphtheria + 0.000513 Adult Mortality\*HIV/AIDS + 0.000000 Adult Mortality\*GDP - 0.001467 Adult Mortality\*Schooling - 0.000545 Year\*Adult Mortality + 0.00674 Alcohol\*BMI - 0.000112 Alcohol\*under-five deaths + 0.00115 Alcohol\*Diphtheria + 0.0455 Alcohol\*HIV/AIDS - 0.000011 Alcohol\*GDP + 0.0330 Alcohol\*Schooling + 0.01507 Year\*Alcohol - 0.000199 BMI\*under-five deaths - 0.001059 BMI\*Diphtheria + 0.00435 BMI\*HIV/AIDS + 0.000001 BMI\*GDP - 0.02209 BMI\*Schooling + 0.00376 Year\*BMI + 0.000000 under-five deaths\*Diphtheria - 0.001693 under-five deaths\*HIV/AIDS + 0.000000 under-five deaths\*GDP - 0.001413 under-five deaths\*Schooling + 0.000337 Year\*under-five deaths - 0.000666 Diphtheria\*HIV/AIDS - 0.000001 Diphtheria\*GDP - 0.00225 Diphtheria\*Schooling - 0.00137 Year\*Diphtheria - 0.000002 HIV/AIDS\*GDP - 0.0306 HIV/AIDS\*Schooling - 0.01465 Year\*HIV/AIDS - 0.000022 GDP\*Schooling + 0.000000 Year\*GDP + 0.0097 Year\*Schooling |

We get the new regression model as shown above.

