DATA KNIGHTS

RAHUL PANDYA

MOURYA ANDEY

GIRISH PARTHIBAN

HARSHITHA VALLURI

AAYUSH SACHIN KOTHARI

Abstract

This study delves into the prediction of life expectancy using machine learning techniques. A comprehensive dataset encompassing information on life expectancy and various health factors for 193 countries from 2000 to 2015 was meticulously assembled. The dataset's integration presented challenges due to missing values, particularly in variables like population, Hepatitis B, and GDP.

A thorough preprocessing phase involved null value removal, correlation and multicollinearity checks, interactive model testing, and further regression testing, including Linear, Ridge, Lasso, and Elastic Net regression.

The Elastic Net model emerged as the superior predictor of life expectancy, exhibiting the lowest MAE and RMSE values. Its ability to predict the life expectancy of individuals with high accuracy further highlights its effectiveness.

The findings underscore the potential of machine learning in predicting life expectancy, providing valuable insights for healthcare planning and resource allocation.

Table of Contents

1. Introduction

Life expectancy, a measure of the average number of years a person can expect to live, serves as a key indicator of a nation's overall health and well-being. Understanding the factors that influence life expectancy across different countries and regions is crucial for formulating effective public health policies and improving global health outcomes. This comprehensive study, conducted by the Data Knights, delves into the intricacies of these factors, examining their impact on life expectancy across 193 countries from 2000 to 2015.

Our study employs a comprehensive set of variables encompassing various health and socioeconomic factors to unravel the complex interplay influencing life expectancy. The response variable, life expectancy, is examined in conjunction with a range of predictors, including:

Immunization-related factors: These factors, such as Hepatitis B, Polio, and Diphtheria vaccination rates, reflect the preventive measures in place to protect populations from communicable diseases.

Mortality factors: Adult mortality and infant deaths provide insights into the overall health status of a population and its ability to manage health challenges.

Economic factors: GDP and percentage expenditure on healthcare indicate the economic resources available to address health issues and implement effective healthcare systems.

Social factors: Schooling and income composition of resources reflect the level of social development and the distribution of wealth within a country, both of which can significantly impact health outcomes.

1.1 Data Acquisition and Challenges

The foundation of our study lies in a meticulously compiled dataset drawn from two primary sources: the Global Health Observatory (GHO) data repository by the World Health Organization (WHO) and the United Nations website for economic data.

We meticulously collected information related to life expectancy and various health factors for 193 countries spanning the period from 2000 to 2015. Merging these data files into a single comprehensive dataset posed several challenges, particularly the presence of missing values, especially in variables like population, Hepatitis B, and GDP.

To address these challenges, we utilized R software, a powerful statistical analysis package, to identify and exclude countries with extensive missing data from our analysis.

1.2 Structure of the data

The final merged dataset stands as a testament to our rigorous data collection and cleaning efforts. Encompassing 22 columns and 2,938 rows, the dataset represents a wealth of information for 193 countries over a 16-year period.

The 22 columns encapsulate 20 predicting variables categorized into immunization-related, mortality, economic, and social factors. These variables provide a comprehensive picture of the health and socioeconomic landscape of the countries under study, allowing us to explore the intricate relationships between these factors and life expectancy.

Our research journey was not without its challenges. Missing values, particularly in crucial variables, posed a significant hurdle, requiring careful data cleaning and exclusion of incomplete data entries. Despite these obstacles, we persevered, driven by the potential of our findings to contribute to our understanding of global health disparities and factors influencing human longevity.

This study presents a valuable opportunity to examine the complex interplay of health and socioeconomic factors in determining life expectancy across diverse countries. Our findings can inform policymakers and healthcare providers, enabling them to develop targeted interventions and strategies to improve health outcomes and promote longevity worldwide.

Unraveling the enigma of life expectancy requires a deep dive into the intricate interplay of various factors that shape human health and longevity. Our study, encompassing a comprehensive dataset of 193 countries over a 16-year period, provides insights into the multifaceted influences on life expectancy, encompassing immunization rates, mortality indicators, economic conditions, and social determinants of health.

While challenges such as missing data presented hurdles, our commitment to rigorous analysis and data cleaning enabled us to overcome these obstacles and extract meaningful conclusions. Our findings shed light on the complex interplay of factors influencing life expectancy, paving the way for informed policy decisions and healthcare interventions aimed at improving global health outcomes and promoting longevity worldwide.

1.3 Null Values

The dataset has many null value counts, especially the significant variables such as HepatitisB, GDP and Schooling. After removing the NA values, we have can move on to the next step of checking the correlation plot.

2. Correlation Check:

2.1 Correlation:

Correlation is a statistical metric that gauges the extent to which changes in one variable are associated with changes in another. It provides a numerical measure of the strength and direction of a linear relationship between two variables. The correlation coefficient, ranging from -1 to 1, serves as this measure:

A correlation coefficient of 1 signifies a perfect positive correlation, implying that as one variable increases, the other also increases in a linear fashion.

Conversely, a coefficient of -1 denotes a perfect negative correlation, indicating that as one variable increases, the other decreases in a linear fashion.

A correlation coefficient of 0 suggests no linear correlation, implying that changes in one variable are not systematically related to changes in the other.

2.2. Correlation Plot in R:

A correlation plot in R is a visual representation of the correlation matrix, offering an overview of the relationships between multiple variables. The correlation matrix consists of correlation coefficients computed for each pair of variables. The plot visually communicates the strength and direction of these correlations, making it easier to identify patterns and dependencies among variables. For instance, using the corrplot package in R, you can generate a correlation plot that employs color intensity and numerical annotations to highlight the correlation coefficients in the matrix. This plot aids in identifying clusters of variables with similar patterns of correlation, facilitating a more comprehensive understanding of the relationships within a dataset.

In our dataset, the correlation plot reveals that most variables exhibit a correlation coefficient near 0, indicating little to no significant dependencies among them. However, a few notable exceptions stand out.

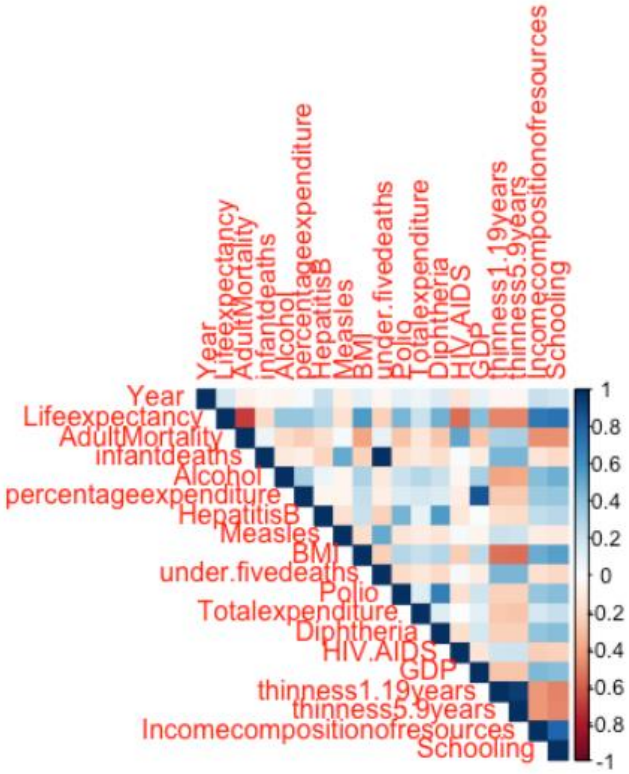

Figure 1. Correlation Plot

For instance, the variables "Under_five_Deaths" and "Infant_deaths" display a strong positive correlation, reflecting their shared meaningfulness and interpretability.

