Emily Rod

ECON0150 Final Project

Effects of Socioeconomic Factors on Life Expectancy

1. Introduction and Data Source

This research focuses on how socioeconomic factors impact life expectancy. The two factors that were investigated were GDP and BMI. My future career in life insurance as an actuarial associate sparks my specific interest in mortality, and mortality is a topic often modelled using regression. The data was compiled on Kaggle, but sourced from the World Health Organization, a reputable organization.

2. Data description

First, the data was filtered to include only the predictors and response variables for our general linear model. In this case, the predictors are GDP and BMI and the response variable is life expectancy. For this model, I chose to use the data from the year 2015.

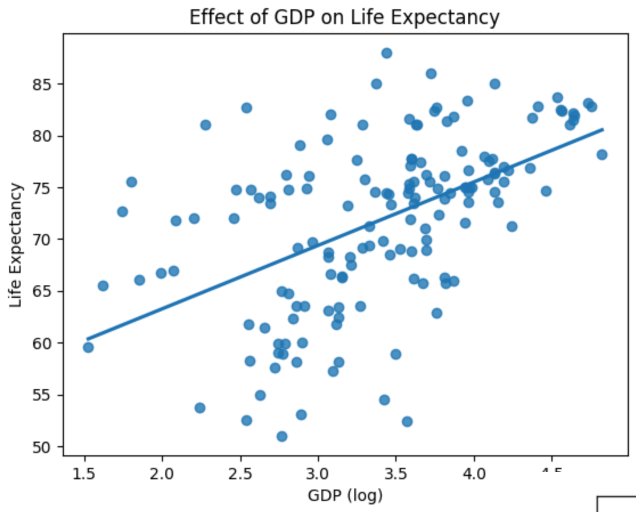
The relationship between GDP and life expectancy, and BMI and life expectancy were visualized using scatterplots to make sure the requirements for linear regression were met. For GDP the relationship visible in the scatterplot was exponential which was likely due to the skewed nature of the histogram of GDP. To solve this issue, a log transformation was performed on the GDP variable to reduce skew. In the scatterplot between log GDP and life expectancy, there was a linear relationship visible. The scatterplot between BMI and life expectancy, there was a linear relationship visible.

3. Methodology

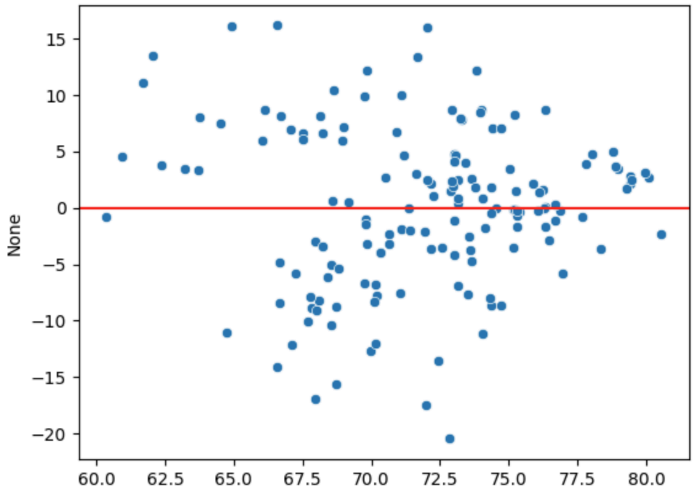
The General Linear Model chosen was a single linear regression model that utilizes ordinary least squares to determine a line of best fit. In linear regression, the data is trained on the training data, and then evaluated using the test data. An 80/20 or 70/30 split between training and test data is common. The model was visualized using a scatterplot with the linear regression line displayed.

Limitation can include the line being skewed by outliers or attempting to model a nonlinear relation. Also the four assumptions of linearity, normality, homoskedasticity, and independence must be met. In this model, a log transformation was used to maintain the linearity assumption and robust errors were used for homoskedasticity. Normality and independence were met.

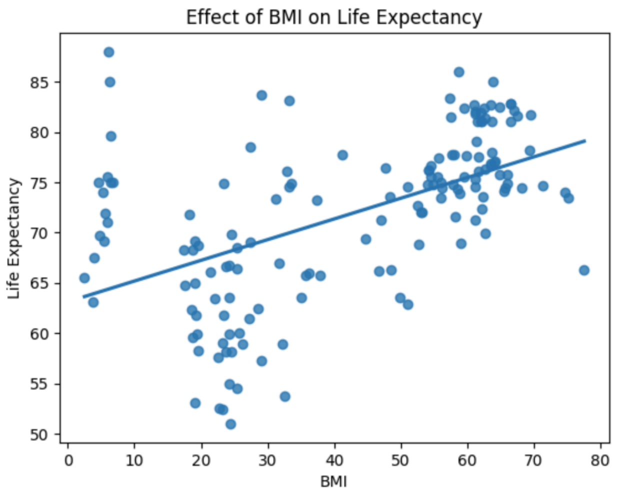
4. Results and analysis



The linear regression line modelling the relationship between GDP(log) and Life Expectancy was calculated as ŷ=6.1233x+51.077+ε. This model implies that if GDP(log) is equal to 0 meaning log is equal to 1, then the predicted life expectancy is 51.077 years. B1 of 6.1223 explains that every increase of one to ln(GDP) is associated with an increase of 6.1223 to life expectancy. The residual plot to the left shows mostly evenly distributed residuals, which indicates a strong model. There are slightly more positive residuals at the lower values, but this is expected since we had to perform a log transformation in the beginning.



The linear regression line modelling the relationship between BMI and Life Expectancy was calculated as ŷ=0.2057x+63.1208+ε. This model implies that if BMI is equal to 0, then the predicted life expectancy is 63.1208 years. B1 of .2057 represents that for every increase of one to BMI is associated with an increase of .2057 to predicted life expectancy. The intercept is most likely significantly higher in this case because of the high outliers with samples with 0-10 BMI as the predictor. This low of a BMI could be due to



missing or incomplete data that requires further investigation.

5. Conclusions

Overall, it was determined that there is a statistically significant relationship between both GDP and life expectancy and BMI and life expectancy. This was determined by conducting two separate tests. For both GLMS, the null hypothesis was that B1=0, implying no relationship between the predictor and the response variable. Both null hypotheses were rejected with p-values of 0.000 (presumably rounded). Using an alpha of .005, we can say with 95% confidence that the B1 of BMI falls in the interval [.145, 0.267] and the B1 of GDP falls in the interval [4.576 , 7.671]. The results of both GLMs indicate a positive relationship between the predictor and response variable, with GDP having a stronger impact noted by the larger magnitude coefficient, even when the natural log is taken. To further investigate this research question, other socioeconomic factors could be modelled or a multilinear regression with both GDP and BMI as predictors could be tested. Additionally, other years worth of data could be tested to confirm this result.

References

varunsaikanuri. “Life Expectancy Visualization.” *Kaggle.com*, Kaggle, 3 Oct. 2022, www.kaggle.com/code/varunsaikanuri/life-expectancy-visualization/input. Accessed 5 Dec. 2025.