Multiple Linear Regression Analysis of Worldwide Life Expectancy

Maxwell Chu

2024-03-16

Introduction

The Life Expectancy (WHO) dataset is a public dataset on kaggle.com made using public data from the World Health Organization and United Nations websites focusing on the life expectancy in all 193 UN member countries. Data for each country from the year 2000 to 2015 are included. The dataset includes variables specifically related to immunization and mortality rates, as well as economic and social factors. Performing regression analysis on this dataset may provide useful statistical insights into the factors which contribute to worldwide life expectancy, ultimately establishing grounds for practical initiatives in increasing life expectancy.

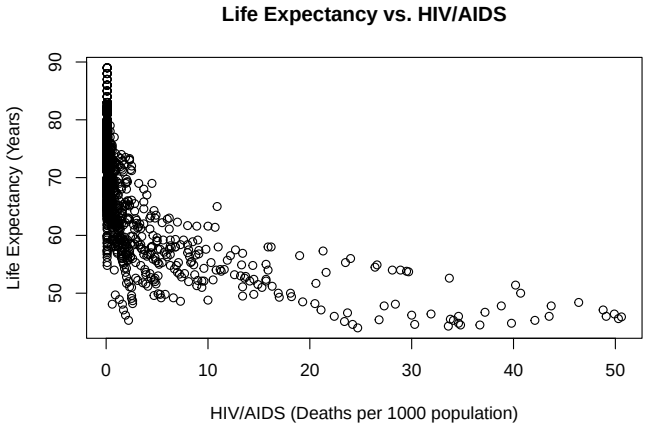
Variable Definitions

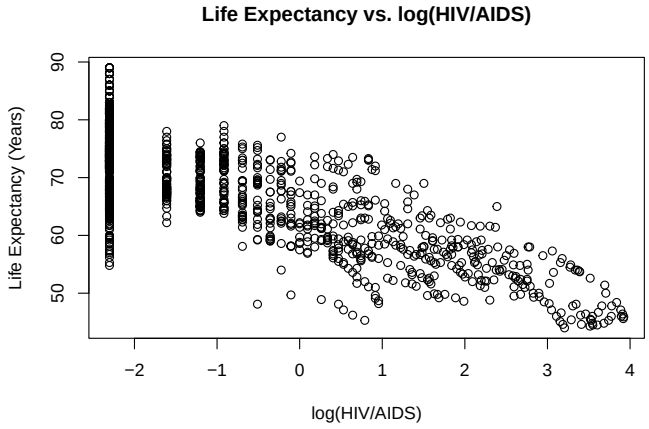
Each variable is defined on kaggle as follows:

1. Country: A categorical variable of the 193 UN member countries. Names the country from which data was collected.
2. Year: Data was collected from 2000 to 2015 for every country.
3. Status: A binary variable denoting whether the given country is developing or developed.
4. Life Expectancy: The life expectancy given by age in years. **This is the response variable.**
5. Adult Mortality: The number of people per 1000 population who die between 15 and 60 years of age.
6. Infant Deaths: The number of infant deaths per 1000 population.
7. Alcohol: Per capita consumption in litres of pure alcohol.
8. Percentage Expenditure: Expenditure on health care as a percentage of GDP per capita.
9. Hepatitis B: Immunization coverage for Hepatitis B among one-year-olds.
10. Measles: Number of reported Measles cases.
11. BMI: Average Body Mass Index of entire population. (This is calculated as weight in kilograms divided by the square of height in meters.)
12. Under-Five Deaths: Number of deaths of children under five years of age per 1000 population.
13. Polio: Immunization coverage for Polio among one-year-olds.
14. Total Expenditure: Government expenditure on health care as a percentage of total gov. expenditure.
15. Diphtheria: Immunization coverage for Diphtheria, Tetanus, and Pertussis among one-year-olds.
16. HIV/AIDS: Deaths from HIV/AIDS per 1000 live births, for children aged 0 to 4 years.
17. GDP: Gross Domestic Product per capita in USD.
18. Population: Population of the given country.
19. Thinness 1-19 Years: Prevalence of thinness in percentage of children aged 10 to 19 years.
20. Thinness 5-9 Years: Prevalence of thinness in percentage of chldren aged 5 to 9 years.
21. Income Composition of Resources: Human Development Index in terms of income composition of resources; a higher number means more equality of the income distribution.
22. Schooling: Average number of years of schooling.

Data Cleaning and Pre-Processing

- The columns were renamed to follow consistent syntax.
- Status variable was made numerical, where "Developed" = 1 and "Developed" = 2.
- Data for the BMI variable was evidently unreliable because 6.5% of observations were below 6, which is fatally low for human beings, especially considering that this represents the average BMI of the entire population. Data for the Population variable was also evidently unreliable because almost every country showed numerical disparities in consecutive years orders of magnitude in severity. Both the BMI and Population variables were removed from use in this report. These faulty data were found by regressing Life_Expectancy on each independent variable and simply perusing the dataset. As a result, all independent variables besides BMI and Population were verified to be valid data.
- Removed any observation for which one value was NA. The number of observations decreased from 2938 to 1657 after this operation.
- The plots for Life_Expectancy versus Infant_Deaths and Life_Expectancy versus HIV_AIDS each showed a scatterplot of values in which most points were bunched near x = 0, so the two independent variables were transformed by taking the logarithm with base e of all values. The resultant relationships between the response and the two predictors became more linear and easier to incorporate into a linear regression model. Testing showed that the adjusted R-squared increased with each of these transformations isolated.