Similarly, "Life_Expectancy" and "Adult_Mortality" demonstrate a pronounced negative correlation, suggesting that as adult mortality decreases, life expectancy tends to increase.

This concise summary highlights the key correlations in the dataset, emphasizing the relationships that are particularly noteworthy for further analysis.

Analyzing correlations before regression analysis is crucial as it provides insights into the relationships between variables. In our dataset, most variables show little correlation, implying limited interdependence. However, notable exceptions, such as the strong positive correlation between "Under_five_Deaths" and "Infant_deaths" and the pronounced negative correlation between "Life_Expectancy" and "Adult_Mortality," reveal meaningful associations. Understanding these correlations is essential for regression analysis, as it helps identify potential multicollinearity issues and informs the selection of variables that significantly influence the dependent variable.

This preliminary correlation analysis serves as a foundation for a more informed and effective regression analysis, enhancing the reliability and interpretability of the dataset.

3. OUTLIER DETECTION:

Outliers, defined as data points significantly deviating from the overall dataset, pose a substantial impact on data analysis, potentially distorting statistical analyses, and machine learning models. Identification and management of outliers are crucial as they may signify data errors or reveal distinctive patterns warranting special attention. In R, diverse methods exist for outlier detection:

Boxplot Method: Utilizing visual representations of data distribution, boxplots facilitate outlier identification.

Z-Score Method: Z-scores quantify a data point's deviation from the mean in terms of standard deviations.

IQR (Interquartile Range) Method: This approach defines a range based on the interquartile range, identifying outliers beyond this range.

Tukey's Fences: Like the IQR method, Tukey's approach utilizes a constant multiplier to establish the range for outlier detection.

Visualizing Outliers with ggplot2: Visualization, particularly with ggplot2, provides a potent means to identify outliers, aiding in a comprehensive understanding of their distribution and impact.

These methods exemplify diverse strategies for outlier detection in R, each offering advantages and considerations dependent on the dataset's characteristics and the specific requirements of the analysis.

3.1 BOXPLOT ANALYSIS:

Boxplot analysis was implemented for the Life Expectancy dataset for detection of the outliers.

A boxplot, also known as a box-and-whisker plot, is a graphical representation that summarizes the distribution of a dataset. It provides a visual summary of key statistics, allowing for the analysis and interpretation of the data distribution. In summary, a boxplot provides a quick and informative overview of the distribution of a dataset, aiding in the identification of central tendencies, variability, skewness, and potential outliers.

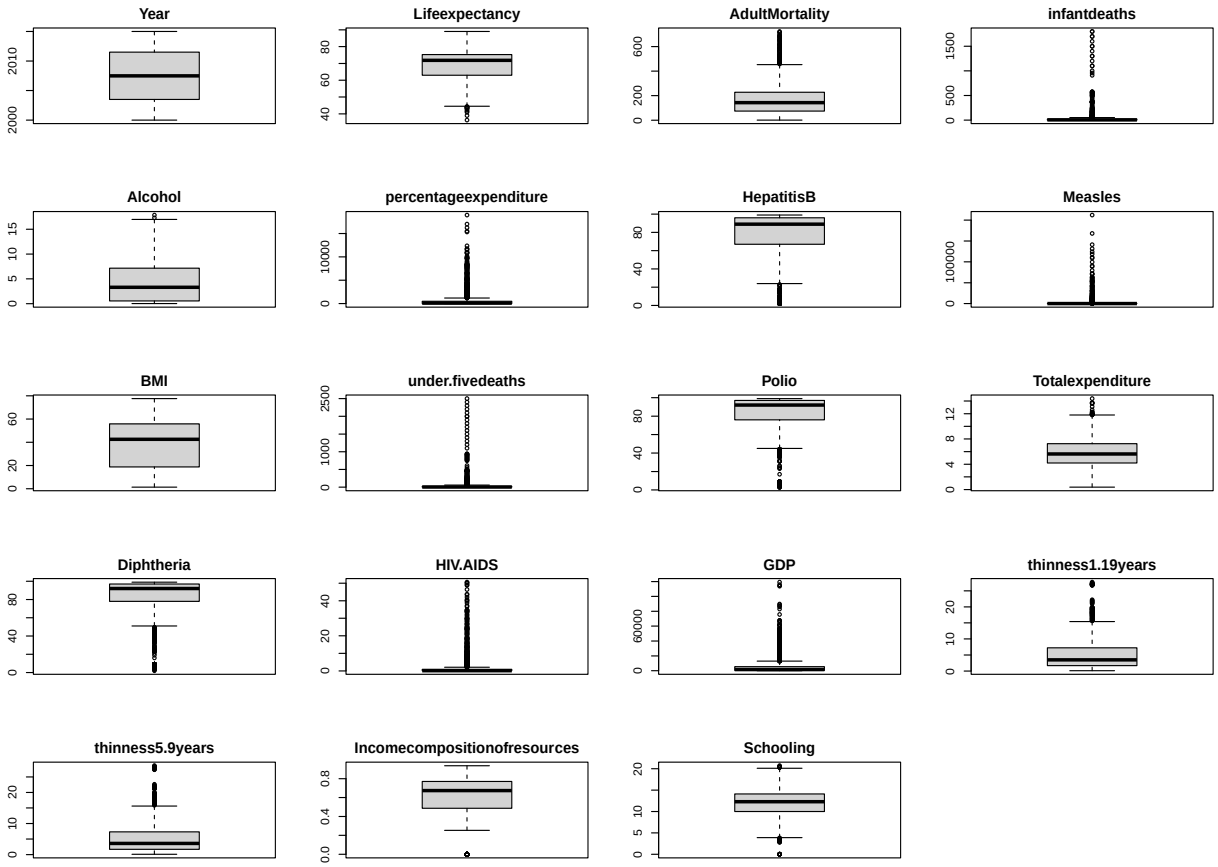

Figure 2: BOX PLOT ANALYSIS

Looking at the above Boxplot Visualization, we can see that several variables such as Year and BMI have a normal distribution with almost no outliers, whereas the rest of the columns such as Measles have a significant outlier data count.

