Andrew Thompson

STAT-266

22 April 2024

Predicting Life Expectancy

Life expectancy is an important statistic that is calculated for many nations due to its significance in understanding standards of life and quality of resources. This value represents the expected lifespan of individuals in a certain demographic. Life expectancy can be calculated with many inputs, such as estimated mortality rates, overall health of a nation, healthcare quality, available resources, culture, lifestyle, diet, and numerous other variables, in order to create an accurate measure of expected longevity of life. Variables that may have significant relationships with life expectancy are commonly related to income, gross domestic product (GDP), and the financial wellbeing of a nation. These financial variables relate to the ability to afford quality healthcare or other impactful goods and services and, correspondingly, play a key role in wellbeing. It is possible for other variables also to have an impact on life expectancy, such as choices that impact the physical, social, and mental health of an individual.

The aim of this project is to determine the strength of the relationship between life expectancy and three input variables: alcohol consumption, schooling, and income composition of resources. Alcohol consumption refers to the number of liters consumed per capita, schooling refers to the standard number of years of schooling, and income composition of resources is rated 0 to 1 on effective utilization of resources using the human development index. A multiple linear regression is used to test the relationship between these variables and life expectancy.

The end goal of this multiple linear regression study is to determine the relationships between the three input variables consisting of alcohol consumption, schooling, and income composition of resources and the response variable, life expectancy. In the multiple linear regression equation, $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$, there are three parameters of interest, β_1, β_2, and β_3. These represent the relationship between life expectancy and the three variables. In the case of this multiple linear regression, the coefficient of each variable is not the same as it would be in a single linear regression. The statistical significance of each variable can be determined using the summary() and anova() functions. In this case, the slope coefficients (and intercept coefficient) are calculated by minimizing the sum of the squared residuals. The residual is the distance between the observed and predicted data values. This is more complex to understand than simple linear regression because it takes place in higher dimensions and can involve linear algebra. The independent variables observed were alcohol consumption per capita in liters, schooling in years, and income composition of resources (rated on the human development index) while the dependent (response) variable was predicted life expectancy. The null hypothesis of this study is that the slope coefficients of the relationship between those three variables and life expectancy is zero, which states that there is no relationship between the independent variables and the dependent variable. If the null hypothesis is true, then these variables do not provide useful information for predicting life expectancy. The alternative hypothesis states that there is statistically significant information that at least one slope coefficient is different from zero and is significantly associated with the dependent variable. This means the linear model can be used to predict life expectancy. This is a two-sided hypothesis test to determine if there is either a positive or negative relationship between the variables of interest.

This study utilized data collected by the World Health Organization (WHO). The data was made accessible to the public for purposes of health data analysis. The dataset, “Life Expectancy (WHO)” was found and retrieved under the name life_expectancy_data.csv from Kaggle; the data was provided by Kumar Rajarshi (Rajarshi, 2016). It provides information for the life expectancy of citizens in 193 countries (by year from 2000 to 2015) as well as various corresponding lifestyle variables for the countries. While there are numerous variables in the dataset, the most relevant items for this study are “Life.expectancy”, “Alcohol”, “Schooling”, and “Income.composition.of.resources”. These variables measured the life expectancies of specific countries by year and the alcohol consumption per capita of countries by liters. Upon removal of any observations with missing data, the data covered 133 countries. The variables included in the study’s hypothesis were used to test the beta slope coefficients and their significance. To do so, a linear model distributed life expectancy by the respective variables by reducing residuals in a higher-dimensional plane. The R function, summary(), was used to extract more information from the model and its calculations.

Descriptive statistics for variables deemed relevant in the life expectancy data are presented via graphs with corresponding analyses to determine more specific relationships and characteristics of the dataset. Afterward, the results of a simple linear regression analysis are displayed to conclude whether there is a significant linear correlation and a level of predictability between alcohol consumption per capita and life expectancy by country. All graph creation and data analysis were executed in R (version 4.3.1).

Results

Descriptive Statistics

Table 1 displays the quantitative variables in this dataset and their respective descriptive statistics. While some of the statistics do not provide much relevant information, it can be helpful to utilize descriptive statistics as a means of determining information about the data being studied. For example, this table provides the units of measurement for each quantitative variable. Some variables are measured in cases per 1000, while others are measured as a percentage or number of years. A few specific variables have their own measurement, such as alcohol consumption, which is measured in liters. Here, it is clear that certain observations, especially population, have extremely large ranges.