The right plot above depicts Life_Expectancy versus log(HIV_AIDS) after the transformation of HIV_AIDS and clearly shows a more linear relationship between the predictor and response than the left plot. A similar difference may be observed between the pre- and post-transformation plots for Life_Expectancy versus Infant_Deaths (not shown).

Descriptive Statistics

Table 1: Summary Statistics for the Dataset

Statistic	N	Mean	St. Dev.	Min	Max
Year	1,657	2,007.851	4.087	2,000	2,015
Status	1,657	1.854	0.353	1	2
Life_Expectancy	1,657	69.300	8.781	44.000	89.000
Adult_Mortality	1,657	168.074	125.151	1	723
Infant_Deaths	1,657	1.847	1.645	0.000	7.378
Alcohol	1,657	4.513	4.030	0.010	17.870
Percentage_Expenditure	1,657	695.599	1,755.644	0.000	18,961.350
Hepatitis_B	1,657	79.162	25.577	2	99
Measles	1,657	2,214.616	10,062.430	0	131,441
Under_Five_Deaths	1,657	44.066	162.520	0	2,100
Polio	1,657	83.422	22.592	3	99
Total_Expenditure	1,657	5.944	2.301	0.740	14.390
Diphtheria	1,657	83.973	21.763	2	99
HIV_AIDS	1,657	−1.202	1.622	−2.303	3.924
GDP	1,657	14,605,687.000	70,295,548.000	34.000	1,293,859,294.000
Thinness_1_19_Years	1,657	4.857	4.589	0.100	27.200
Thinness_5_9_Years	1,657	4.913	4.643	0.100	28.200
Income_Composition_of_Resources	1,657	0.632	0.183	0.000	0.936
Schooling	1,657	12.116	2.791	4.200	20.700

Stepwise Linear Regression Model

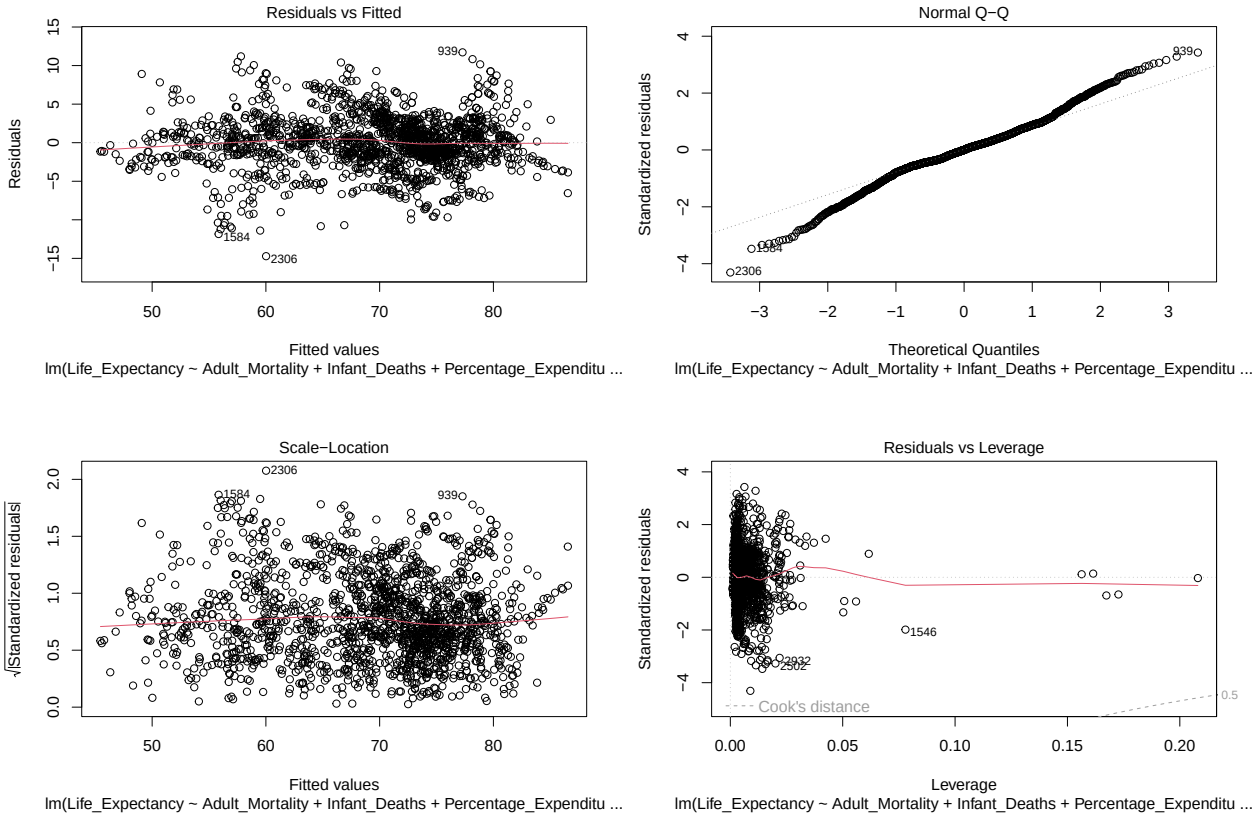
According to the results of stepwise regression, the model with the lowest AIC includes ten predictors as follows:

Table 2: Stepwise Linear Regression Model Results

	Dependent variable:
	Life_Expectancy
Adult_Mortality	−0.015*** (0.001)
Infant_Deaths	−0.456*** (0.067)
Percentage_Expenditure	0.0004*** (0.0001)
Total_Expenditure	0.090** (0.039)
Diphtheria	0.013*** (0.004)
HIV_AIDS	−2.233*** (0.074)
GDP	0.000** (0.000)
Thinness_5_9_Years	−0.097*** (0.022)
Income_Composition_of_Resources	10.069*** (0.780)
Schooling	0.454*** (0.055)
Constant	56.571*** (0.694)
Observations	1,657
R^2	0.848
Adjusted R^2	0.847
Residual Std. Error	3.430 (df = 1646)
F Statistic	920.623*** (df = 10; 1646)
Note:	*p<0.1; **p<0.05; ***p<0.01

For each predictor variable on the left of the table, the estimated coefficient, significance indicator, and standard error are displayed to the right, in order. All of the model's predictors are statistically significant, each with a sufficiently low p-value derived from a T-test, as denoted by the asterisks next to each estimated coefficient. The model observed an adjusted R^2 of 0.847, so the model was able to explain 84.7% of the variance in the response variable Life_Expectancy. This is a high figure and suggests that the model has high explanatory power for the dataset. Finally, the statistical significance of the F-statistic derived from an F-test indicates that at least some of the variance in Life_Expectancy was explained by the independent variables and there is a relationship between them.

Checking Model Assumptions



The plot of residuals versus fitted values for the stepwise linear regression model displays constant variance (homoscedasticity) and no pattern in the residuals, which suggests that the model was able to capture an underlying pattern in the dataset. The plot of standardized residuals versus fitted values is similarly verifies the validity of the model assumptions.

The Q-Q plot above shows that the errors are nearly consistent with the normal distribution, since it follows a fairly straight line. Finally, the plot of standardized residuals versus leverage indicates that almost all samples with high leverage are good leverage points, as their errors are near zero. Moreover, no points are anywhere near the Cook's distance threshold, suggesting that there are few outliers and the impact of those few on the model is minimal.

Conclusions

Since the regression model indicates that a high proportion of the total variance in the response Life_Expectancy is explained by the predictors, we may infer how each of the predictors in the model affected the response.

- The negative coefficients for Adult_Mortality and Infant_Deaths in the final model suggest that they have an inverse relationship with the response Life_Expectancy. This makes sense mathematically, as the more people who die at a younger age, the less time people are expected to live. Higher adult mortality and infant deaths may also point to a larger problem that causes more people to die, such as lack of immunization coverage or general health care deficits.

- The positive coefficients for Percentage_Expenditure and Total_Expenditure in the final model suggest that the more the government spends on health care, the higher life expectancy becomes. This makes sense since health care in general would improve, allowing people more accessbility to and higher quality coverage.

- Diphtheria also has a positive coefficient, which indicates that as more babies are immunized from diphtheria, tetanus, and pertussis, life expectancy increases. This may speak to the effectiveness of immunization to save lives and allow babies to grow up, ultimately contributing to a higher life expectancy.

- HIV_AIDS has a negative coefficient. As more people die from HIV/AIDS, life expectancy decreases since people on average live for less time.

- Thinness_5_9_Years also has a negative coefficient. As more children are deemed thin, likely due to malnutrition, more health issues may arise for them, causing them to die earlier than they would otherwise, thereby lowering the life expectancy.

- Income_Composition_of_Resources and Schooling both have positive coefficients. The more equal the income distribution is, the more people have means to obtain paid health care, food, and other beneficial things that contribute to a higher life expectancy. Meanwhile, schooling could be speculated to provide useful skills and learning to people, who then leverage their abilities and knowledge to secure a higher standard of living than they would have being completely uneducated. According to the Human Development Index, a better standard of living in general heightens the life expectancy of a country.

Overall, this linear regression model and report confirms the importance of decreasing mortality rates, illness, and malnutrition in a population to improve life expectancy. It is also vital to increase health care expenditure, quality, and accessibility, as well as immunization coverage, income composition equality, and schooling. All of these initiatives have different immediate effects but are ultimately conducive a higher life expectancy.