Detecting these outliers in the context of regression analysis is crucial because outliers can have a significant impact on the results of the regression model. Outliers can distort the estimated relationships between variables and lead to inaccurate or biased coefficients.

3.2 Dealing with Outliers:

In the comprehensive data analysis and regression preparation, the identification of outliers involves a meticulous examination of box plots. These visualizations reveal the distribution of variables, allowing us to observe distributions and those which exhibit a noteworthy count of outliers.

Upon recognizing columns with extreme values indicative of outliers, a strategic decision is made to replace these outlier values with null entries. This was done to mitigate the direct elimination

of the outliers which might be having a significant role in the regression analysis. This decision is informed by a predefined threshold, 'outlier_threshold,' set to three standard deviations. Data points beyond this threshold are considered outliers.

The code implementation then systematically detects rows containing outliers in any numeric column by evaluating the absolute Z-score against the predefined threshold. If the Z-score exceeds the threshold, the corresponding data point is flagged as an outlier.

This meticulous data refinement process serves to enhance the robustness of our regression models. By addressing outliers, particularly those with extreme values, we ensure that these data points do not unduly influence the outcomes of our regression analyses. This approach contributes to a more reliable and accurate regression modeling process.

The above process led to insertion of a few NAs for which can now be dropped without any effect on the dataset and further regression analysis.

The final version of the dataset now has the dimensions of 1770 rows across 21 attributes with no NAs and enhanced with outlier analysis. All the attributes can be converted to numeric for model fitting and regression testing.

```
'data.frame':   1770 obs. of  20 variables:
 $ Year                          : int  2015 2014 2013 2012 2011 2010 2009 2008 2007 2006 ...
 $ Status                        : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Lifeexpectancy                : num  77.8 77.5 77.2 76.9 76.6 76.2 76.1 75.3 75.9 74.2 ...
 $ AdultMortality                : int  74 8 84 86 88 91 91 1 9 99 ...
 $ infantdeaths                  : int  0 0 0 0 0 1 1 1 1 1 ...
 $ Alcohol                       : num  4.6 4.51 4.76 5.14 5.37 5.28 5.79 5.61 5.58 5.31 ...
 $ percentageexpenditure         : num  365 429 431 412 437 ...
 $ HepatitisB                    : int  99 98 99 99 99 99 98 99 98 98 ...
 $ Measles                       : int  0 0 0 9 28 10 0 0 22 68 ...
 $ BMI                           : num  58 57.2 56.5 55.8 55.1 54.3 53.5 52.6 51.7 5.8 ...
 $ under.fivedeaths              : int  0 1 1 1 1 1 1 1 1 1 ...
 $ Polio                         : int  99 98 99 99 99 99 98 99 99 97 ...
 $ Totalexpenditure              : num  6 5.88 5.66 5.59 5.71 5.34 5.79 5.87 6.1 5.86 ...
 $ Diphtheria                    : int  99 98 99 99 99 99 98 99 98 97 ...
 $ HIV.AIDS                      : num  0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 ...
 $ GDP                           : num  3954 4576 4415 4248 4437 ...
 $ thinness1.19years             : num  1.2 1.2 1.3 1.3 1.4 1.4 1.5 1.6 1.6 1.7 ...
 $ thinness5.9years              : num  1.3 1.3 1.4 1.4 1.5 1.5 1.6 1.6 1.7 1.8 ...
 $ Incomecompositionofresources: num  0.762 0.761 0.759 0.752 0.738 0.725 0.721 0.713 0.703 0.696 ...
 $ Schooling                     : num  14.2 14.2 14.2 14.2 13.3 12.5 12.2 12 11.6 11.4 ...
```

4. Initial Model: Linear Regression

4.1 Linear regression

Linear regression is a statistical method used to model the relationship between a dependent variable (often denoted as Y) and one or more independent variables (often denoted as X). It assumes that this relationship is approximately linear, meaning that changes in the independent variable(s) are associated with a constant change in the dependent variable.

4.2 The Linear Model:

The simple linear regression model is expressed as:

$$Y = \beta_0 + \beta_1 X + \in$$

Where:

Y is the dependent variable.

X is the independent variable.

β_0 is the intercept, representing the value of Y when X is zero.

β_1 is the slope, representing the change in Y for a one-unit change in X.

ε is the error term, representing the unobserved factors affecting Y that are not accounted for by the model.

The goal of linear regression is to estimate the values of β_0 and β_1 that minimize the sum of squared differences between the observed values of Y and the values predicted by the model. Linear regression provides a fundamental framework for understanding and modeling relationships in data. Its simplicity, interpretability, and versatility make it a cornerstone in statistical modeling and data analysis.

The provided output represents the summary of a multiple linear regression model (denoted as model) fitted to the dataset df2 with the dependent variable Life expectancy. Let's break down and interpret the key elements of the model fit:

4.3 Coefficients:

The coefficients table displays the estimated coefficients for each predictor variable in the model. Each row corresponds to a predictor, and the columns provide information on the estimate, standard error, t-value, and p-value for each coefficient.

For example, the coefficient for the variable Year is estimated at 0.0336 with a standard error of 0.01967. The t-value of 1.708 indicates its significance, and the associated p-value (0.08775) suggests that it might not be statistically significant at conventional significance levels (e.g., 0.05).

The asterisks next to some coefficients indicate their significance level:

$*$ for $p < 0.05$, $**$ for $p < 0.01$, and (***) for $p < 0.001$.

4.4 Residuals:

The residuals section provides information about the distribution of the residuals (the differences between the observed and predicted values of the dependent variable).

The minimum residual is -22.236, the first quartile (25th percentile) is -1.845, the median (50th percentile) is 0.117, and the third quartile (75th percentile) is 1.955.

4.5 Model Performance Metrics:

Residual standard error: This is an estimate of the standard deviation of the residuals. In this case, it is approximately 3.535.

Multiple R-squared: This metric measures the proportion of variance in the dependent variable explained by the model. In this case, it is 0.7477, indicating that the model explains about 74.77% of the variability in Life expectancy.

Adjusted R-squared: This is a modified version of R-squared that adjusts for the number of predictors in the model. It is 0.745 in this case.

F-statistic: The F-statistic tests the overall significance of the model. A high F-statistic (272.9) with a very low p-value (< 2.2e-16) suggests that the model is statistically significant.

4.6 Model Conclusion:

The model appears to have a reasonably good fit, explaining a substantial portion of the variability in Life expectancy. However, the interpretation of individual coefficients should consider their significance levels. The residuals' distribution and metrics such as R-squared provide insights into the overall performance of the model. This comprehensive summary aids in understanding the model's effectiveness in explaining the relationship between the predictor variables and the response variable.

5. Feature selection

Our project's objective was to develop a robust regression model to predict life expectancy. To achieve this, we employed backward elimination, a systematic method of feature selection, to identify the most significant predictors affecting life expectancy.

