Predicting Life Expectancy Using Global Data in 2015

May 1, 2024

Abstract
Life expectancy can be used as an indicator of people's overall quality of life in a nation. Even though previous studies researched the key factors that influence the average life expectancy of nations, the data were unrepresentative of the whole world. Hence, we analyzed the 2015 data from 179 nations, including both developed and developing nations, to figure out crucial factors that determine a nation's average life expectancy. Specifically, we employed a multiple linear regression model and found that BMI, childhood and adolescence thinness, vaccination coverage, education level, and GDP were significant explanatory variables for predicting a nation's average life expectancy. Based on the result, we suggest that governments should implement policies that encourage healthy eating habits to control BMI and prevent thinness during childhood and adolescence, provide essential vaccinations, and increase educational funding to elevate the overall educational level for their people.

Introduction

Due to technological advancements, especially in medical science, the average life expectancy in the world has been steadily increasing. In 2023, the national life expectancy was reported to range from a high of 89.6 years (in Monaco) to a low of 54.1 years (in Afghanistan) (CIA 2023). Despite the increasing trend worldwide, a large discrepancy in life expectancy exists among nations. An International Journal of Health Economics and Management study, which conducted statistical inference on 36 Organization for Economic Co-operation and Development (OECD) nations, concluded that two main factors contributed to population life expectancy. These two factors were per capita healthcare expenditure and the quality of the healthcare system, which they measured by the number of available physicians per capita (Roffia 2023). Even though other factors, including GDP, population size, population behavior, and weather, also bear some impact on life expectancy, these factors are either uncontrollable or more difficult to address, as they require a long-term plan or increased effort to result in an improvement in life expectancy.

Another study investigated the determinants of life expectancy in 91 developing countries (Kabir 2008). The study concluded that the improvement of socioeconomic factors, such as per capita income, education, health expenditures, access to safe water, and urbanization, did not necessarily lead to increased life expectancy, which contradicts results from previous studies and further supports the International Journal study discussed previously. To increase life expectancy, the study suggested that governments should instead strategize policies to increase physicians' availability and alleviate adult illiteracy and undernourishment to increase life expectancy.

While both studies showed that the medical system is the key factor influencing a nation's life expectancy, the sample used in these studies seemed to be unrepresentative of the world as a whole. Specifically, the former study only used 36 OECD nations as its sample, and the latter study used 91 developing nations, excluding developed nations. This paper incorporates data from 179 nations (including both developed and developing countries) in 2015, analyzes the relationship between various explanatory variables and life expectancy, and compares the results with those of previous studies. Since life expectancy is one of the indicators of overall quality of life, this paper may provide a useful guideline on how governments should allocate budgets and implement policies to enhance the general standard of living.

Methods

The data in our dataset were originally from World Bank Data, the World Health Organization, and the University of Oxford. Population, Gross Domestic Product (GDP), and life expectancy were collected from World Bank Data (Gochiashvili 2023). Health-related data, including Body-Mass-Index (BMI), vaccinations, average annual alcohol consumption, mortality rates, and thinness, were collected from WHO public datasets. Finally, data regarding average educational level were collected from the University of Oxford. After refinements and data cleaning, this dataset was uploaded to Kaggle by a data scientist named Lasha Gochiashvili in 2023 (Gochiashvili 2023).

Our dataset included subjects from 179 nations, whose information was recorded from 2000 to 2015, and we decided to use the most recent year's dataset. In this analysis, we wish to predict life expectancy using predictor variables such as BMI, GDP per capita, thinness among children and adolescents, schooling, vaccinations, and population (Gochiashvili 2023). For these predictor variables, the population was measured in millions, GDP was measured in per capita converted to US dollars, and vaccinations (Hepatitis B, Measles, Polio, and Diphtheria) were measured as the proportion of the population that received each of these vaccinations. BMI, which can be measured as weight (in kilograms) divided by height (in meters) squared, indicates the average BMI of the population of each nation. Thinness (five to nine years and ten to nineteen years) measured the proportion of children from five to nine years old and ten to nineteen years old whose BMIs were two standard deviations below the average (World Health Organization). Schooling was measured as the average number of school years the population spent, indicating

a nation's overall educational level. Lastly, life expectancy was measured as the average period one expects to live in years (Gochiashvili 2023). Since infant, premature, and adult mortality rates overlap with the measurement of life expectancy and the incidence of HIV was insignificant (less than 1% from our dataset), we excluded these predictor variables from our data analysis.

