Nini Curcione

An Investigation of Life Expectancy in Various Countries

Introduction

How life expectancy varies from country to country and which factors contribute to this gap is one of the most frequently asked questions about human life. In this investigation, a set of variables that seem to affect life expectancy will be chosen. The goal is to find out which measures are most important for predicting how long a newborn child would live in a given country on average. The first step was to investigate the factors that contribute to a person's high quality of life. For a healthy life, one of the most important aspects is having access to food, education, and basic sanitation.

Numerous articles discuss the factors that contribute to the higher life expectancy of people living in developed nations compared to those living in developing nations. An example of this is a study that was published in the International Journal of Physical Education and Sports. It argues that health education and promotion are directly linked to people's life expectancy around the world. In addition, daily conditions like peace and terrorism in nations are linked to people's life expectancy due to the uncountable number of terrorist victims (Factors Affecting Life Expectancy in Developed and Developing Countries of the World, 2016). There are additional factors in people's routines that should also be highlighted, in addition to all the measures that result from some countries' higher level of development. Consumption of alcoholic beverages and smoking pose health risks to consumers. A 1998 study titled "Alcohol-Related Mortality by Age and Sex and its Impact on Life Expectancy" found that in addition to the high number of deaths from health conditions brought on by excessive alcohol consumption, there are also a significant number of accidental and violent alcohol-related deaths. This investigation aims to identify the most important factors that influence life expectancy and develop a model that can accurately predict it.

Methods

In collaboration with universities, public agencies, non-governmental organizations, and the United Nations, the Gapminder Foundation is a non-profit organization that produces free resources to make the world understandable based on reliable statistics. Eight different datasets from the Gapminder foundation, in addition to the response variable itself, were selected as the most relevant predictors for life expectancy for this study. Gapminder consolidates information from different authority sources, and the information utilized for this study was gathered from

World Bank, World Wellbeing Association, Worldwide Wellbeing Information Trade, UNICEF and different sources. Since the goal of this investigation is to predict life expectancy, it is important to keep in mind that the observations will be country-specific, meaning that each measure will be based on average values for each country at a specific year. We can assume that the data are representative because organizations take into account all countries when compiling information like life expectancy and it is important to understand this context. However, it is necessary to select the year with the greatest number of observations and only consider databases that are as complete as possible because there is no dataset that contains measures for every country every year. The values from the year 2015 were selected from each dataset because this was the year with the fewest empty cells in order to have as many observations as possible. To generate a clean dataset, it is still necessary to eliminate observations with empty cells. In order to achieve this, inner join operations were used to combine the eight tables in Excel. The total number of observations after these operations was 127, which is a greater number than most considering the number of variables we have and there are 10 observations per variable.

Food Supply, GDP per Capita, Broadband Subscribers, Basic Sanitation, Average School Years, High-Tech Exports, Alcohol Consumption, and MCV Immunization were analyzed as predictors in this investigation. Continuous variables were identified as potential life expectancy predictor for example the average number of years a newborn child is expected to live while maintaining constant mortality patterns is referred to as life expectancy, which is measured in years. The percentage of people in a nation who have access to basic sanitation services is the definition of basic sanitation, which is measured in percentage. The percentage of children under one year old who received at least one dose of the measles-containing vaccine in 2015 is also recorded as MCV immunized. The number of people who subscribe to high-speed public internet access (speeds greater than 256 kbit/s) per 100 people is known as broadband subscribers. The average annual consumption of pure alcohol by a person aged 15 or older is defined as alcohol consumption, which is measured in liters. The number of kilograms of food consumed per person per day is recorded in kilocalories. The gross domestic product divided by the population is the GDP per Capita. It is adjusted for inflation and measured in US dollars. The average number of years spent in primary, secondary, and tertiary education by men between the ages of 15 and 24 is represented by school years. Last but not least, the percentage of manufactured exports with a high R&D intensity—such as computers, pharmaceuticals, scientific instruments, and electrical machinery—is represented by high-technology exports.

Since the response variable- Life Expectancy is a binary variable, multiple linear regression models were used to analyze this data. By plotting scatterplots and checking for linear associations, the first step was to examine each variable separately against the response to discover linear relationships. All subsets, backward elimination, forward selection, and stepwise regression were used to create the best model. The final model was constructed by combining variables from both to achieve a balance between complexity and effectiveness based on the models discovered through these approaches. Furthermore, since the conditions for multiple regression were already met and the model could not be significantly improved, no transformations or interactions were required.

Results

Five-number statistics and there corresponding histograms have been computed for each of the variables that were found to be potential predictors for life. The response's summary statistics and any other relevant variables can be viewed from the table. The variable *LifeExpectancy* has a standard deviation of 7.35 years, and the mean life expectancy for a newborn child is 73.45 years. The majority of this variable's values fall between 50.5 and 84.1, as shown by its distribution. It is a right skewed histogram caused by a number of outliers.