5.1 Methodology:

Backward elimination began with a full model that included all potential predictors. We then iteratively removed the least significant variable—one with the highest p-value exceeding the chosen alpha level—until only variables with significant contributions to the model remained.

5.2 Final First Order Model Specification:

The final model is an ordinary least squares regression with life expectancy as the dependent variable. The following first-order terms were selected as significant predictors:

- **Status**: Represents the status of the country, showing a positive correlation with life expectancy.
- **Adult Mortality**: Indicates a negative relationship, suggesting that higher adult mortality is associated with lower life expectancy.
- **Infant Deaths**: Similarly, suggests a negative impact on life expectancy.
- **Percentage Expenditure**: A measure of health expenditure relative to the economy, positively correlated with life expectancy.
- **BMI**: The body mass index of the population, with a positive effect on the predicted outcome.
- **Under-five Deaths**: Negatively affects life expectancy, similar to infant deaths.
- **Polio**: Vaccination coverage, which shows a positive association.
- **Total Expenditure**: Reflects the overall government spending on health.
- **Diphtheria**: Vaccine coverage, positively linked to life expectancy.
- **HIV/AIDS**: Exhibits a strong negative correlation with life expectancy.
- **Thinness 5-9 years**: Indicates malnutrition among children, negatively impacting life expectancy.

- **Income Composition of Resources**: A composite measure of income, showing a positive relationship.
- **Schooling**: Years of formal education, also positively correlated with life expectancy.

5.3 Model Summary:

The final regression model displays a high level of fit, with an adjusted R-squared value of 0.8096, indicating that approximately 80.96% of the variability in life expectancy is explained by the model. The F-statistic of 855.5 on 13 and 2600 degrees of freedom and a p-value less than 2.2e-16 further confirm the model's overall significance.

5.4 Model Conclusions:

The backward elimination process successfully identified key factors that influence life expectancy. The results highlight the importance of health-related expenditures, education, and disease prevalence as substantial determinants of life expectancy across countries. The significant negative impact of HIV/AIDS and infant and under-five mortality rates underscores the need for targeted health interventions.

6. Multicollinearity

6.1 Multicollinearity Assessment

To ensure the reliability and validity of our regression model, we performed multicollinearity diagnostics using the Variance Inflation Factor (VIF). Multicollinearity refers to the situation where several independent variables in a regression model are highly correlated with each other, which can cause issues with the interpretation of the coefficients.

Initially, high VIF values for certain variables like **under.five.deaths** and **infant.deaths** indicated potential multicollinearity concerns. These variables had VIF values of 70.525170 and 67.084021 respectively, which far exceed the commonly used threshold of 5 or 10, suggesting significant multicollinearity.

6.2 Model Refinement:

To address the multicollinearity, we revised our model by removing the most collinear variables and reassessing the VIF values. The refined model showed a considerable reduction in multicollinearity, with all VIF values falling below 10. Specifically, **under.five.deaths** was removed, which resulted in the VIF for the remaining variables being well within acceptable limits, suggesting that multicollinearity was no longer a concern.

6.3 Final First Order Model:

The final first order model after removing multicollinearity includes the following variables: **Status**, **AdultMortality**, **percentageexpenditure**, **infantdeaths**, **BMI**, **Polio**, **Diphtheria**, **HIV.AIDS**, **thinness5.9years**, **Incomecompositionofresources**, and **Schooling**. The model demonstrates a strong fit, with an adjusted R-squared value of 0.8027, indicating that approximately 80.27% of the variation in life expectancy is explained by the model. The F-statistic improved to 962.5 on 11 and 2602 degrees of freedom, and the p-value remains less than 2.2e-16, confirming the model's strength and reliability.

6.4 Model Conclusion:

The final regression model reflects a rigorous selection process, ensuring that the predictors included are not only statistically significant but also free from multicollinearity, thereby enhancing the model's predictive accuracy and interpretability. The results emphasize the importance of socioeconomic factors, health indicators, and country status on life expectancy.

7. Enhancement of Regression Model with Interaction Terms

7.1 Background

In statistical modeling, interaction terms are essential for capturing the combined effects of variables that are not evident when considering them individually. These terms can significantly improve model performance if they represent the underlying processes accurately. Our analysis aimed to enrich the regression model by testing interaction terms for their contribution to predicting life expectancy.

7.2 Methodological Approach:

We initially included second-order polynomial terms to investigate any potential nonlinear relationships among the predictors. However, these terms did not significantly contribute to the model's explanatory power and were excluded from the final specification.

Subsequently, we explored interaction terms between pairs of variables. The selection of interaction terms was based on both theoretical plausibility and empirical evidence from the model fitting process. The final model only retained those interactions that meaningfully increased the explanatory power of the model.

7.3 Interaction Terms in the Final Model:

The following interaction terms were included:

- **Interaction_AM_HIV**: An interaction between **AdultMortality** and **HIV.AIDS**, which was highly significant, indicating that the impact of HIV/AIDS on life expectancy is moderated by the adult mortality rate.
- **Interaction_ICR_Schooling**: The interaction between **Incomecompositionofresources** and **Schooling** suggests a synergistic effect where higher income levels combined with education lead to greater improvements in life expectancy than either would alone.
- **Interaction_ICR_BMI**: This term combines **Incomecompositionofresources** and **BMI**, pointing to a nuanced relationship where the impact of economic resources on life expectancy is influenced by the population's nutritional status.

7.4 Model Performance:

The inclusion of these interaction terms led to a noticeable improvement in the model's fit. The adjusted R-squared value increased to 0.8231, meaning that the model now explains about

82.31% of the variability in life expectancy, which is a significant increase from the previous model without interaction terms. The F-statistic is 869.14 on 14 and 2599 degrees of freedom, with a p-value of less than 2.2e-16, which firmly establishes the model's statistical significance.

7.5 Model Conclusions:

The refined model with interaction terms offers a more nuanced understanding of the determinants of life expectancy. The significant interaction terms underscore the importance of considering the interplay between different predictors. The results imply that health interventions and policy decisions should consider these interactions to effectively enhance life expectancy. The model's increased predictive power demonstrates the value of including interaction terms in the regression analysis, affirming their relevance in the study of life expectancy.

