# ALY 6015 Intermediate Analytics

## Module 6

## Final Project Report

## Analysis of Life Expectancy: Exploring Socio-Economic and Health Factors

**For: Prof. Samuel Li**

**Prepared by: Zaili Gu, Benren Cai, Qiwei Guo**

**Date:February 12, 2025**

# Introduction

Life expectancy is a crucial indicator of a country's overall health and socio-economic development. Understanding the factors influencing life expectancy can help policymakers and public health officials implement targeted interventions to improve population well-being. This project aims to analyze the Life Expectancy Dataset to identify key determinants affecting life expectancy across different countries and assess their significance using statistical and machine learning techniques.

## Objectives

**1. Identify significant predictors of life expectancy** – Examining factors such as health indicators (e.g., Adult Mortality, Alcohol Consumption), socio-economic variables (e.g., GDP per Capita, Schooling), and demographic aspects (e.g., Population, Status: Developed/Developing) to determine their impact.

**2. Perform statistical hypothesis testing** – Conducting T-test and Chi-Square tests to explore differences in life expectancy based on categorical factors and relationships between key variables.

**3. Develop predictive models** – Implementing Logistic Regression to classify countries into high and low life expectancy groups and LASSO Regression to identify the most influential variables for life expectancy prediction.

**4. Evaluate model performance** – Assessing predictive accuracy using AUC, RMSE, and misclassification rates to determine the effectiveness of different models.

**5. Provide data-driven insights** – Interpreting results to understand how different variables contribute to life expectancy and suggesting potential improvements in data preprocessing and model selection for better predictive performance.

## Qestions

1. Does a country's development status (Developed vs. Developing) impact life expectancy?
2. Which factors significantly influence life expectancy?
3. Is there an association between a country's development status and child mortality?
4. Can we predict whether a country has high or low life expectancy based on health and economic factors?
5. Does adding Ridge/LASSO Regularization improve the predictive power of our model?

# Exploratory Data Analysis

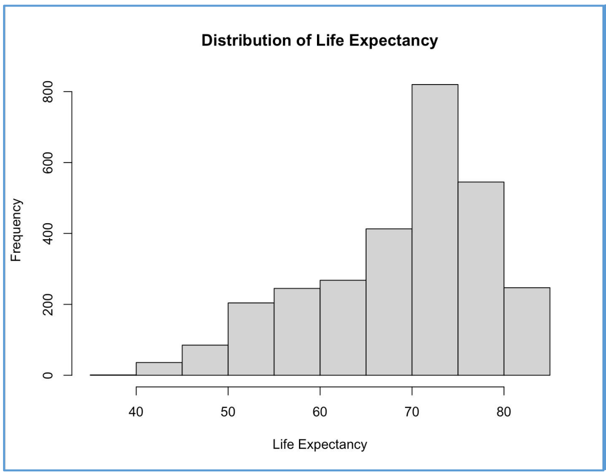
## Dataset Description

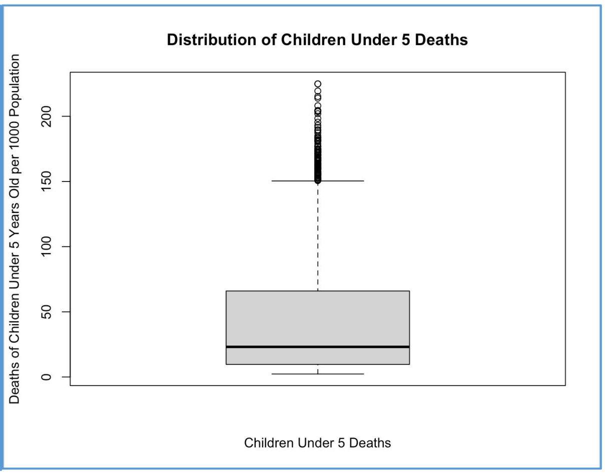
The **Life Expectancy** Dataset from Kaggle contains health, economic, and demographic indicators for multiple countries over several years. It includes variables such as Life Expectancy, Adult Mortality, Schooling, GDP per Capita, Alcohol Consumption, and Population, among others. The dataset categorizes countries as Developed or Developing and provides insights into how various factors impact life expectancy. Some variables contain missing values, requiring preprocessing. This dataset is suitable for statistical analysis and predictive modeling to identify key determinants of life expectancy and assess their significance.

## Distribution of main variables

Distribution of Life Expectancy: The life expectancy values are primarily concentrated between 70 and 80 years, indicating a significant portion of countries in the dataset have life expectancies within this range.

Distribution of Children Under 5 deaths: The number of deaths of children under five years old per 1000 population predominantly falls between 20 and 60 deaths per 1000, highlighting a common range for child mortality across the countries in the dataset.