**Coded Coefficients**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Term** | **Coef** | **SE Coef** | | | **95% CI** | **T-Value** |
| Constant | 69.500 | 0.202 | | | (69.104, 69.896) | 344.07 |
| Adult Mortality | -0.02121 | 0.00123 | | | (-0.02363, -0.01879) | -17.19 |
| Alcohol | -0.0697 | 0.0347 | | | (-0.1377, -0.0017) | -2.01 |
| BMI | 0.04414 | 0.00753 | | | (0.02937, 0.05892) | 5.86 |
| Total expenditure | 0.0922 | 0.0392 | | | (0.0153, 0.1691) | 2.35 |
| Diphtheria | 0.01272 | 0.00619 | | | (0.00057, 0.02486) | 2.05 |
| HIV/AIDS | -0.6771 | 0.0513 | | | (-0.7777, -0.5765) | -13.20 |
| GDP | 0.000253 | 0.000042 | | | (0.000170, 0.000336) | 5.98 |
| Year | -0.1023 | 0.0232 | | | (-0.1478, -0.0568) | -4.41 |
| Hepatitis B | -0.00084 | 0.00441 | | | (-0.00950, 0.00781) | -0.19 |
| Measles | 0.000026 | 0.000011 | | | (0.000004, 0.000049) | 2.33 |
| under-five deaths | -0.00857 | 0.00416 | | | (-0.01673, -0.00041) | -2.06 |
| Polio | 0.00512 | 0.00492 | | | (-0.00453, 0.01477) | 1.04 |
| Population | -0.000000 | 0.000000 | | | (-0.000000, 0.000000) | -0.26 |
| thinness 5-9 years | -0.0676 | 0.0271 | | | (-0.1207, -0.0144) | -2.49 |
| Schooling | 0.9577 | 0.0573 | | | (0.8453, 1.0701) | 16.71 |
| Adult Mortality\*Alcohol | -0.000953 | 0.000314 | | | (-0.001569, -0.000337) | -3.04 |
| Adult Mortality\*BMI | -0.000030 | 0.000069 | | | (-0.000165, 0.000104) | -0.44 |
| Adult Mortality\*under-five deaths | 0.000012 | 0.000006 | | | (0.000000, 0.000024) | 1.98 |
| Adult Mortality\*Diphtheria | -0.000079 | 0.000040 | | | (-0.000158, 0.000001) | -1.95 |
| Adult Mortality\*HIV/AIDS | 0.000513 | 0.000080 | | | (0.000356, 0.000671) | 6.39 |
| Adult Mortality\*GDP | 0.000000 | 0.000000 | | | (-0.000000, 0.000001) | 1.88 |
| Adult Mortality\*Schooling | -0.001467 | 0.000534 | | | (-0.002514, -0.000420) | -2.75 |
| Year\*Adult Mortality | -0.000545 | 0.000259 | | | (-0.001052, -0.000038) | -2.11 |
| Alcohol\*BMI | 0.00674 | 0.00162 | | | (0.00357, 0.00991) | 4.17 |
| Alcohol\*under-five deaths | -0.000112 | 0.000338 | | | (-0.000775, 0.000552) | -0.33 |
| Alcohol\*Diphtheria | 0.00115 | 0.00170 | | | (-0.00218, 0.00449) | 0.68 |
| Alcohol\*HIV/AIDS | 0.0455 | 0.0101 | | | (0.0257, 0.0652) | 4.51 |
| Alcohol\*GDP | -0.000011 | 0.000003 | | | (-0.000017, -0.000004) | -3.30 |
| Alcohol\*Schooling | 0.0330 | 0.0131 | | | (0.0072, 0.0588) | 2.51 |
| Year\*Alcohol | 0.01507 | 0.00703 | | | (0.00128, 0.02886) | 2.14 |
| BMI\*under-five deaths | -0.000199 | 0.000127 | | | (-0.000449, 0.000051) | -1.56 |
| BMI\*Diphtheria | -0.001059 | 0.000293 | | | (-0.001634, -0.000484) | -3.61 |
| BMI\*HIV/AIDS | 0.00435 | 0.00220 | | | (0.00003, 0.00866) | 1.98 |
| BMI\*GDP | 0.000001 | 0.000000 | | | (-0.000000, 0.000002) | 1.66 |
| BMI\*Schooling | -0.02209 | 0.00269 | | | (-0.02736, -0.01682) | -8.22 |
| Year\*BMI | 0.00376 | 0.00134 | | | (0.00113, 0.00639) | 2.80 |
| under-five deaths\*Diphtheria | 0.000000 | 0.000028 | | | (-0.000054, 0.000055) | 0.01 |
| under-five deaths\*HIV/AIDS | -0.001693 | 0.000423 | | | (-0.002522, -0.000864) | -4.01 |
| under-five deaths\*GDP | 0.000000 | 0.000001 | | | (-0.000001, 0.000002) | 0.23 |
| under-five deaths\*Schooling | -0.001413 | 0.000653 | | | (-0.002694, -0.000132) | -2.16 |
| Year\*under-five deaths | 0.000337 | 0.000231 | | | (-0.000117, 0.000791) | 1.45 |
| Diphtheria\*HIV/AIDS | -0.000666 | 0.000878 | | | (-0.002388, 0.001057) | -0.76 |
| Diphtheria\*GDP | -0.000001 | 0.000001 | | | (-0.000003, 0.000001) | -1.33 |
| Diphtheria\*Schooling | -0.00225 | 0.00245 | | | (-0.00705, 0.00256) | -0.92 |
| Year\*Diphtheria | -0.00137 | 0.00106 | | | (-0.00344, 0.00071) | -1.29 |
| HIV/AIDS\*GDP | -0.000002 | 0.000012 | | | (-0.000025, 0.000022) | -0.13 |
| HIV/AIDS\*Schooling | -0.0306 | 0.0160 | | | (-0.0619, 0.0007) | -1.91 |
| Year\*HIV/AIDS | -0.01465 | 0.00618 | | | (-0.02676, -0.00253) | -2.37 |
| GDP\*Schooling | -0.000022 | 0.000005 | | | (-0.000031, -0.000013) | -4.74 |
| Year\*GDP | 0.000000 | 0.000003 | | | (-0.000005, 0.000005) | 0.09 |
| Year\*Schooling | 0.0097 | 0.0121 | | | (-0.0140, 0.0334) | 0.80 |
|  |  |  | | |  |  |
|  |  |  | | |  |  |
|  |  |  | | |  |  |
|  |  |  | | |  |  |
|  |  |  | | |  |  |
|  |  |  | | |  |  |
| **Term** | **P-Value** | | **VIF** |
| Constant | 0.000 | |  |
| Adult Mortality | 0.000 | | 3.48 |
| Alcohol | 0.044 | | 2.84 |
| BMI | 0.000 | | 3.23 |
| Total expenditure | 0.019 | | 1.18 |
| Diphtheria | 0.040 | | 2.60 |
| HIV/AIDS | 0.000 | | 13.96 |
| GDP | 0.000 | | 34.33 |
| Year | 0.000 | | 1.31 |
| Hepatitis B | 0.849 | | 1.86 |
| Measles | 0.020 | | 1.91 |
| under-five deaths | 0.040 | | 67.00 |
| Polio | 0.298 | | 1.78 |
| Population | 0.797 | | 2.23 |
| thinness 5-9 years | 0.013 | | 2.32 |
| Schooling | 0.000 | | 3.74 |
| Adult Mortality\*Alcohol | 0.002 | | 2.76 |
| Adult Mortality\*BMI | 0.658 | | 3.34 |
| Adult Mortality\*under-five deaths | 0.048 | | 2.28 |
| Adult Mortality\*Diphtheria | 0.051 | | 2.08 |
| Adult Mortality\*HIV/AIDS | 0.000 | | 4.19 |
| Adult Mortality\*GDP | 0.060 | | 11.35 |
| Adult Mortality\*Schooling | 0.006 | | 4.52 |
| Year\*Adult Mortality | 0.035 | | 2.58 |
| Alcohol\*BMI | 0.000 | | 2.42 |
| Alcohol\*under-five deaths | 0.741 | | 4.44 |
| Alcohol\*Diphtheria | 0.497 | | 2.50 |
| Alcohol\*HIV/AIDS | 0.000 | | 3.05 |
| Alcohol\*GDP | 0.001 | | 7.05 |
| Alcohol\*Schooling | 0.012 | | 3.30 |
| Year\*Alcohol | 0.032 | | 1.98 |
| BMI\*under-five deaths | 0.118 | | 30.15 |
| BMI\*Diphtheria | 0.000 | | 2.15 |
| BMI\*HIV/AIDS | 0.048 | | 5.15 |
| BMI\*GDP | 0.097 | | 2.13 |
| BMI\*Schooling | 0.000 | | 2.96 |
| Year\*BMI | 0.005 | | 1.62 |
| under-five deaths\*Diphtheria | 0.993 | | 2.62 |
| under-five deaths\*HIV/AIDS | 0.000 | | 3.79 |
| under-five deaths\*GDP | 0.815 | | 67.18 |
| under-five deaths\*Schooling | 0.031 | | 9.78 |
| Year\*under-five deaths | 0.146 | | 2.19 |
| Diphtheria\*HIV/AIDS | 0.448 | | 1.90 |
| Diphtheria\*GDP | 0.183 | | 3.92 |
| Diphtheria\*Schooling | 0.359 | | 3.34 |
| Year\*Diphtheria | 0.197 | | 1.30 |
| HIV/AIDS\*GDP | 0.898 | | 20.78 |
| HIV/AIDS\*Schooling | 0.056 | | 6.27 |
| Year\*HIV/AIDS | 0.018 | | 4.21 |
| GDP\*Schooling | 0.000 | | 5.64 |
| Year\*GDP | 0.929 | | 2.19 |
| Year\*Schooling | 0.423 | | 2.67 |