Results

A multiple linear regression model was employed for the data analysis since the response variable (life expectancy) is continuous and quantitative and there are more than one potential predictor variables for the model. The purpose of using this model was to figure out which combination of explanatory variables produced a strong but simple regression model for predicting the life expectancy of each nation.

First, we examined the interaction effect in the predictor variables. Specifically, a conjecture was made that there might exist interaction effects between thinness from 5 to 9 years old and thinness from 10 to 19 years old, the coverage of Hepatitis B and Measles vaccinations, the coverage of Hepatitis B and Polio vaccinations, and the coverage of Measles and Polio vaccinations. The rationale for this conjecture was that being significantly underweight from 5 to 9 years old would not affect life expectancy to a great extent, but if this status continues up to 19 years old, then this may significantly decrease life expectancy. Moreover, although one vaccination may not considerably influence one's life expectancy, a series of vaccinations together may boost it. To test whether these interaction effects exist, interaction plots were plotted in Figures 1, 2, 3, and 4. From these interaction plots, non-parallel, intersecting lines on the interaction plot of thinness from 5 to 9 and 10 to 19 indicated that there exists an interaction between these two predictor variables. Indeed, the interaction term in the regression model turned out to be significant, and, thus, was included in the regression model along with the corresponding two predictor variables, even though the thinness from 5 to 9 years old turned out to be not significant. Parallel lines were presented on the interaction plot of the coverage of Hepatitis B and Measles vaccinations, indicating that seldom interaction exists between these two predictor variables. The rest of the interaction plots regarding vaccinations displayed non-parallel and intersecting lines, indicating some interactions. However, with a 5% significance level, all the interaction terms regarding vaccinations turned out to be not significant, and, thereby, were omitted in the regression model.

Next, we conducted a regression analysis on the dataset, using the backward model selection based on AIC for the variable selection. The result shows that BMI, the coverage of Measles and Polio vaccinations, Schooling, GDP per capita, and the interaction between thinness from 5 to 9 years old and from 10 to 19 years old should be included to yield the lowest AIC, whereas the other predictor variables, such as the number of population and the amount of alcohol consumption, should be removed. After the variable selection, the following equation was deduced as the best-fitting model according to the variable selection using AIC and all the predictor variables turned out to be significant except for the thinness from 5 to 9 years old.

$$y_i = 37.66 + 0.4359 * BMI + 0.6527 * Measles + 0.1420 * Polio + 0.0001149 * GDP_{percapita} + 0.6504 * Schooling - 1.069 * Thinness_{10-19} + 0.2985 * Thinness_{5-9} + 0.03916 * Thinness_{10-19} * Thinness_{5-9} + \epsilon_i$$

Before adopting this model, we tested whether this regression model satisfied the following four assumptions. First, we tested whether the residuals were approximately normally distributed. The Shapiro-Wilk test suggested that the residuals are not normally distributed with a low p-value of 0.003. Hence, we undertook possible transformations on the response variable, including taking log, square root, and square. We found out that squaring the response variable resolves the non-normality issue while producing an increased p-value of 0.09956. After undertaking the squared transformation, the QQ plot and QQ line confirmed that the residuals are approximately normally distributed, substantiated by the straight-line pattern in Figure 5.

Next, we used the diagnostic plot of fitted values vs. residuals to detect if severe heteroscedasticity exists and if the residuals are distributed around 0. The diagnostic plot of the regression model was created as in Figure 6. As shown in the diagnostic plot, the residuals were mostly centered around 0 and were randomly scattered throughout the plot, indicating that the residuals were approximately centered around 0 with roughly constant variance.