## Correlation Analysis

A correlation matrix was generated for all the numeric variables in the dataset to examine the relationships between life expectancy and other factors. The analysis revealed that certain variables exhibit a strong positive correlation with life expectancy. Specifically:

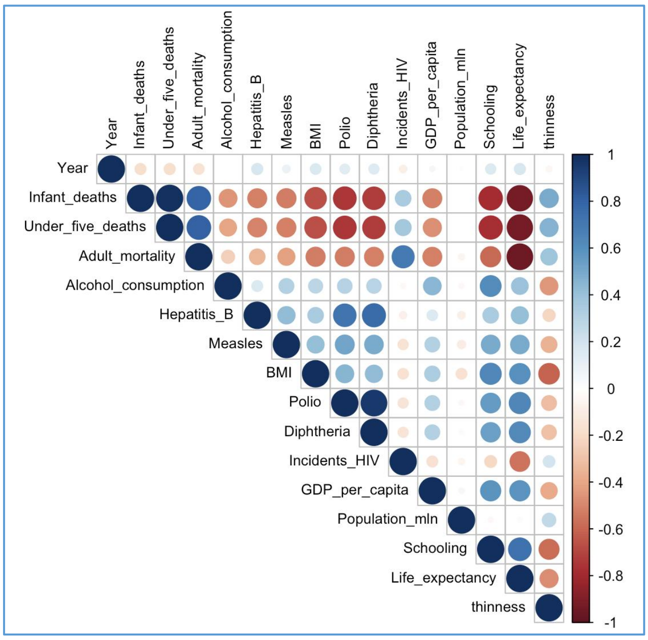
- Schooling: A higher level of education is strongly associated with increased life expectancy, suggesting that access to education plays a critical role in improving health outcomes.

- GDP: Greater Gross Domestic Product (GDP) per capita is positively correlated with life expectancy, reflecting the potential influence of economic prosperity on public health and healthcare access.

- BMI (Body Mass Index): A higher average BMI is associated with longer life expectancy, possibly indicating a correlation with better overall nutrition and healthcare.

- Polio: A strong positive relationship with life expectancy indicates the significant impact of polio vaccination programs on improving health and survival rates.

This correlation matrix highlights these factors as key determinants of life expectancy in the dataset.



# Methods

## Methods Will be Applied in This Analysis

1. **T-test:** Comparing mean life expectancy across two group.
2. **Multiple Linear Regression:** Identify the variables which influence the life expectancy.
3. **Chi-Square Test for Independence:** Explore the association between country's state and child morality.
4. **Logistic Regression:** To predict the high or low life expectancy by health and economic factors.
5. **Regularization (Ridge & Lasso):** Prevent to model from over-fitting and improve model.

# Analysis

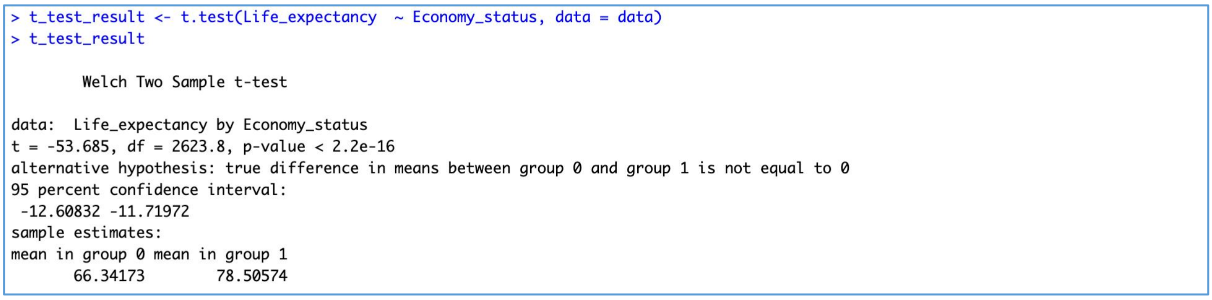
## Question 1: Does a country's development status (Developed vs. Developing) impact life expectancy?