We have a few VIFs which are greater than 10.

* This was not mitigated by standardizing predictors using subtract the means.
* Removing GDP and under-five deaths improved the VIF for all the factors and interactions, but they made the fit a bit worse. So, I chose to stick with the current model.

**Model Summary**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **S** | **R-sq** | **R-sq(adj)** | **PRESS** | **R-sq(pred)** | **AICc** | **BIC** |
| 3.36176 | 85.85% | 85.40% | 19589.9 | 84.64% | 8735.12 | 9018.15 |

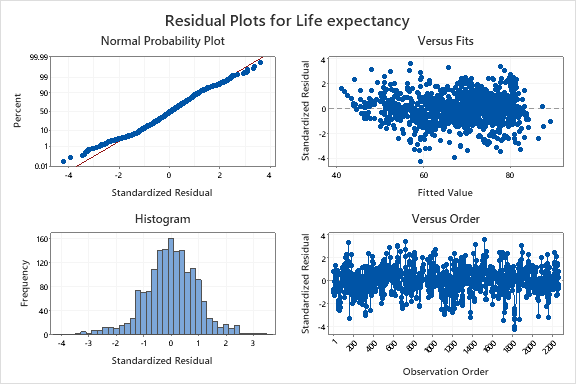
* The model fit of 85% is fairly better than the previous model with no interactions. The PRESS value decreased a bit too.