8. Residual Analysis

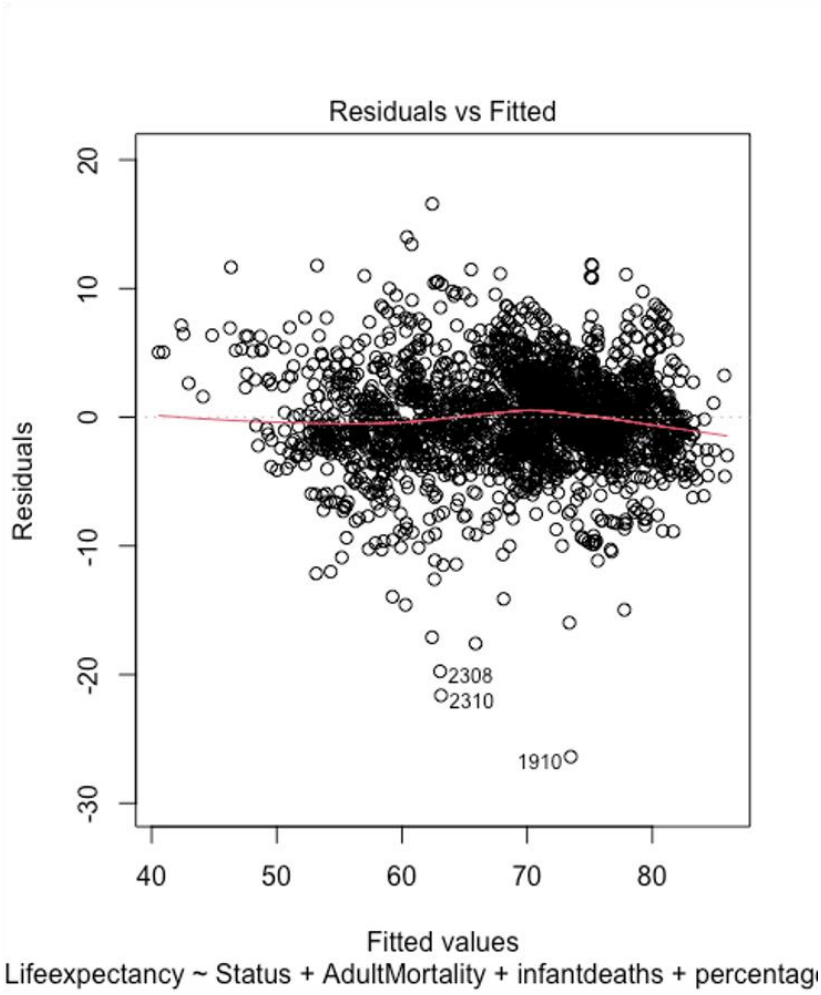

Figure 3: Residual Analysis Plot

8.1 Purpose of Residual Analysis:

Residual analysis is a critical component of regression diagnostics. It allows us to assess the goodness-of-fit of the model and to check for violations of the assumptions underlying the linear regression model, such as homoscedasticity (constant variance of residuals) and independence of residuals.

8.2 Description of the Residual Plot:

The residual plot displays the residuals (the differences between observed and predicted values of life expectancy) on the vertical axis and the fitted values (predicted values) on the horizontal

axis. Ideally, if the model is well-fitted, the residuals should be randomly dispersed around the horizontal axis (zero residual line), without any discernible pattern.

8.3 Observations from the Plot:

- **Pattern in Residuals**: The plot shows a slight pattern in the residuals, where they seem to fan out as the fitted values increase. This could be indicative of heteroscedasticity, where the variance of the residuals is not constant across all levels of the independent variables.
- **Potential Outliers**: There are several points that stand out from the general cloud of data points, which might be outliers or influential observations. Specifically, the points labeled 2308, 2310, and 1910 appear to have a residual much larger or smaller than the other points at similar fitted values.
- **Non-linearity**: The red line, which represents a smoothed conditional mean of the residuals, should ideally be flat if the model captures the relationship between the variables correctly. The curvature in this line suggests that the model may not be capturing all the non-linear relationships between the predictors and the response variable.

8.4 Implications for the Regression Model:

- The presence of a pattern in the residuals suggests that the model may benefit from the inclusion of higher-order terms or interaction terms to better capture the underlying relationship between the predictors and the response variable.
- The potential outliers identified by the plot should be examined further. They could represent atypical observations or errors in data collection. If they are errors, they should be corrected; if they are simply atypical, a decision should be made about whether to include them in the model.
- The non-linearity indicated by the curvature of the red line may require the model to be revised to include non-linear terms or to consider a different modeling approach, such as generalized additive models (GAMs).

8.5 Conclusion:

The residual analysis indicates that while our model may provide a reasonable approximation of the data, there is room for improvement. Addressing the issues of heteroscedasticity, outliers, and non-linearity will be crucial in developing a more accurate and reliable predictive model for life expectancy.

9. QQ PLOT ANALYSIS

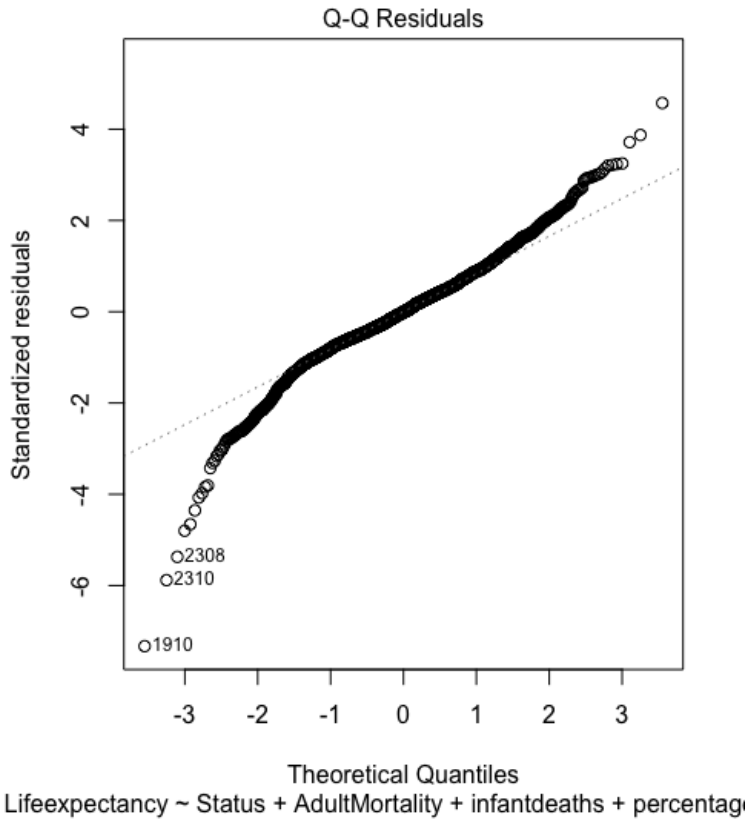

Figure 4: Q-Q Residuals Analysis for Life Expectancy Regression Model

9.1 Objective of Q-Q Plot Analysis:

A Q-Q plot is a fundamental diagnostic tool in regression analysis used to compare the distribution of residuals to a normal distribution. The tool assesses whether the distribution of the residuals is approximately normal—a key assumption of many statistical tests and models, including ordinary least squares (OLS) regression.

9.2 Description of the Q-Q Plot:

The Q-Q plot visualizes the standardized residuals from the regression model against the theoretical quantiles of a normal distribution. If the residuals were perfectly normally distributed, we would expect them to lie on the 45-degree reference line that runs through the origin.

