Final Write Up

The multiple regression analysis conducted sought to investigate the complex interplay between various socioeconomic and health-related factors and their collective impact on life expectancy. This approach was predicated on the hypothesis that life expectancy is not solely determined by economic performance, as measured by GDP, but is also significantly influenced by educational attainment, health behaviors (e.g., alcohol consumption), health interventions (e.g., immunization rates), and the prevalence of major diseases (e.g., HIV/AIDS).

The outcome of the exploratory data analysis (EDA) and subsequent regression analysis was revealing. By incorporating a broader set of variables—GDP, schooling, alcohol consumption, BMI, total health expenditure, diphtheria immunization rates, and HIV/AIDS prevalence—the model's explanatory power increased substantially, with an R-squared value of approximately 0.793. This indicated that nearly 79.3% of the variance in life expectancy could be explained by these combined factors, a significant improvement over the simple regression model that considered only GDP. The mean squared error (MSE) also decreased markedly, suggesting that the predictions made by the multiple regression model were much closer to the actual observed values.

Despite these improvements, the analysis could have benefited from the inclusion of additional variables that were not considered. For instance, factors such as access to clean water and sanitation, nutritional status of the population, income inequality, and healthcare access (beyond expenditure and specific immunizations) might also have a significant impact on life expectancy. These additional variables could potentially provide a more comprehensive understanding of the determinants of life expectancy.

One assumption inherent in the analysis was the linear relationship between the independent variables and life expectancy. In reality, these relationships may be non-linear or may involve complex interactions among the variables that weren't accounted for in the model. For example, the impact of schooling on life expectancy could vary at different levels of GDP or in the presence of certain health interventions.

The primary challenges faced during the analysis included the selection of relevant variables from a wide array of potential predictors and the interpretation of their collective impact on life expectancy. Additionally, dealing with missing data and deciding how to handle it without introducing bias or losing significant information was a critical step in the preparation of the dataset for analysis.

In summary, the multiple regression analysis provided valuable insights into the factors affecting life expectancy, demonstrating the importance of a multifaceted approach when examining complex health outcomes. However, the analysis also highlighted the limitations of regression modeling in capturing the full complexity of human health and the social determinants of health. Future research could benefit from incorporating a wider range of variables, exploring non-linear relationships, and considering the potential for variable interactions to gain a deeper understanding of the factors that influence life expectancy.