Finally, to check the existence of a significant correlation among the residuals, we conducted the Durbin-Watson test. The result showed that there are no notable correlations among the residuals with a test statistic of 2.0161 and a high p-value of 0.5537. Conclusively, the four basic assumptions of the regression model were met after the squared transformation.

We also checked whether there were outliers and/or influential points in our dataset. For detecting the outliers, we tested whether the predicted value was more than three standard deviations away from the actual value for each data point. Furthermore, for detecting the influential points, we checked whether Cook's distance was greater than the 50th percentile of the F distribution with 9 and 179 degrees of freedom for each data point. Consequently, there were two outliers (Eswatini and Lesotho) and one influential point that was not an outlier (India). Specifically, the z-scores for Eswatini and Lesotho were -3.29 and -3.86, respectively, and Cook's distance from India was 2.02. However, as there were no errors detected in the data measurement and collection, we kept these outliers and an influential point during data analysis.

Discussion/Conclusions

$$Life\_expectancy^2 = 866.4 + 49.35 * BMI + 9.539 * Measles + 17.99 * Polio + 0.01810 * GDP_{percapita} + 89.29 * Schooling - 138.7 * Thinness_{10-19} + 30.65 * Thinness_{5-9} + 5.326 * Thinness_{10-19} * Thinness_{5-9} + \epsilon_i$$

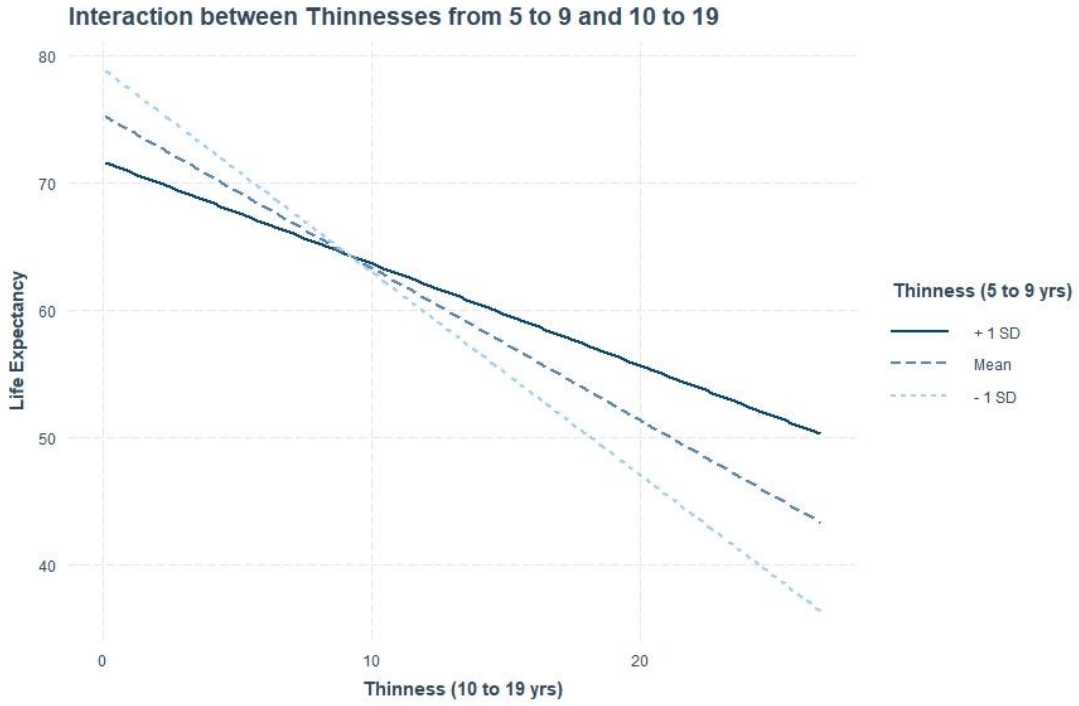
This multiple linear regression model suggested that predictor variables BMI, Measles, Polio, Schooling, Thinness from 10 to 19, and the interaction term between Thinness from 10 to 19 and 5 to 9 are statistically significant for predicting the life expectancy of each nation. Interestingly, education level was statistically significant and influential in predicting life expectancy, most likely because the average education level in a nation has a large impact on people's overall lifestyle. As people are more educated, they are more likely to know how to live a healthier life, helping them to avoid various health problems discussed in our paper. Indeed, from the above regression model, the square of life expectancy is expected to increase by 89.29, on average, for an additional year of education.