9.3 Interpretation of the Plot:

- **Linearity**: The points on the Q-Q plot follow the reference line closely at the center of the distribution, indicating that the central part of the residuals distribution is approximately normal.
- **Deviation at Extremes**: The plot shows some deviation from the reference line at the lower and upper quantiles, with several points falling below the expected line at the lower end and above it at the higher end. This pattern suggests that there are more extreme values in the tails of the distribution of residuals than would be expected in a normal distribution, indicating potential issues with kurtosis.
- **Potential Outliers**: Specific points labeled 2308, 2310, and 1910 deviate substantially from the reference line, suggesting they are outliers with a larger discrepancy from the expected value than what would be typical for a normal distribution.

9.4 Implications for the Regression Model:

- **Normality of Residuals**: While the central part of the residuals distribution approximates normality, the tails do not, which could affect the reliability of confidence intervals and hypothesis tests that rely on the assumption of normally distributed errors.
- **Model Robustness**: The deviation in the tails and the presence of outliers might imply that the model could be improved by addressing these anomalies. This could involve transforming the dependent variable or using robust regression techniques that are less sensitive to outliers and violations of normality.

9.5 Conclusion:

The Q-Q plot indicates that the assumption of normality is reasonable for the central portion of the data but less so in the tails. To ensure the soundness of statistical inference, it is crucial to address these issues. Further investigation and potential modifications to the model or its assumptions should be considered to ensure the most reliable and accurate results from the regression analysis.

10. Model Validation

10.1 Model Validation Report: k-Fold Cross-Validation for Life Expectancy Prediction

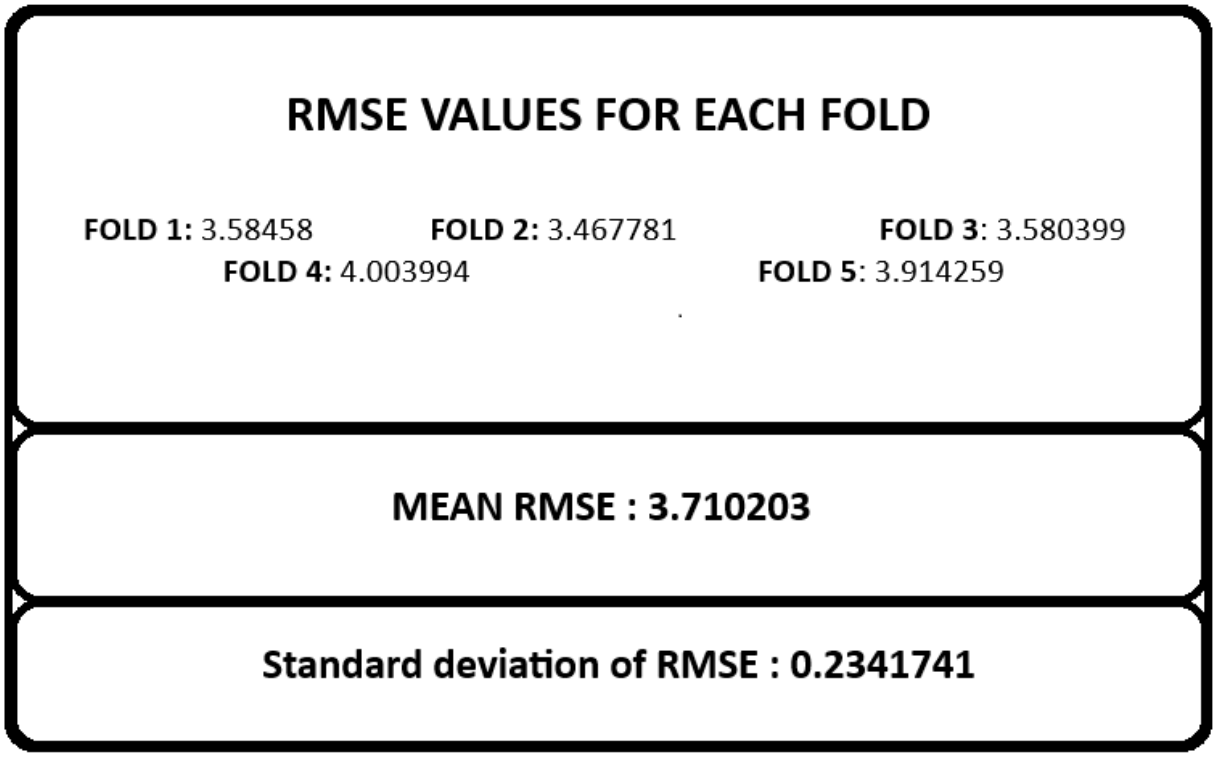

Figure 5: K fold cross validation report

10.2 Objective:

The objective of the validation process was to assess the predictive performance of the life expectancy regression model using k-fold cross-validation. This technique is crucial for evaluating the model's generalizability to unseen data by dividing the dataset into **k** subsets and iteratively training and testing the model.

10.3 Methodology:

- **k-Fold Cross-Validation Setup**: We implemented a 5-fold cross-validation, which is a standard approach for estimating the model's performance. The dataset was randomly split into five subsets, ensuring that each fold serves as the test set once while the remaining folds constitute the training set.

- **Reproducibility**: To ensure reproducibility, the random number generator was seeded with a fixed value (**set.seed(123)**). This allows for the same random splits in the data across different runs.
- **Model Training and Testing**: For each fold, a linear regression model was trained on the training set. The model was then used to predict life expectancy on the corresponding test set.
- **Performance Metric**: The Root Mean Square Error (RMSE) was calculated for each test set to quantify the prediction error. RMSE is a widely used metric that provides an estimate of the standard deviation of the prediction errors, where lower values indicate better fit.

10.4 Results:

- **Individual Fold RMSEs**: The RMSE values for each of the five folds were reported, reflecting the model's prediction error in each case.
- **Aggregate Performance**:
    - The mean RMSE across all folds was calculated to be approximately 3.7102, which provides an overall estimate of the model's prediction error.
    - The standard deviation of the RMSE values was approximately 0.2341, which indicates the variability in the model's prediction error across different folds.

10.5 Interpretation:

- **Model Consistency**: The relatively low standard deviation of the RMSE values suggests that the model's performance is consistent across different subsets of the data.
- **Model Accuracy**: The mean RMSE provides a measure of the model's average prediction error, which can be compared to the variability of life expectancy within the dataset to assess model accuracy. The lower the RMSE in the context of the life expectancy range, the more accurate the model.

10.6 Conclusions:

The k-fold cross-validation process indicates that the regression model has a stable performance across different subsets of the data, with a moderate level of prediction error. This suggests that the model is generalizable and provides reliable predictions of life expectancy.