Table 1: *Comprehensive Descriptive Statistics Table of Relevant Quantitative Variables*

		min	max	median	IQR	mean	sd	skew
Year		2000	2015	2008	6	2007.84	4.09	-0.2
Life Expectancy		44	89	71.7	10.6	69.3	8.8	-0.63
Schooling	(number of years)	4.2	20.7	12.3	3.7	12.12	2.8	-0.13
BMI	(mean)	2	77.1	43.7	36.3	38.13	19.75	-0.23
Alcohol Consumption	(liters)	0.01	17.87	3.79	6.53	4.53	4.03	0.66
GDP	(USD)	1.68	119173	1592.57	4256.36	5566.03	11475.9	4.51
Infant Deaths		0	1600	3	21	32.55	120.85	8.46
Adult Mortality		1	723	148	150	168.22	125.31	1.27
Measles		0	131441	15	373	2224.49	10085.8	7.94
Under Five Deaths		0	2100	4	28	44.22	162.9	8.33
HIV AIDS	(per 1000)	0.1	50.6	0.1	0.6	1.98	6.03	4.97
Percentage Expenditure of GDP		0	18961.35	145.1	471.95	698.97	1759.23	4.97
Hepatitis B		2	99	89	22	79.22	25.6	-1.79
Polio		3	99	93	16	83.56	22.45	-2.36
Diphtheria		2	99	92	15	84.16	21.58	-2.48
Thinness 1-19 years		0.1	27.2	3	5.5	4.85	4.6	1.82
Thinness 5-9 years		0.1	28.2	3.2	5.4	4.91	4.65	1.86
Total Expenditure	(percent)	0.74	14.39	5.84	3.06	5.96	2.3	0.21
Income Composition of Resources	(0-1)	0	0.94	0.67	0.24	0.63	0.18	-1.15
Population		34	1293859294	1419631	7467075	14653625.9	70460393.4	14.16

As there was only one categorical variable for this dataset, a table was unjustified. The "Status" variable had only two categories: developed and developing. There were only 242

(17.4%) developed country observations in the dataset, while there were 1407 (82.6%) developing country observations. The graphs in Figure 1 provide a more in depth look at the categorical variable because they display the trend that developed countries generally place higher on scales that may be related to life expectancy. Developed countries tend to drink more per capita, have more hours of schooling, and have a higher placement on the human development index regarding income composition of resources. There are multiple possible explanations, ranging from access to safer alcohol, access to schooling, or better job opportunities. This distribution of the "Status" variable regarding alcohol consumption is important to keep in mind when looking at the upcoming multiple linear regression plot.