The multiple R-squared value was 0.7488 and the adjusted R-squared value was 0.737. In other words, our model was able to explain approximately 74.88% of the variation in the square of life expectancy by its linear relationships with the predictor variables. High multiple and adjusted R-squared values, which are over 0.7, suggest that the above model has strong predictive power for estimating a nation's average life expectancy.

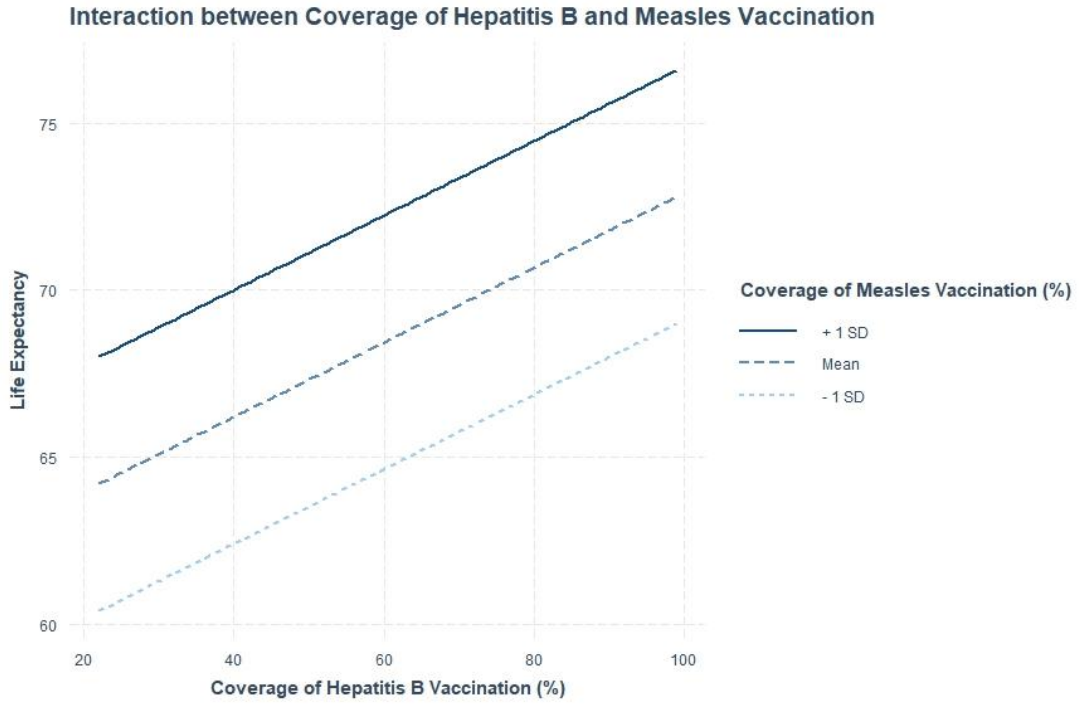
However, since we only used 2015 data for this project, the above model may not be representative of an extended timeframe. Due to the limitation of knowledge to deal with time series data, the problem regarding correlated residuals was not addressed in this paper. For future analysis, constructing a multiple linear regression for an extensive time horizon using an autoregression model may provide a better model to predict a nation's life expectancy.