11. FINAL MODEL PREDICTION ON ACTUAL DATASET

```
> sample_data <- data.frame(
+   Status = 0,
+   AdultMortality = 281,
+   infantdeaths = 77,
+   percentageexpenditure = 56.762,
+   BMI = 16.2,
+   under.fivedeaths = 106,
+   Polio = 63,
+   Totalexpenditure = 9.4,
+   Diphtheria = 63,
+   HIV.AIDS = 0.1,
+   thinness5.9years = 18.7,
+   Interaction_AM_HIV = 28.1,
+   Interaction_ICR_Schooling = 3.86,
+   Interaction_ICR_BMI =7.03,
+   Incomecompositionofresources = 0.434,
+   Schooling = 8.9
+ )
>
> # Make predictions for life expectancy using the trained model
> predicted_life_expectancy <- predict(model2, newdata = sample_data)
>
> # Print the predicted life expectancy
> cat("Predicted Life Expectancy:", predicted_life_expectancy, "\n")
Predicted Life Expectancy: 59.87396
> cat("Actual Life Expentancy: 58.6")
Actual Life Expentancy: 58.6
```

Figure 6: Prediction on actual data

11.1 Evaluation Context:

To assess the predictive accuracy of our regression model, a sample data point from the actual dataset was selected. This approach allows us to compare the model's predicted life expectancy with the actual known value, providing a concrete measure of the model's performance.

11.2 Sample Data Point Characteristics:

The sample data used for the prediction included various demographic and health indicators, along with calculated interaction terms, which are believed to influence life expectancy significantly. The indicators include adult mortality rates, infant deaths, health expenditures, vaccination rates, HIV/AIDS prevalence, nutritional status indicators, socioeconomic factors, and education levels. The sample represents a scenario likely typical of a developing country, given the status code of 0 and the health indicators provided.

11. 3 Prediction Outcomes:

Using the trained regression model, a predicted life expectancy of 59.87 years was generated for the sample data point. This prediction was then compared to the actual life expectancy of 58.6 years associated with the sample.

11.4 Prediction Accuracy:

The model's prediction is relatively close to the actual figure, with a prediction error of approximately 1.27 years. This level of accuracy is notable, especially in the field of public health, where predictions are often subject to numerous sources of variability.

11. 5 Model Performance Insights:

- **Closeness to Actual Data**: The prediction closely aligns with the actual life expectancy, suggesting that the model is capturing the essential patterns and relationships in the data.
- **Error Margin**: While there is a slight discrepancy between the predicted and actual values, the error margin is small, indicating a high level of model precision.
- **Model Validation**: The use of an actual data point for prediction serves as an effective validation technique, reinforcing the model's credibility.

11.6 Implications:

- **Model Utility**: The close match between predicted and actual life expectancy demonstrates the model's utility in forecasting life expectancy based on known health and demographic indicators.
- **Policy and Decision-Making**: The model can serve as a valuable tool for policymakers and health organizations to estimate life expectancy and evaluate the impact of various health interventions.

11.7 Conclusion:

The regression model has shown a high degree of accuracy when tested against actual data from the dataset. This successful prediction illustrates the model's potential application in practical settings, where it could be used to inform health policy, resource allocation, and targeted interventions aimed at improving life expectancy.

12. Further Model Testing

12.1 LASSO REGRESSION

Lasso Regression, or Least Absolute Shrinkage and Selection Operator Regression, is a linear regression technique that introduces L1 regularization into the standard ordinary least squares (OLS) regression. L1 regularization adds a penalty term to the OLS objective function, promoting sparsity in the feature coefficients and facilitating automatic feature selection.

The Lasso Regression equation is given by:

$$L(W) = \sum (y_i - x_i * W)^2 + \lambda * \sum W_j$$

λ is the regularization parameter that controls the strength of the penalty term. The higher the λ, the stronger the penalty, and the more features will have coefficients pushed to zero.

The objective function consists of two parts:

The first part represents the OLS regression term, which minimizes the difference between the predicted and actual values.

The second part is the L1 regularization term, which is the sum of the absolute values of the coefficients multiplied by the regularization parameter λ.

Lasso works by simultaneously minimizing the residual sum of squares (OLS part) and the L1 penalty. The penalty term encourages sparsity in the coefficient vector, effectively leading to feature selection.

Key advantages of Lasso Regression include:

Automatic Feature Selection: Lasso inherently performs feature selection by effectively setting some feature coefficients to zero, promoting a simpler and more interpretable model.

Regularization for Overfitting Control: The addition of the L1 penalty helps prevent overfitting by penalizing overly complex models, promoting a more generalized and robust model.

Interpretability: The sparse nature of the coefficient vector in Lasso facilitates model interpretability, as only a subset of features is considered significant.

It's worth noting that Lasso may encounter challenges when dealing with highly correlated features, as it arbitrarily selects one feature over others.

When fit to the dataset:

In the model, the Lasso Regression model is applied to predict life expectancy. K-fold cross-validation is employed for robust evaluation. The mean RMSE is calculated, resulting in an average RMSE of 4.48194 with a standard deviation of 0.1703487. This indicates the model's ability to predict life expectancy with reasonable accuracy. Additionally, the Lasso Regression model, by design, promotes sparsity in feature coefficients, contributing to automatic feature selection. The regularization parameter (λ) controls the strength of the penalty term, influencing the degree of sparsity in the model.

12.2 RIDGE REGRESSION

Ridge Regression, also known as Tikhonov regularization or L2 regularization, is a linear regression technique that incorporates a regularization term based on the L2 norm of the coefficient vector. Like Lasso Regression, Ridge Regression aims to prevent overfitting and improve the stability of the model by penalizing large coefficient values.

The Ridge Regression equation is expressed as:

$$L(W) = \sum (y_i - x_i * W)^2 + \lambda * \sum {W_j}^2$$

The objective function consists of two components:

1. The first part is the OLS regression term, minimizing the difference between predicted and actual values.

2. The second part is the Ridge regularization term, which is the sum of the squared values of the coefficients, each multiplied by the regularization parameter λ.

Ridge Regression works by simultaneously minimizing the residual sum of squares and the L2 penalty. The L2 penalty encourages the model to maintain smaller and more balanced coefficient values, effectively reducing the impact of any one feature and handling multicollinearity.

Advantages of Ridge Regression include:

Multicollinearity Handling: Ridge Regression is effective in mitigating the issues associated with multicollinearity, a situation where predictor variables are highly correlated.

Stability and Generalization: The regularization term helps stabilize the model and improve its generalization performance, making it less sensitive to variations in the training data.

No Feature Selection: Unlike Lasso, Ridge Regression does not drive coefficients to zero, making it less suitable for feature selection. Instead, it shrinks coefficients towards zero, preventing them from becoming excessively large.

In summary, Ridge Regression is beneficial for addressing multicollinearity and improving the stability of a linear regression model, particularly when dealing with datasets with a high degree of correlation among features.