Figure 1: *Descriptive Statistics for Categorical Variable*

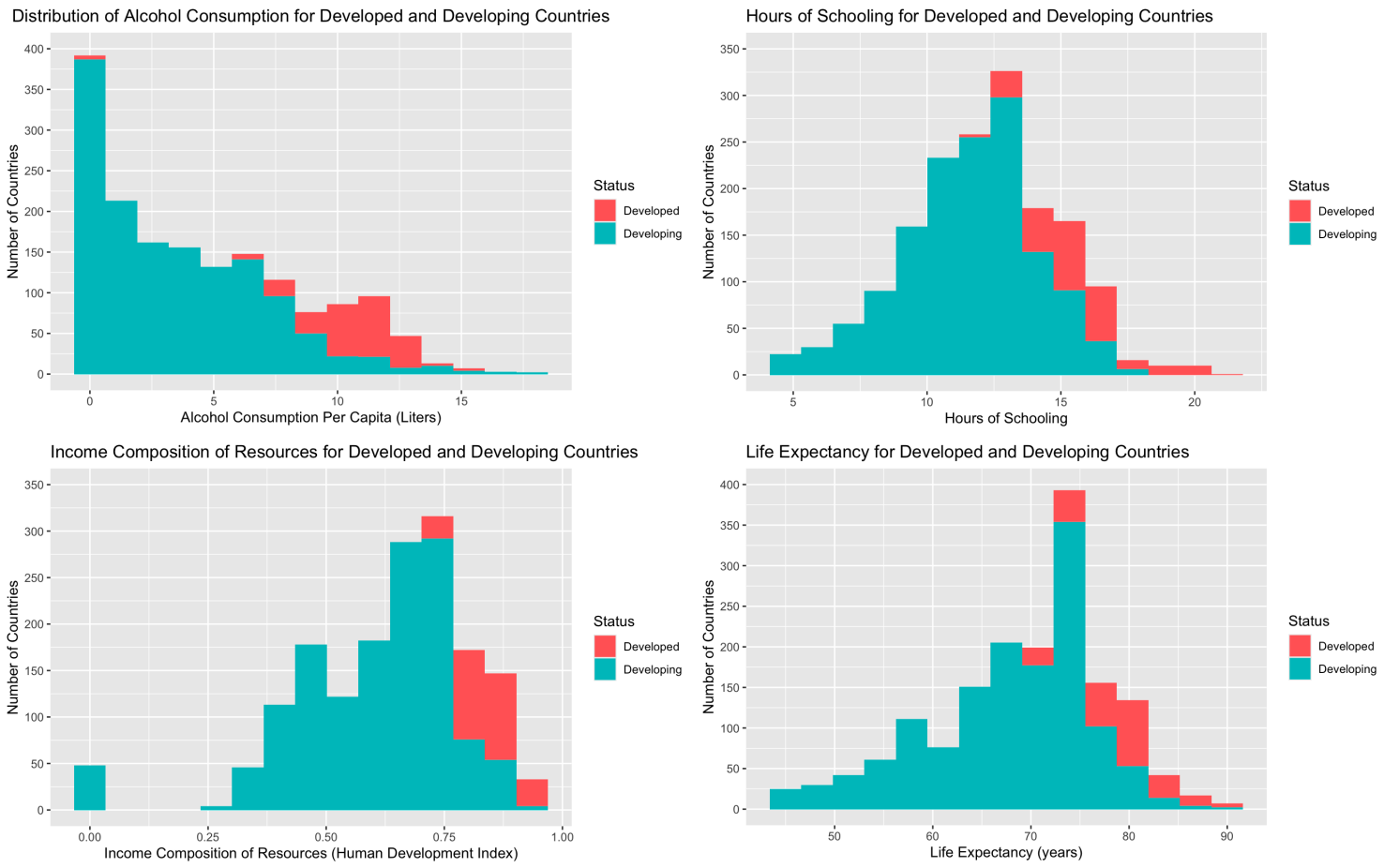


Table 2 holds a lot of information for this specific linear regression analysis. The ANOVA table is significant for determining *F*-statistics for each variable, and calculating the *p*-value from the *F*-test. This *p*-value helps to show which variable relationships are significant to the model and which are not. In the table, all three *p*-values are less than 0.001. This implies that the three variables used in the model are significant.

Table 2: *ANOVA Table Determining Coefficient Significance to Model*

Anova Table	Sum Sq	DF	Mean Sq	F value	Pr(>F)
Alcohol	20683	1	20683	663.17	< 2.2e-16
Schooling	47277	1	47277	1515.85	< 2.2e-16
Income Composition of Resources	8265	1	8265	265.01	< 2.2e-16
Residuals	51305	1645	31		

Contained in Table 3 is an organization of each variable in our dataset and its correlation value with regard to life expectancy. The correlations were calculated using the cor() function and were rounded to three digits for readability purposes. By taking note of the higher and lower values in the table, it is notable that certain variables do portray a much higher correlation to life expectancy than others. Variables with low correlations include “Year”, “Measles”, and “Population”, which implies that these variables have a weaker relationship with life expectancy. On the other hand, the correlation values imply that certain variables such as “Adult Mortality”, “Schooling”, “BMI”, and others have a strong correlation, and therefore relationship with life expectancy. This was influential in determining variables that could be effectively tested with hypothesis testing and linear regression to determine if the relationship was statistically significant and strongly correlated enough for the linear regression to display significant results.

Table 3: *Correlation Table Representing Each Variable's Correlation with Life Expectancy*

Variable Name	Correlation to Life Expectancy
Year	0.051
Life.expectancy	1.000
Adult.Mortality	-0.703
infant.deaths	-0.169
Alcohol	0.403
percentage.expenditure	0.410
Hepatitis.B	0.200
Measles	-0.069
BMI	0.542
under.five.deaths	-0.192
Polio	0.327
Total.expenditure	0.175
Diphtheria	0.341
HIV.AIDS	-0.592
GDP	0.441
Population	-0.022
thinness..1.19.years	-0.458
thinness.5.9.years	-0.458
Income.composition.of.resources	0.721
Schooling	0.728

Figure 2 displays the individual scatterplots for the three variables in this linear regression. The visualizations display the results of our fitted multiple linear regression model in terms of how the three predictor variables affect the outcome, years of life expectancy (Breheny, 2017). The relationship between alcohol and life expectancy is the only negative relationship, and appears to be the weakest, though it is still significant. The slope of the three plots represents the three coefficients in the multiple linear regression model, so the negative slope of the alcohol plot contrasts the positive slope that would be observed in a single linear regression of alcohol and life expectancy. This is one of the three slope coefficients that will be tested against a null hypothesis at the 5% significance level.

