STAT 630-02 SUMMARY REPORT- Life Expectancy

By Group 6 -Vikas Reddy,Naveen Kuparaju

Introduction

With changing lifestyles and development, the life expectancy value in years also changes from the last 70 years i.e from the era of industrial revolution period. So in order to find how major factors like health, economy, social status, etc. play a major role in effecting the life expectancy value, 2 researchers namely Deeksha Russell and Duan Wang. Have taken the health, social factors, immunisation factors, mortality factors into consideration and collected some data from WHO and UN for years 2000 to 2015. With which they have done multiple testing on which factor effects to lower life expectancy and where each country should improve.

The data collected consists of 22 columns, and 2938 rows for a total of 193 countries(for each country from 2000 to 2015).

Research Question: Does Immunization factor like Alcohol, and Social factor like schooling have any effect on Life expectancy values individually?

This area is worth studying as alcohol is one of the major factors that is said to degrade life, while some also say alcohol is something that relieves stress and helps in betterment of life.

On the other side there is a speculation that schooling provides discipline and the values to be followed that helps in betterment of life(indirectly increases life expectancy).

In order to find up to what extent these two factors individually affect life we conducted linear regression analysis on these topics.

Null Hypothesis(H_0):

For Life expectancy with respect to Schooling $H_{0.}$: There is no relationship between Life expectancy(Response Variable) and Schooling values(Predictor Variable)
$\beta_1 = 0$ & $Y= \beta_o + \epsilon$. Y = Life expectancy value, β_1 is slope for predictor of interest(Schooling) , β_o is Intercept, ϵ is the error term.

For Life expectancy with respect to Alcohol $H_{0.}$: There is no relationship between Life expectancy(Response Variable) and Alcohol (Predictor Variable)
$\beta_2 = 0$ & $Y= \beta_o + \epsilon$. Y = Life expectancy value, β_2 is slope for predictor of interest(Alcohol), β_o is Intercept, ϵ is the error term.

Alternate Hypothesis(H_A):

For Life expectancy with respect to Schooling (H_A): There is some relationship between Life expectancy(Response Variable) and Schooling values(Predictor Variable)

$\beta_1 \neq 0$ then Y= β_o+β_1*Schooling+ϵ. Y = Life expectancy value, β_1 is predictor of interest(Schooling) , β_o is Intercept, ϵ is the error term.

For Life expectancy with respect to Alcohol (H_A): There is some relationship between Life expectancy(Response Variable) and Alcohol (Predictor Variable)

$\beta_2 \neq 0$ then Y= β_o+β_2*alcohol+ϵ. Y = Life expectancy value, β_2 is slope for predictor of interest(Alcohol), β_o is Intercept, ϵ is the error term.

Exploratory Data Analysis:

Before we go through the Major part of study let's explore the Summary statistics and General information pertaining to our study.

There were more than 40% of missing values and we have used mean imputation to fill these missing values as if all the missing values are removed then there will be no variables available for some years like 2015, 2005 etc where there are major missing values in Alcohol, Schooling etc.

After Imputation the summary statistics are as follows.

Variable(Description)	Mean or n(frequency)	SD or %(percentage)	Median
Life Expectancy(Years)	69.22	9.50	72
Schooling(Years in schooling	11.99	3.26	12.1
Alcohol(Number of litres in year)	4.60	3.91	4.16

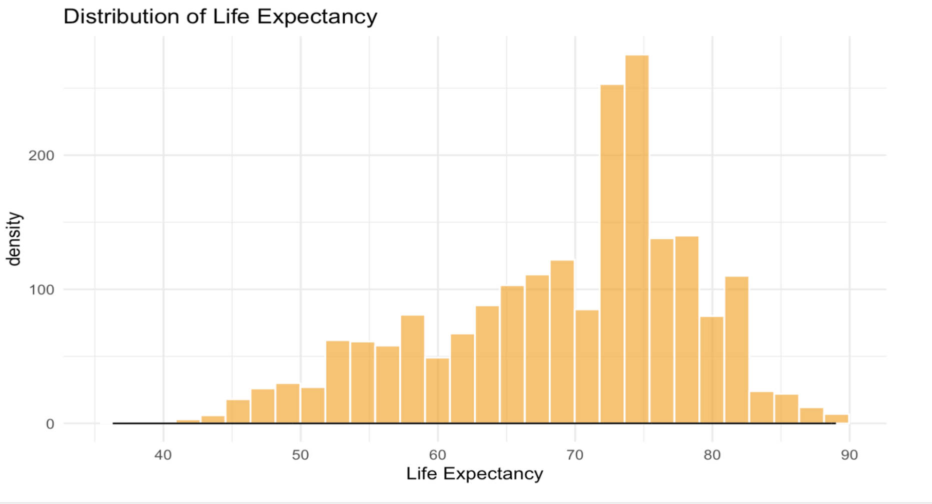
The mean values for primary variables of interest (Life expectancy rate, Schooling, Alcohol and few other variables are shown) here shows that Life expectancy mean values are 69.22 years, and have a Standard variance of 9.5 years.