**Analysis of Variance**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Seq SS** | **Contribution** | **Adj SS** | **Adj MS** | **F-Value** | **P-Value** |
| Regression | 51 | 109481 | 85.85% | 109481 | 2146.68 | 189.95 | 0.000 |
| Adult Mortality | 1 | 62941 | 49.35% | 3341 | 3341.08 | 295.63 | 0.000 |
| Alcohol | 1 | 10272 | 8.05% | 46 | 45.72 | 4.05 | 0.044 |
| BMI | 1 | 7199 | 5.64% | 388 | 388.13 | 34.34 | 0.000 |
| Total expenditure | 1 | 127 | 0.10% | 63 | 62.55 | 5.53 | 0.019 |
| Diphtheria | 1 | 2442 | 1.92% | 48 | 47.67 | 4.22 | 0.040 |
| HIV/AIDS | 1 | 8515 | 6.68% | 1969 | 1968.76 | 174.20 | 0.000 |
| GDP | 1 | 2742 | 2.15% | 404 | 404.19 | 35.76 | 0.000 |
| Year | 1 | 11 | 0.01% | 220 | 219.99 | 19.47 | 0.000 |
| Hepatitis B | 1 | 2 | 0.00% | 0 | 0.41 | 0.04 | 0.849 |
| Measles | 1 | 7 | 0.01% | 61 | 61.16 | 5.41 | 0.020 |
| under-five deaths | 1 | 413 | 0.32% | 48 | 47.93 | 4.24 | 0.040 |
| Polio | 1 | 259 | 0.20% | 12 | 12.23 | 1.08 | 0.298 |
| Population | 1 | 228 | 0.18% | 1 | 0.75 | 0.07 | 0.797 |
| thinness 5-9 years | 1 | 125 | 0.10% | 70 | 70.21 | 6.21 | 0.013 |
| Schooling | 1 | 8408 | 6.59% | 3156 | 3156.08 | 279.26 | 0.000 |
| Adult Mortality\*Alcohol | 1 | 104 | 0.08% | 104 | 104.11 | 9.21 | 0.002 |
| Adult Mortality\*BMI | 1 | 26 | 0.02% | 2 | 2.22 | 0.20 | 0.658 |
| Adult Mortality\*under-five deaths | 1 | 1 | 0.00% | 44 | 44.15 | 3.91 | 0.048 |
| Adult Mortality\*Diphtheria | 1 | 57 | 0.04% | 43 | 42.93 | 3.80 | 0.051 |
| Adult Mortality\*HIV/AIDS | 1 | 1739 | 1.36% | 461 | 460.85 | 40.78 | 0.000 |
| Adult Mortality\*GDP | 1 | 102 | 0.08% | 40 | 39.94 | 3.53 | 0.060 |
| Adult Mortality\*Schooling | 1 | 138 | 0.11% | 85 | 85.33 | 7.55 | 0.006 |
| Year\*Adult Mortality | 1 | 253 | 0.20% | 50 | 50.24 | 4.45 | 0.035 |
| Alcohol\*BMI | 1 | 32 | 0.02% | 196 | 196.36 | 17.37 | 0.000 |
| Alcohol\*under-five deaths | 1 | 314 | 0.25% | 1 | 1.23 | 0.11 | 0.741 |
| Alcohol\*Diphtheria | 1 | 37 | 0.03% | 5 | 5.21 | 0.46 | 0.497 |
| Alcohol\*HIV/AIDS | 1 | 475 | 0.37% | 230 | 230.08 | 20.36 | 0.000 |
| Alcohol\*GDP | 1 | 82 | 0.06% | 123 | 122.73 | 10.86 | 0.001 |
| Alcohol\*Schooling | 1 | 42 | 0.03% | 71 | 71.34 | 6.31 | 0.012 |
| Year\*Alcohol | 1 | 147 | 0.11% | 52 | 51.91 | 4.59 | 0.032 |
| BMI\*under-five deaths | 1 | 0 | 0.00% | 28 | 27.64 | 2.45 | 0.118 |
| BMI\*Diphtheria | 1 | 451 | 0.35% | 148 | 147.53 | 13.05 | 0.000 |
| BMI\*HIV/AIDS | 1 | 16 | 0.01% | 44 | 44.18 | 3.91 | 0.048 |
| BMI\*GDP | 1 | 5 | 0.00% | 31 | 31.25 | 2.77 | 0.097 |
| BMI\*Schooling | 1 | 892 | 0.70% | 764 | 764.14 | 67.61 | 0.000 |
| Year\*BMI | 1 | 96 | 0.08% | 89 | 88.67 | 7.85 | 0.005 |
| under-five deaths\*Diphtheria | 1 | 7 | 0.01% | 0 | 0.00 | 0.00 | 0.993 |
| under-five deaths\*HIV/AIDS | 1 | 201 | 0.16% | 181 | 181.29 | 16.04 | 0.000 |
| under-five deaths\*GDP | 1 | 1 | 0.00% | 1 | 0.62 | 0.05 | 0.815 |
| under-five deaths\*Schooling | 1 | 35 | 0.03% | 53 | 52.87 | 4.68 | 0.031 |
| Year\*under-five deaths | 1 | 29 | 0.02% | 24 | 23.91 | 2.12 | 0.146 |
| Diphtheria\*HIV/AIDS | 1 | 7 | 0.01% | 6 | 6.50 | 0.57 | 0.448 |
| Diphtheria\*GDP | 1 | 21 | 0.02% | 20 | 20.08 | 1.78 | 0.183 |
| Diphtheria\*Schooling | 1 | 12 | 0.01% | 10 | 9.52 | 0.84 | 0.359 |
| Year\*Diphtheria | 1 | 18 | 0.01% | 19 | 18.84 | 1.67 | 0.197 |
| HIV/AIDS\*GDP | 1 | 25 | 0.02% | 0 | 0.19 | 0.02 | 0.898 |
| HIV/AIDS\*Schooling | 1 | 93 | 0.07% | 41 | 41.43 | 3.67 | 0.056 |
| Year\*HIV/AIDS | 1 | 66 | 0.05% | 64 | 63.55 | 5.62 | 0.018 |
| GDP\*Schooling | 1 | 260 | 0.20% | 254 | 254.13 | 22.49 | 0.000 |
| Year\*GDP | 1 | 0 | 0.00% | 0 | 0.09 | 0.01 | 0.929 |
| Year\*Schooling | 1 | 7 | 0.01% | 7 | 7.26 | 0.64 | 0.423 |
| Error | 1597 | 18048 | 14.15% | 18048 | 11.30 |  |  |
| Total | 1648 | 127529 | 100.00% |  |  |  |  |