Figure 2: *Life Expectancy by Alcohol, Schooling, and Income Composition of Resources*

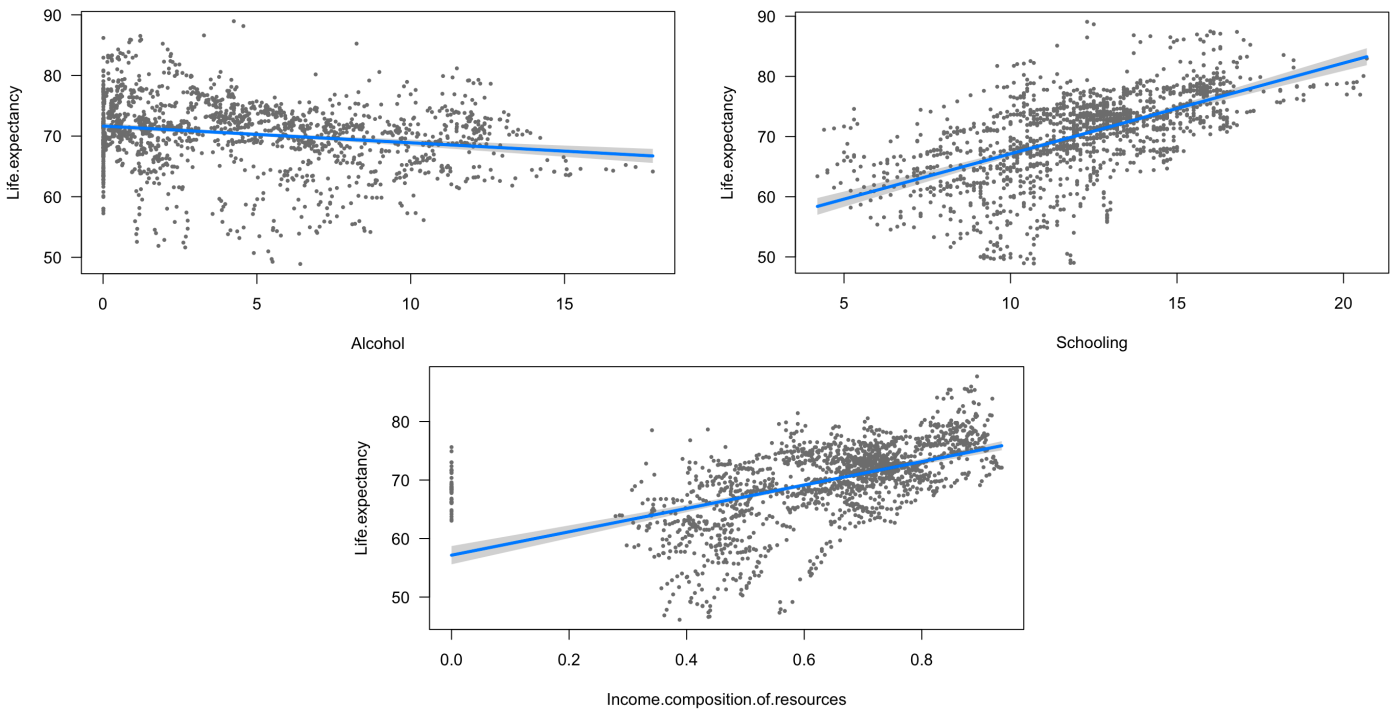
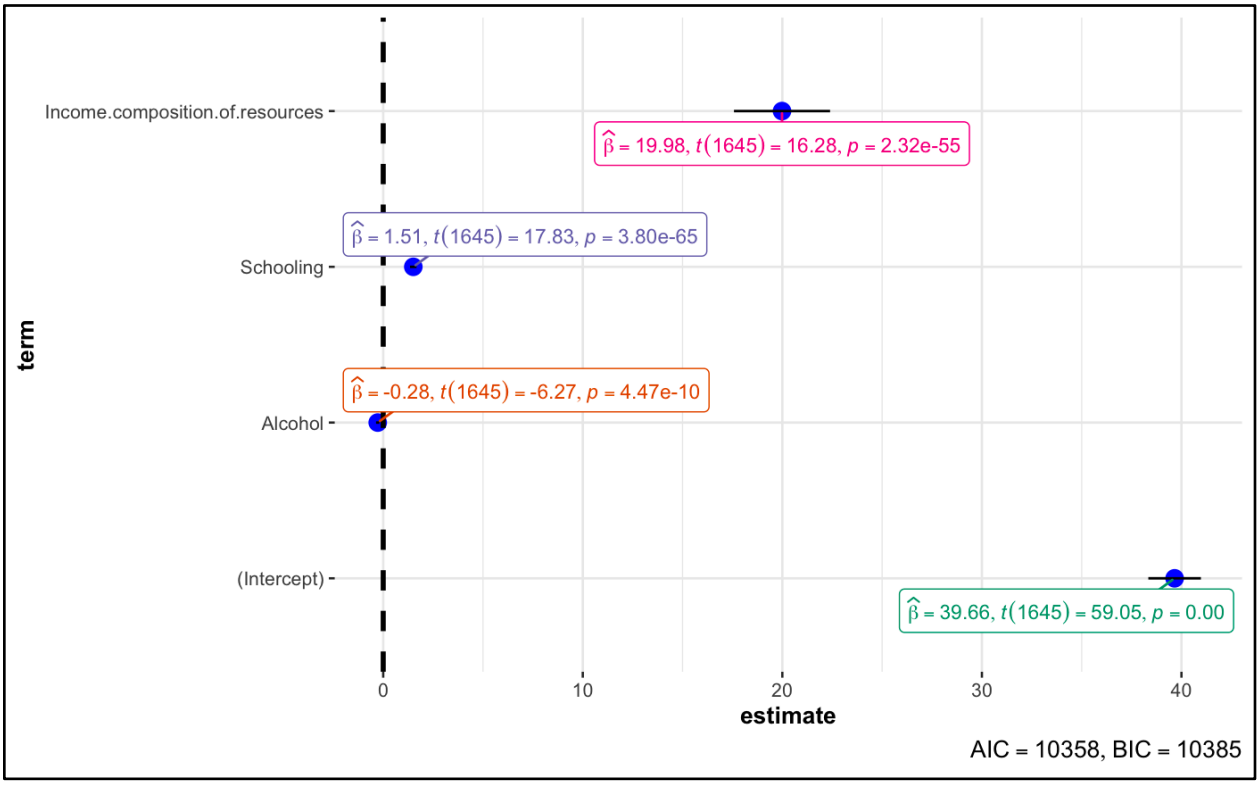