While schooling and Alcohol have a mean value of 12 years schooling Approximately and standard deviation in Schooling years of 3.2 years, While for alcohol, the mean number of litres consumed is 4.6 litres per year with a standard deviation of 3.91 Litres.

The Other variables are not being interpreted as they are not have interest in the project.

The EDA of our primary variables of interest gives the following plots that check if there is any Linear relationship between the variables.

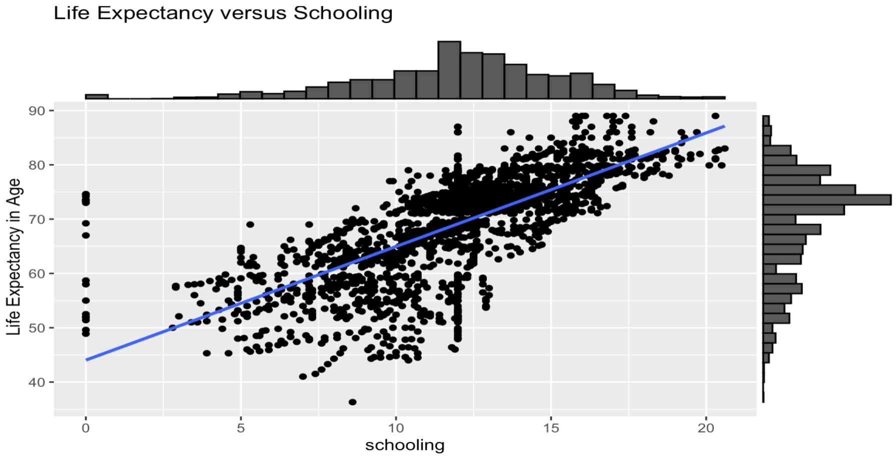
Life Expectancy Histogram: The histogram shows that life expectancy is highly left skewed, we may use a transformation to change this but, as it is real life data and the data may be dependent on the countries and several factors we cannot change the data. Here X represents age and Y represents how many rows(by countries have the life expectancy values).



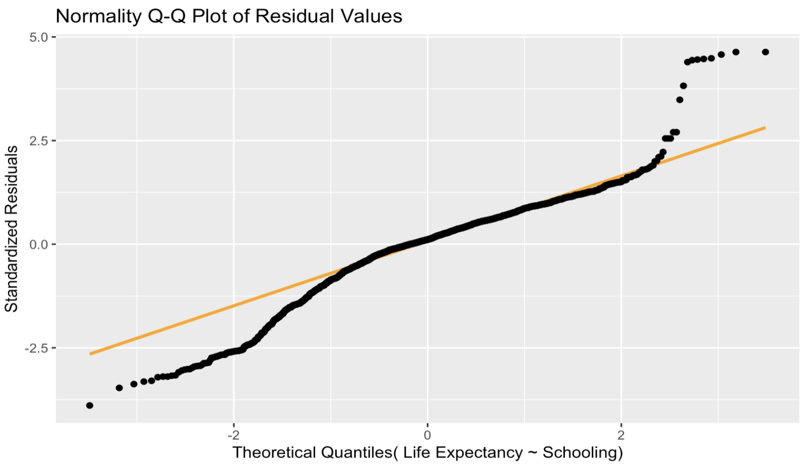
Independence: Independence is Satisfied as all data is collected country wise and does not depend on other countries and also data does not depend on previous years data.

SLR Analysis for Life Expectancy and Schooling:

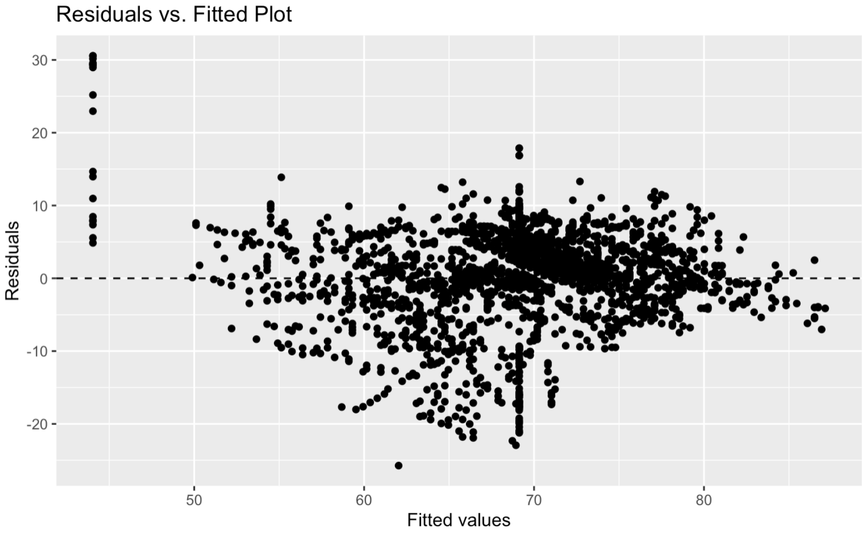
Linearity: Satisfied



Normality: Though Normality has fat Tails, most points lie on the Slope line. As this is real life data.