From the table above, we can show that

* **Adult mortality** has a significant interaction **with Alcohol consumption, under-five deaths, HIV/AIDS deaths/1000 live births, Schooling years and Year**.
* **Alcohol** has a significant interaction with **BMI, HIV/AIDS deaths/1000 live births, GDP, Schooling years and Year.**
* **BMI** has a significant interaction with **Diphtheria immunization coverage, HIV/AIDS deaths/1000 live births, Schooling Years and Year.**
* **Under-five deaths** has a significant interaction with **HIV/AIDS deaths/1000 live births and Schooling years.**
* **HIV/AIDS** has a significant interaction with **Year**.
* **GDP** has a significant interaction with **Schooling Years**.
* Also, **Total expenditure, Measles immunization coverage and thinness for children aged 5-9 years** has an effect on **Life Expectancy.**
* The effect of other factors on Life expectancy depends on the interaction between them as mentioned above.



The model is adequate as the residuals are mostly normally distributed and there are no unusual patterns in the versus fits plot.

*Note: I thought about using Added Variable Plots, Principal Component Analysis and Partial Least Squares regression, Best Subsets Regression. At one point I did add them to this report but eventually I found the simpler model with interactions to be the most appropriate model as all the other techniques gave me results as good or worse than the model above.*

# Section 5: Conclusion

Final conclusion: From the above model, we can say that Life Expectancy is a linear function of Adult mortality, Alcohol consumption, BMI, under-five deaths, HIV/AIDS deaths/1000 live births, Schooling years, GDP, HIV/AIDS deaths/1000 live births, Total expenditure, Measles immunization coverage and thinness for children aged 5-9 years and Year.

It means to improve Life Expectancy, a country must

* reduce the Adult mortality, Alcohol consumption, under-five deaths, deaths by HIV/AIDS, prevalence of thinness among children aged 5-9 years and/or
* increase Schooling years, GDP, Total expenditure on health, Measles immunization coverage**.**
* **Factors such as Adult mortality, alcohol consumption, under five deaths, GDP, Schooling years have some level of interdependence. Improving or preventing one factor will not directly improve Life Expectancy and thus they must be tweaked based on the other factors.**
* The factors are independent of country.

# Section 6: References

Books: INTRODUCTION TO LINEAR REGRESSION ANALYSIS by Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining.

Articles and Webpages:

<https://statisticsbyjim.com/regression/interaction-effects/>

<https://www.researchgate.net/post/Is_it_a_problem_to_have_multicollinearity_with_interactions>

<https://support.minitab.com/en-us/minitab/18/help-and-how-to/calculations-data-generation-and-matrices/calculator/calculator-functions/text-calculator-functions/replace-and-substitute-functions/>

<https://support.minitab.com/en-us/minitab/18/help-and-how-to/modeling-statistics/multivariate/how-to/principal-components/before-you-start/data-considerations/>

<https://blog.minitab.com/blog/understanding-statistics/handling-multicollinearity-in-regression-analysis>

<https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/>

<https://support.minitab.com/en-us/minitab/19/macro-library/macro-files/anova-and-regression-macros/avplots/>

# Section 7: Appendices