Figure 3 was created using the ggcoefstats() function, which visualizes a statistical analysis of β_0, β_1, β_2, and β_3. The vertical dotted line is the y-axis and specifies the estimated $\beta = 0$. If the black line spanning from any of the variables crosses that dotted line, it means that the 95% confidence interval includes 0, and the respective β coefficient value is not statistically significant. This visualization, however, shows that all the β values in the analysis are statistically significant, because none of the intervals include zero. The closest value to 0 is the coefficient for alcohol, which is also the only negative coefficient, but its confidence interval does not intersect or include 0. This means that alcohol is the least important for predicting the relationship, but its inclusion in the model is still statistically significant.

Figure 3:



Regression Analysis

The linear regression model proposed in this project is designed to determine the relationship between alcohol consumption, schooling, income composition of resources, and life expectancy. The goal is to determine if the coefficients for each of these variables are statistically significant when predicting life expectancy. In other words, the aim is to determine whether the variables are useful for predicting life expectancy. The equation determined using the linear model function, in the form of $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$, was:

$$Life\ Expectancy = 39.65669 - 0.27567(Alcohol) + 1.50779(Schooling)$$
$$+ 19.98407(Income\ Composition\ of\ Resources)$$

This equation takes liters of alcohol consumed per capita, years of schooling, and income composition of resources, and outputs life expectancy in years. Because the alternative hypothesis proposes that the slope of at least one coefficient is not equal to zero, a hypothesis test using the *F*-test is important to determine whether the null hypothesis, $\beta_1 = \beta_2 = \beta_3 = 0$ can be rejected.