The goal of increasing people's average life quality can be a challenging task for the government. One of the indicators of the overall quality of life is life expectancy, and the government may strive to increase this quantity for the wellness of its people. Based on our multiple linear regression model, we suggest that the government should implement policies that encourage healthy eating habits to control BMI, provide essential vaccinations, and increase educational funding to raise the overall educational level. Even though these efforts cannot be realized immediately, making gradual progress over the long term would help people enjoy a more prosperous life eventually.

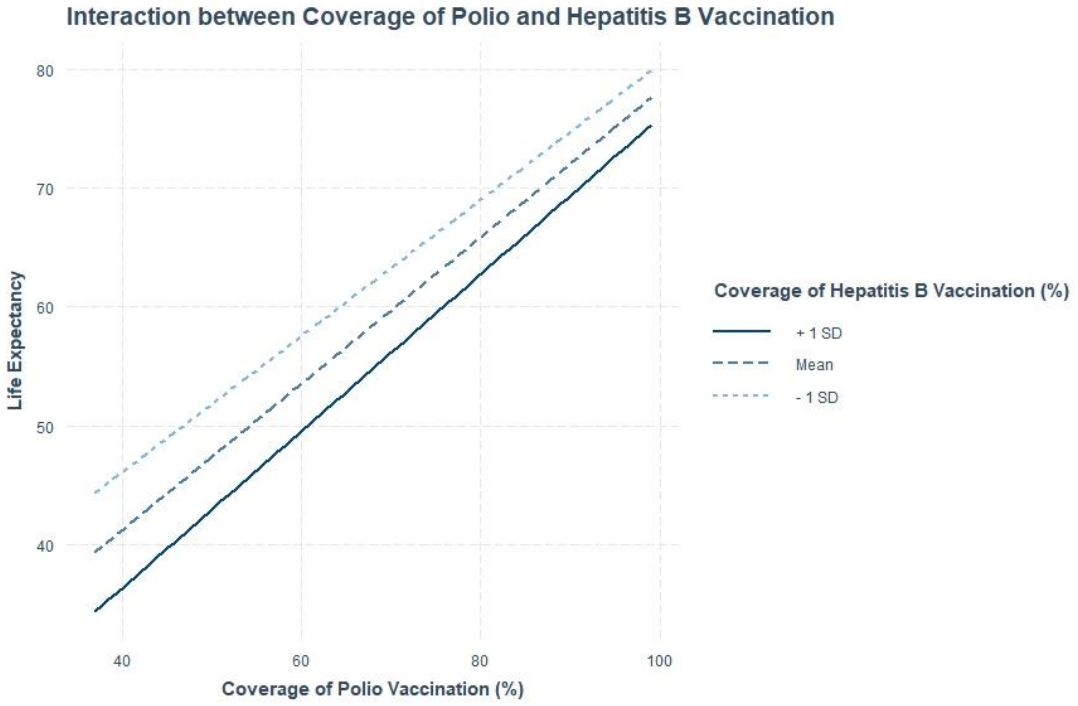
Appendix



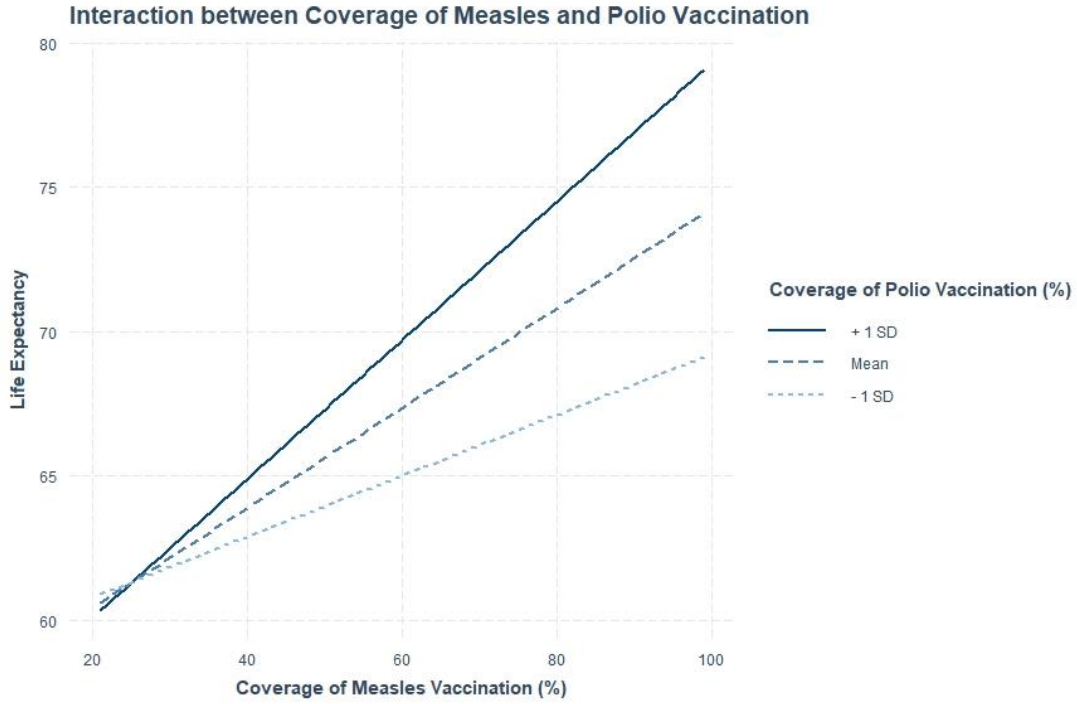
[Figure 1]



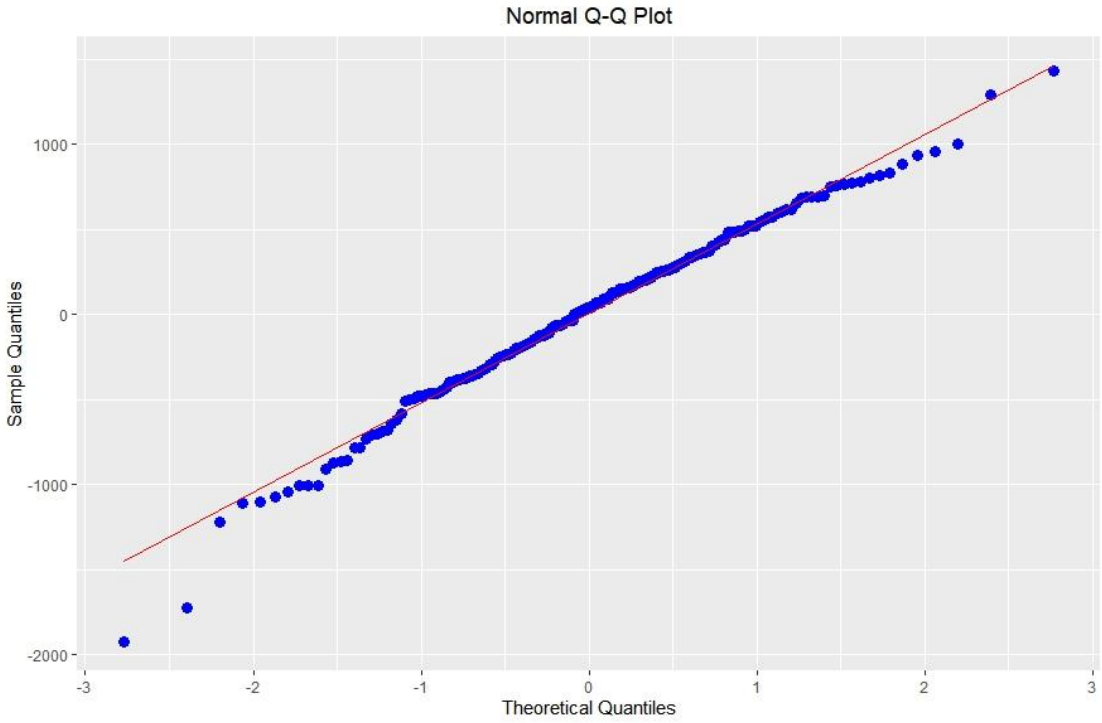
[Figure 2]