With a standard deviation of 20224.24, the sample's mean average Gross Domestic Product per Capita (*GDPperCapita*) is 15547.69 dollars. This value is very high, so you can see a right skewness in the histogram. Most of the values are between 347 and 108,000 American dollars, but there are some outliners that are significantly different from most countries. This is also true for the variables exports of high technology and broadband subscribers; the substantial right skewness can be explained by the fact that a small number of wealthy nations have significantly higher measures in these categories. In contrast, the distributions of Basic Sanitation (*BasicSanitation*) and MCV (*mcvImmunized*) histograms are rather left-skewed, indicating the opposite. According to Table 1, 78% of people have access to basic sanitation, with a standard deviation of 27%. We have a mean of 89% and a standard deviation of 11% for immunized MCV. The majority of nations were able to achieve relatively high averages in these measures compared to broadband subscribers, high-technology exports, and GDP per capita, according to these statistics. In addition, we can observe more normal distributions in the histograms for the variables average school years for men, food supply, and alcohol consumption. According to table 1, men spend an average of 10.56 years in school. The mean is very close to the median in this instance. This also applies to food supply and alcohol consumption: adults consume an average of 6.87 liters of alcohol annually and 2923.70 kilocalories per day, respectively.

Figure 1: Distribution of Life Expectancy

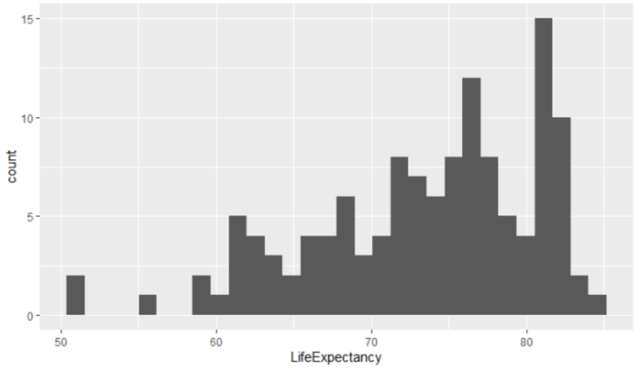

Figure 2: Distribution of Broadband Subscribers

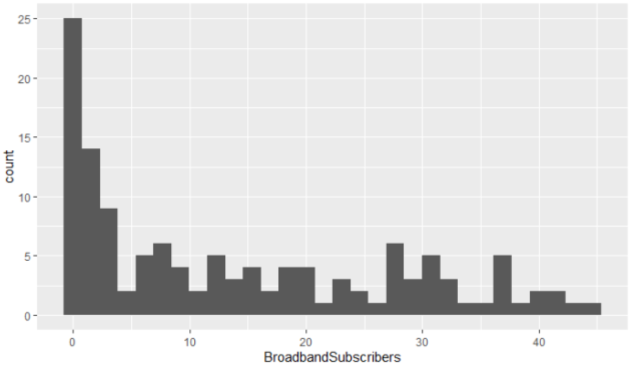

Table 1: Summary statistics for Life Expectancy, GDP per Capita, Food Supply, and School Years

Variables	**Mean**	**Standard Deviation**
LifeExpectancy (years)	73.45	7.35
GDPperCapita (US$ adjusted)	15547.69	20224.24
FoodSupply (Kcal/ person &day)	2923.70	447.795
SchoolYears (years)	10.56	2.50

BasicSanitation	.78	.27
AlcoholConsumption	6.87	4.04
BroadBandConsumption	13.87	13.41
mcvImmunized	.90	.11
HighTechExport	.099	.10

From the matrix, it seems that the correlation between life expectancy and basic sanitation is the highest (0.83), followed by broadband subscribers (0.75), school years (0.72), and food supply (0.72). Nevertheless, this does not imply that they would all be significant in the same model because they may all provide essentially the same information regarding the response variable. Other variables that are less correlated with the response might be more beneficial to the model as a whole. With a correlation coefficient of 0.55, these variables are MCV immunized, alcohol consumption, and GDP per capita (0.63).

The simple two-variable model with Broadband Subscribers and Basic Sanitation as predictors was found to be the best after running the best subset. On the other hand, the best model constructed through all backward elimination, forward selection, and stepwise regression had the same two variables as the most significant predictors in addition to three other predictors that were only marginally significant: MCV Immunization, Alcohol Consumption, and GDP per Capita. The more complex model had a much lower mallow CP, despite having a R-squared that was comparable to that of the simpler model. As a result, the final model was chosen because it was able to strike a balance between R-squared, mallow Cp, and complexity by analyzing all of the combinations of the final three variables that were added to the first two.