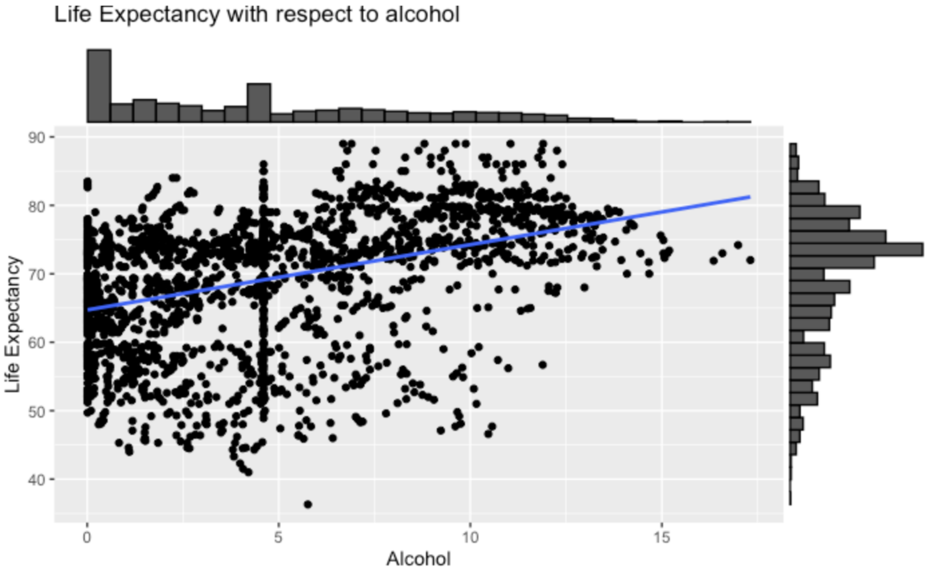


Residuals: the residuals seem to be evenly spread with a small fan effect. So constant variance satisfied.

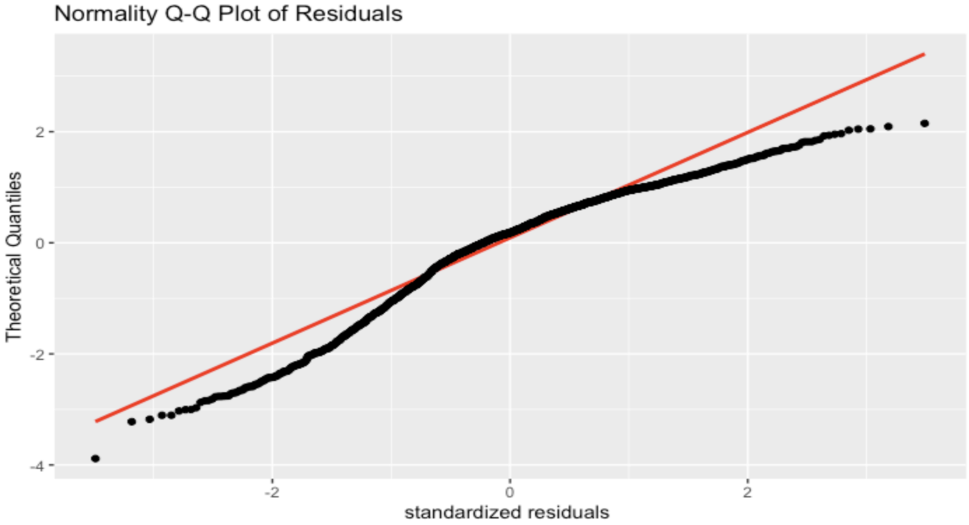


SLR for Life Expectancy and Alcohol:

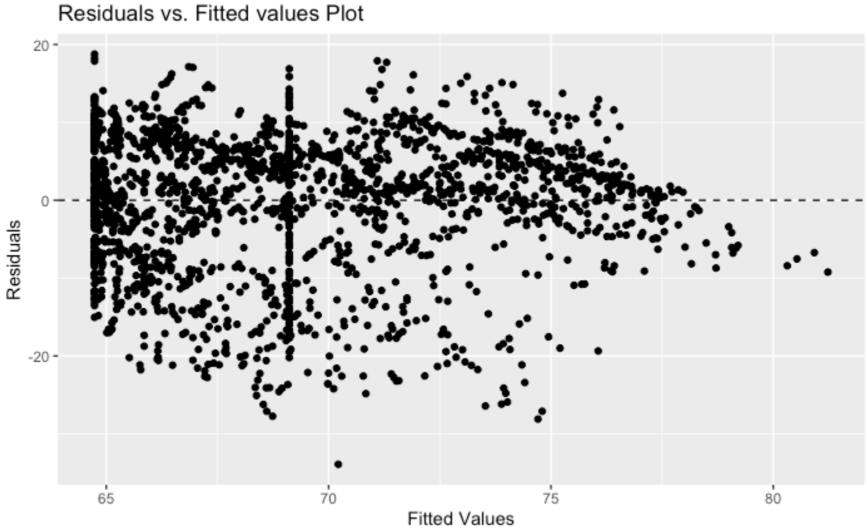
Linearity:



Normality: From the graph we can see that the points violate Normality and form u shaped patterns. We have tried using Transformations to this variable, and Transformation does not change the Normality. So we continued with the untransformed Model. As most points are nearer to the Slope Line.



Residual Vs Fitted Plot: Though the values of Residual vs fitted values shows most points are clustered at the left side of plot, but other points are scattered all around the plot so it partially satisfies the Equal variance assumption.



Results: From our test we find that both Schooling and Alcohol have a positive linear relationship with Life Expectancy Individually .

For Life Expectancy VS Schooling:The confidence interval for the average schooling value of 12 gives the life expectancy value in confidence intervals of 68.8 and 69.42 years. With T statistic value of 46.83 and P value ≈0 (2.2e-16),
From p value we reject Null Hypothesis and Conclude that there is a linear relationship between Schooling and Life Expectancy. With final model equation as

Life Expectancy = 44.03+2.09*Schooling
means for every year increase in schooling, life expectancy value increases by 2.09 years.

For Life Expectancy VS alcohol: The confidence interval for the average Alcohol consumption in litres of 4.6 gives the life expectancy value in confidence Interval range of 68.7 and 69.5 years. With T statistic of 19.51and p value ≈0 (2.2e-16). Considering the P value we reject the Null Hypothesis(H_0) and conclude that there is a positive linear relationship between Life Expectancy and Alcohol.
With final equation as
Life Expectancy = 64.72 + 0.95*Alcohol, means for every litre increase in alcohol consumption the Life expectancy value in years increases by 0.95 years.

Final Conclusion: Conclusion:

The exploratory data analysis (EDA) and subsequent analysis of the data have revealed insightful patterns regarding the relationship between life expectancy and various factors. Notably, there appears to be a positive association between life expectancy and schooling, a key social factor. Additionally, a positive linear relationship was observed between life expectancy and alcohol consumption. These findings underscore the complex interplay between social and lifestyle factors and their impact on life expectancy.

Limitations:

However, it is crucial to acknowledge the limitations inherent in the analysis. One significant constraint is the presence of numerous missing values for certain years, which compromises the completeness of the dataset. To attain more robust conclusions, a dataset with a larger sample size and fewer missing values would be indispensable. Furthermore, the analysis focused solely on simple linear regression (SLR), and adopting a more comprehensive approach, such as multiple linear regression (MLR), could provide deeper insights into how various variables collectively influence life expectancy.

Another limitation stems from the temporal scope of the data, covering the years 2000 to 2015. This timeframe provides a historical perspective on life expectancy trends but fails to capture the present scenario. It is essential to recognize that the results pertain to past life expectancy values, and a more contemporary dataset, such as the revised data from the World Health Organization (WHO), would offer insights into current life expectancy trends for diverse countries.

Future Studies and Analyses:

In light of the limitations identified, future studies should prioritize obtaining a dataset with reduced missing values and an extended temporal range. This would facilitate a more comprehensive exploration of the intricate relationships between life expectancy, social factors, and lifestyle variables. Adopting advanced analytical techniques, such as MLR, could unravel nuanced interactions among multiple variables.

Additionally, expanding the scope beyond the provided timeframe to include more recent data would contribute to a holistic understanding of present life expectancy dynamics. Comparative analyses with updated WHO data could unveil shifts in global health patterns and inform policymakers and researchers about the efficacy of interventions over time.

In conclusion, while the current analysis provides valuable insights, addressing these limitations and incorporating advancements in analytical methodologies will undoubtedly enhance the depth and applicability of future studies in this field.

Source of the Dataset

Kumar, R. (2022). Life Expectancy WHO. Kaggle.
https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who.

World Health Organization. (n.d.). Global Health Observatory (GHO) data: Life expectancy and healthy life expectancy.
https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-life-expectancy-and-healthy-life-expectancy