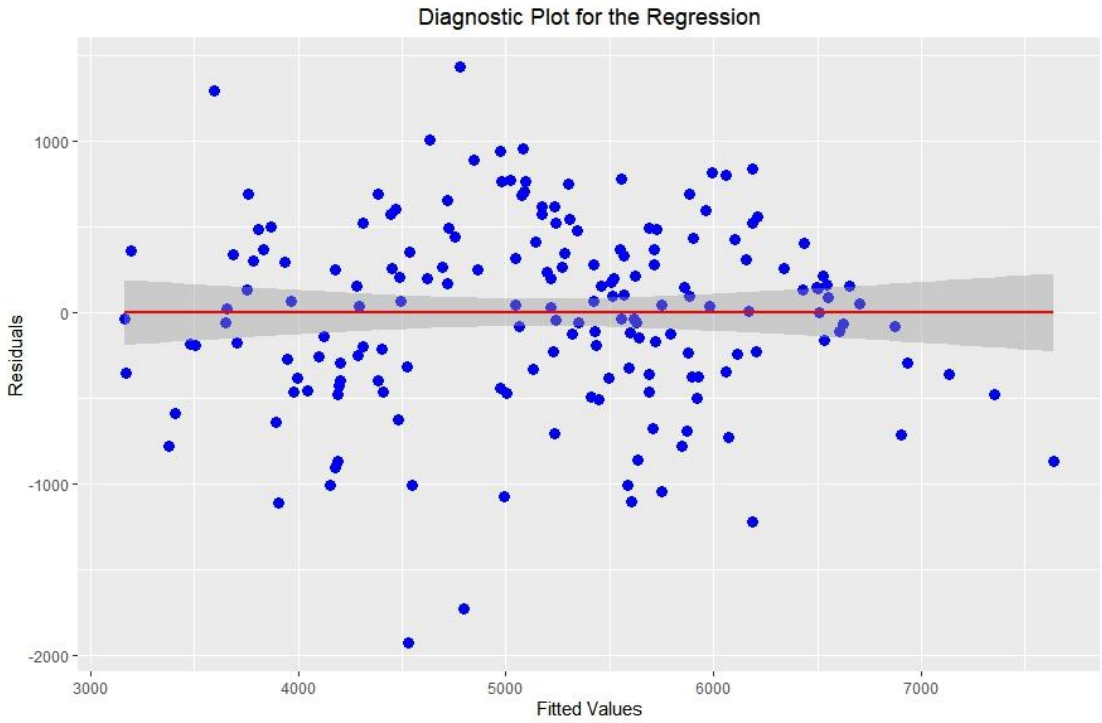
[Figure 3]



[Figure 4]



[Figure 5]



[Figure 6]

Works Cited

Central Intelligence Agency (CIA), 2023, www.cia.gov/the-world-factbook/field/life-expectancy-at-birth/country-comparison/.

Gochiashvili, Lasha. 2023, *Life expectancy (WHO) fixed*. Kaggle. https://www.kaggle.com/datasets/lashagoch/life-expectancy-who-updated/data.

Kabir, Mahfuz. "Determinants of Life Expectancy in Developing Countries." *The Journal of Developing Areas*, vol. 41 no. 2, 2008, p. 185-204. *Project MUSE,* https://doi.org/10.1353/jda.2008.0013.

Roffia, Paolo, et al. “Determinants of life expectancy at birth: a longitudinal study on OECD countries.” *International journal of health economics and management* vol. 23,2 (2023): 189-212. doi:10.1007/s10754-022-09338-5

World Health Organization. *Prevalence of thinness among children and adolescents, bmi < -2 standard deviations below the median (crude estimate) (%)*. World Health Organization. https://www.who.int/data/gho/data/indicators/indicator-details/GHO/prevaleNce-of-thinness-among-children-and-adolescents-bmi--2-standard-deviations-below-the-median-(crude-estimate)-(-).