The selected specification consists of:

$$\boldsymbol{LifeExpectancy_i = \beta_0 + \beta_1 BasicSanitation_i + \beta_2 BroadbandSubscribers_i}$$
$$\boldsymbol{+ \beta_3 mcvImmunized_i + \varepsilon_i}$$

The regression summary for our preferred model that is shown in Table 3. Our adjusted R-squared value demonstrates that the chosen model correctly predicts 76% of the variation away from the mean life expectancy. The model has a mallow Cp of 8.09, which is neither very small nor very large. Additionally, it is possible to observe that each of the obtained p-values is statistically significant. However, mcv immunized is only weakly significant with a p-value of 0.097, while only broadband subscribers and basic sanitation can be considered highly significant predictors with p values close to zero. The coefficients of broadband supporters, fundamental disinfection, and mcv inoculated demonstrate positive associations with future. The coefficients show that a membership increment for each 100 individuals in broadband endorsers is related with an increment of 0.20 long stretches of future, while a rate expansion in essential disinfection and mcv vaccinated with increments of 14.57 and 6.39 years, separately. In addition, the confidence intervals indicate that we are 95 percent certain that the coefficient for broadband subscribers is between 0.13 and 0.26, for basic sanitation between 10.81 and 18.34, and for mcv immunized between -1.17 and 13.95.

The relationship between the predictors and the response variables is relatively linear, and the error terms of our preferred model reasonably satisfy the constant variance, zero mean, and normality conditions, according to analyses of residuals vs. fitted plots and normal quantile plots. The sample also appears to satisfy the requirements for randomness and representativeness in the prior analysis. The fact that one nation's actions have no effect on those of other nations suggests that independence has also been achieved. Additionally, there are no points considered to be leveraged in accordance with Cook's-D criteria in the residual’s vs leverage plot, so there are no influential ones.

Multicollinearity is yet another issue that requires consideration. In a multiple regression model, multicollinearity occurs when two or more explanatory variables have high correlations with one another. This can make it difficult to determine which variables actually affect life expectancy. The correlation between broadband subscribers and basic sanitation was found to be 0.667. This indicates that there is not a strong correlation between the predictors. In addition, when the model was tested without either of these variables, its effectiveness significantly improved. It was discovered that all of the other predictor combinations had a lower correlation than this one did.

Table 3: Ordinary Least Square Estimates of the Relationship between Life Expectancy, Broadband Subscribers, Basic sanitation, and MCV Immunized.

Variables	**Coefficient Estimates**	**Standard Errors**	**Confidence Intervals (95%)**	**P values**
Intercept	53.62***	2.86	[47.96, 59.28]	< 2e-16
BroadbandSubscribers	0.20***	.032	[0.13, 0.26]	1.57e-08
BasicSanitation	14.57***	1.90	[10.81, 18.34]	4.71e-12
mcvImmunized	6.39	3.82	[-1.17, 13.95]	0.097
Observations	127			
Adjusted R^2	0.76			
Mallow Cp	8.09			

*** $p < .001$; ** $p < 0.01$; * $p < 0.05$; $p < 0.1$

Discussion

The final model for predicting Life Expectancy includes Broadband Subscribers, MCV Immunization, and Basic Sanitation. Broadband Subscribers and Basic Sanitation were found to be the most significant predictors of Life Expectancy, according to this investigation. These results are significant because they point to the future direction of health policy and reveal that there is a significant connection between health and economics at the country level.

Also in the final model, the coefficient for basic sanitation is positive and statistically significant. The linear regression was significantly impacted by this predictor. Health is fundamentally influenced by proper sanitation. Research has demonstrated a connection between poor sanitation, hygiene, stunting, and anemia, which are risk factors for deficits in the development of children. As a result, there is a strong correlation between life expectancy and access to basic sanitation. (Water, sanitation, and hygiene, environmental enteropathy, nutrition, and early child development: making the links, 2014).

The other highly significant predictor of Life Expectancy was the number of Broadband Subscribers. Notable, cofounding variables play a significant role here, but it is difficult to see how access to high-speed internet would ever affect the average number of years a newborn child is expected to live. The majority of people have easy access to the internet in wealthy nations. As a result, the positive coefficient of the variable in the final model indicates that those nations are generally more developed in all respects and, as a result, are anticipated to have higher life expectancy measures.

The strength of this study is the meaning of the indicators alongside a high changed R squared. The flawed datasets that were gathered, on the other hand, are one of the study's limitations. Even though all of the data used came from official sources, not all of the datasets that were recorded included a lot of countries. A smaller dataset is produced when all of the datasets must be combined using only the countries that match the eight variables. If there is a tendency in the nations in which it was not possible to collect measures, arguments could be made against this study. Therefore, in order for the validity of the inferences to be unquestionable, future research ought to expand the sample of nations, either by waiting for organizations to record the measures in a more complete manner in the future or by carrying out additional research to discover additional predictors for which the datasets might already be complete enough to address this issue.

Bibliography

Ngure FM, Reid BM, Humphrey JH, Mbuya MN, Pelto G, Stoltzfus RJ, "Water, sanitation, and hygiene, environmental enteropathy, nutrition, and early child development: making the links" (2014, January). https://doi.org/10.1111/nyas.12330

Pia Makela, European Journal of Public Health 199S: S: 43-51, "Alcohol-related mortality by age and sex and its impact on life expectancy" (1998, September). https://doi.org/10.1093/eurpub/8.1.43

Alamgir Khan, Dr. Salahuddin Khan, Manzoor Khan, International Journal of Physical Education and Sports. "Factors effecting life expectancy in developed and developing countries of the world", Volume 1, Issue 1 (2016, November).