The overall regression model was statistically significant, $F(3, 1645) = 814.7$, $p < 0.001$, $R^2 = 0.5977$, $R^2_{Adjusted} = 0.597$. In addition, a hypothesis test on the individual variables shows that all three predictor variables were statistically significant in this model. With other variables held constant, alcohol consumption had a statistically significant relationship with life expectancy such that for each liter increase in alcohol consumption, predicted life expectancy decreases by 0.276 years, $t(1645) = -6.275$, $p < 0.001$. Similarly, for the model, if schooling was the only variable to change, for each year increase in schooling, life expectancy is expected to increase by 1.508 years, $t(1645) = 17.829$, $p < 0.001$. The last coefficient showed that, for all other variables held constant, if income composition of resources increases by 0.1 on the human development index, predicted life expectancy increases by 1.998 years, $t(1645) = 16.279$, $p < 0.001$. This means that the null hypothesis, $\beta_1 = \beta_2 = \beta_3 = 0$, can be rejected, and that there is a significant relationship between each of the explanatory variables and life expectancy, as displayed by Figure 3. Finally, the adjusted R^2 value means that 59.7% of the variability in life expectancy can be explained by the three statistically significant predictor variables used in the model.

Discussion

Based on our analysis at the 5% significance level, the multiple linear regression model using alcohol consumption, years of schooling, and income composition of resources to predict life expectancy was statistically significant. Additionally, using $\alpha = 0.05$, we can reject the null hypothesis that there is no relationship between each of those individual predictor variables and life expectancy. In fact, all three were significant predictors in the model, because $p < 0.001$ for each coefficient, therefore $p < \alpha$, and the null hypothesis could be rejected. This data provides evidence that the true slope coefficient between alcohol consumption and life expectancy, the true slope coefficient between hours of schooling and life expectancy, and the true slope coefficient between income composition of resources and life expectancy are all three statistically significantly different than zero. This means the multiple linear regression model can effectively be used to predict life expectancy.

It is important to note that the data cannot imply causation, so the model does not mean any of the variables cause a higher (or lower) life expectancy. Because the data collection was not intended to determine causation, confounding variables could be present. It is possible that the variables used to predict life expectancy in the current analysis reflect another variable such as GDP which is a confounding variable, and therefore other variables could be the cause of (or a contributor to) increased or decreased life expectancy.

This statistical analysis utilizes multiple linear regression as a means for constructing a model of life expectancy and determining if there is a statistically significant relationship between the chosen input and output variables. The multiple linear regression analysis in this report displays a useful application: determining the variables that are capable of predicting the

output variable. It is useful for determining which variables in the model have a statistically significant relationship and which variables are not statistically significant for prediction. One limitation of this methodology is that higher dimensional analysis with many variables can be difficult to visualize, or even understand. Another limitation is that this form of analysis requires quantitative variables, so the status variable was not applied, despite a potentially helpful trend from Figure 1 that indicated developed countries appear to be on the higher end of the life expectancy scale (but it would have been possible to include the status variable using one-hot encoding if required). Multiple linear regression's inability to determine causation is a large limitation. This study focuses on determining relationships and trends but does not account for confounding variables that may affect the studied response variable. In the future, this study could be enhanced and the model could be utilized more effectively by starting with all variables in the model and removing only the statistically insignificant input variables. This can be applied until all variables in the model are statistically significant. This would be a more efficient model and an effective way to determine the most statistically significant variables (R on Stats, 2021).

References

Breheny, P. & Burchett, W. (2017). Visualization of regression models using visreg. *The R Journal*, 9(2): 56-71.

Rajarshi, K., Russel, D., & Wang, D. (2018). Life Expectancy (WHO), Version 1. Retrieved 18 April, 2024 from www.kaggle.com/datasets/kumarajarshi/life-expectancy-who/data.

R on Stats and R. (2021, October 4). Multiple linear regression made simple. R-bloggers. https://www.r-bloggers.com/2021/10/multiple-linear-regression-made-simple/

World Health Organization Department of Data and Analytics. (2020, December). "WHO methods and data sources for life tables 1990-2019" Retrieved from https://www.who.int/docs/default-source/gho-documents/global-health-estimates/ghe2019_life-table-methods.pdf?sfvrsn=c433c229_5#:~:text=World%20Health%20Organization.,Population%20Prospects%202019.;%202019.&text=11.,1987;1(3).&text=14.,2020;396(10258).