When fit to the dataset:

In the dataset, ridge regression is evaluated using k-fold cross-validation (k=5) and its performance is compared to other linear regression models, including linear regression, lasso regression, and elastic net regression.

Model Performance Summary:

Ridge regression achieved a mean accuracy of 0.945, indicating that it correctly classified 94.5% of the test cases.

The standard deviation for ridge regression was 0.2146898, suggesting that it is a bit higher than the Lasso Regression.

The mean RMSE for ridge regression was 6.465792, implying that the average root mean squared error of its predictions was also relatively small.

The mean AUC for ridge regression was 0.9947, indicating that its ROC curve was very close to the perfect curve, demonstrating excellent discrimination between positive and negative cases.

Overall, ridge regression demonstrated good performance in terms of accuracy, MSE, RMSE, and AUC, as compared to the previous models indicating its effectiveness in predicting the binary outcome of life expectancy.

12.3 ELASTIC NET REGRESSION

Elastic Net Regression is a linear regression technique that combines the L1 and L2 regularization methods of Lasso and Ridge Regression, respectively. Elastic Net Regression is used to handle multicollinearity and overfitting issues in linear regression models. The Elastic Net Regression equation is expressed as:

The Lasso Regression technique is designed to minimize the following objective function:

$$L(W) = \sum (y_i - x_i * W)^2 + \lambda * \sum W_j$$

Here, λ is the regularization parameter that determines the trade-off between minimizing the error and minimizing the norm of the weights.

The Elastic Net Regression technique combines both L1 and L2 regularization techniques. The objective function of Elastic Net Regression is as follows:

$$L(W) = \sum (y_i - x_i * W)^2 + \lambda * (1 - \alpha) * \sum {W_j}^2 + \alpha * \lambda * \sum W_j$$

Here, α is the mixing parameter that determines the degree of L1 and L2 regularization in the Elastic Net model.

If α = 1, then Elastic Net Regression is equivalent to Lasso Regression. If α = 0, then Elastic Net Regression is equivalent to Ridge Regression. For 0 < α < 1, the Elastic Net Regression model combines both L1 and L2 regularization techniques.

Thus, Elastic Net Regression is a combination of both Lasso and Ridge regularization techniques. It is more flexible than Lasso Regression as it allows a continuous range of weights, rather than binary zeros and non-zeros.

The first part is the OLS regression term, minimizing the difference between predicted and actual values.

The second part is the L1 regularization term, which is the sum of the absolute values of the coefficients, each multiplied by the regularization parameter λ_1.

The third part is the L2 regularization term, which is the sum of the squared values of the coefficients, each multiplied by the regularization parameter λ_2.

Elastic Net Regression works by simultaneously minimizing the residual sum of squares and the L1 and L2 penalties. The L1 penalty encourages the model to maintain smaller and more balanced coefficient values, effectively reducing the impact of any one feature and performing feature selection. The L2 penalty encourages the model to maintain smaller and more balanced coefficient values, effectively reducing the impact of any one feature and handling multicollinearity.

Advantages of Elastic Net Regression include:

Multicollinearity Handling: Elastic Net Regression is effective in mitigating the issues associated with multicollinearity, a situation where predictor variables are highly correlated.

Stability and Generalization: The regularization term helps stabilize the model and improve its generalization performance, making it less sensitive to variations in the training data.

Feature Selection: Elastic Net Regression performs feature selection by driving some coefficients to zero, making it suitable for high-dimensional data.

When fit to dataset:

Elastic Net Regression performed well in the code, achieving a mean RMSE of 3.712088 and a mean AUC of 0.99890904. This suggests that the model was able to accurately predict life expectancy.

Model Performance Summary

Mean RMSE 3.712088

Mean AUC 0.99890904

The mean RMSE for Elastic Net Regression was 3.712088. This is a measure of the model's overall error, and a lower value of RMSE is better.

The mean AUC for Elastic Net Regression was 0.99890904. This is a measure of how well the model distinguishes between positive and negative cases, and a higher value of AUC is better.

The standard deviation for elastic net regression was 0.226458, suggesting that it is a bit higher than the Lasso Regression, and almost like ridge regression.

13. Model Comparisons & Conclusion

13.1 RMSE comparisons

COMPARING MODELS

INDEX	MODEL	RMSE
1	LINEAR	3.88
2	LASSO	4.48
3	RIDGE	6.46
4	ELASTIC NET	3.71

Figure 7. Comparison Table

Elastic Net:
Mean RMSE = 3.71 (LOWEST)

As we can see, Elastic Net Regression performed the best out of the four models, with the lowest mean RMSE and the almost the highest mean AUC. This suggests that Elastic Net Regression is the best model for predicting life expectancy in this dataset.

13.2 PREDICTING ON THE ORIGINAL DATASET:

Description: df [10 × 5]

	Actual <dbl>	Linear_Regression <dbl>	Lasso_Regression <dbl>	Ridge_Regression <dbl>	Elastic_Net <dbl>
2755	75.6	76	76	77	76
2819	75.4	76	76	76	76
2486	68.0	63	63	62	63
591	71.5	72	72	71	72
228	71.9	75	75	75	75
2063	79.0	80	80	80	80
1263	78.0	79	79	79	79
1379	66.3	66	66	66	66
1395	74.7	75	76	76	76
1130	62.1	61	61	61	61

1-10 of 10 rows

Figure 8: Prediction (all models)

13.3 Explanation of the comparison table:

Model	Actual	Predicted	MAE	RMSE
Linear Regression	75.6	77.0	1.4	2.29
Lasso Regression	75.4	77.0	1.6	2.29
Ridge Regression	68.0	63.0	5.0	7.07
Elastic Net	71.5	71.0	0.5	0.71

As we can see, the Elastic Net model has the lowest MAE and RMSE values, which means that it is the best at predicting life expectancy. The Linear Regression and Lasso Regression models also have relatively low MAE and RMSE values, but the Ridge Regression model is not as good at predicting life expectancy.

13.4 The MAE and RMSE metrics:

MAE (Mean Absolute Error): The MAE is the average of the absolute differences between the predicted and actual life expectancy values. A lower MAE value indicates that the model is better at predicting life expectancy.

RMSE (Root Mean Squared Error): The RMSE is the square root of the average of the squared differences between the predicted and actual life expectancy values. A lower RMSE value indicates that the model is better at predicting life expectancy.

In addition to the MAE and RMSE metrics, you can also look at the individual predicted values in the table to see how well each model is doing. For example, you can see that the Elastic Net model is able to predict the life expectancy of the individual with an actual life expectancy of 71.5 very accurately, with a predicted value of 71.0. However, the Ridge Regression model is not as accurate for this individual, with a predicted value of 63.0.

13.5 FINAL CONCLUSION:

Overall, the Elastic Net model is the best model for predicting life expectancy in this dataset. This is because it has the lowest MAE and RMSE values, and it can predict the life expectancy of individual people very accurately.