### Method: T-test (comparing mean life expectancy across status)

**State the hypothesis**

- *H0: There is no significant difference in mean life expectancy between developed and developing countries.*
- *H1: There is a significant difference in mean life expectancy between developed and developing countries.*

## Perform a T-test



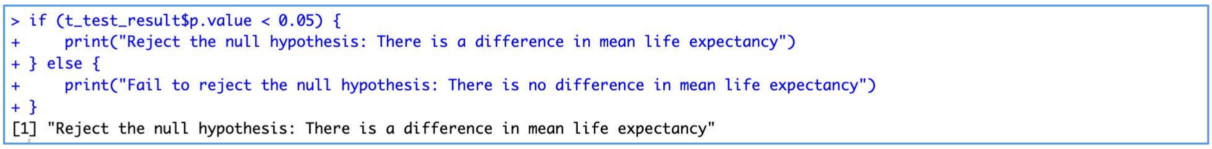
```
> t_test_result <- t.test(Life_expectancy  ~ Economy_status, data = data)
> t_test_result

	Welch Two Sample t-test

data:  Life_expectancy by Economy_status
t = -53.685, df = 2623.8, p-value < 2.2e-16
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
 -12.60832 -11.71972
sample estimates:
mean in group 0 mean in group 1 
       66.34173        78.50574 
```

The Welch Two-Sample T-test was performed to compare the means of Life Expectancy between two groups: Developing (Group 0) and Developed (Group 1) countries.

## Make a decision



```
> if (t_test_result$p.value < 0.05) {
+     print("Reject the null hypothesis: There is a difference in mean life expectancy")
+ } else {
+     print("Fail to reject the null hypothesis: There is no difference in mean life expectancy")
+ }
[1] "Reject the null hypothesis: There is a difference in mean life expectancy"
```

## Summarize the result

The p-value is extremely small (less than 0.05), indicating strong evidence to reject the null hypothesis. This suggests a statistically significant difference in life expectancy between developed and developing countries.

• Developed countries have a significantly higher mean life expectancy (78.51 years) compared to developing countries (66.34 years).

• The 95% confidence interval for the difference in means suggests that life expectancy in developed countries is between 11.72 and 12.61 years higher than in developing countries.

## Conclusion

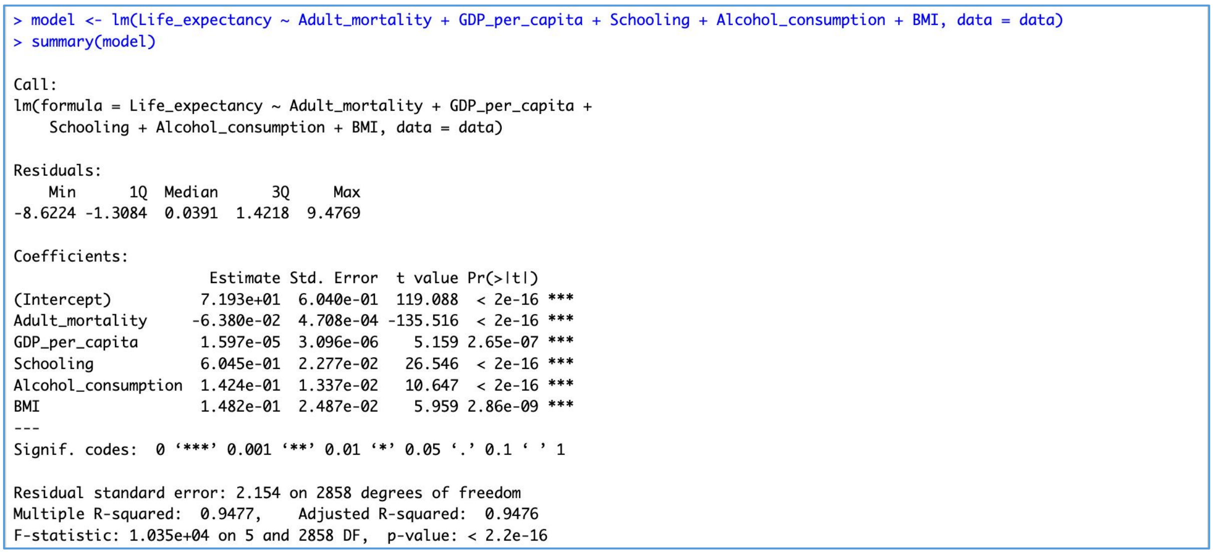
This result confirms that economy status (developed vs. developing) has a significant impact on life expectancy, with developed countries consistently showing higher life expectancy.

# Question 2: Which factors significantly influence life expectancy?

### Method: Multiple Linear Regression

## Fit a regression model

A linear regression model was fitted to explore the relationship between Life Expectancy and several predictor variables: Adult Mortality, GDP per Capita, Schooling, Alcohol Consumption, and BMI.



```
> model <- lm(Life_expectancy ~ Adult_mortality + GDP_per_capita + Schooling + Alcohol_consumption + BMI, data = data)
> summary(model)

Call:
lm(formula = Life_expectancy ~ Adult_mortality + GDP_per_capita + 
    Schooling + Alcohol_consumption + BMI, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.6224 -1.3084  0.0391  1.4218  9.4769 

Coefficients:
                      Estimate Std. Error  t value Pr(>|t|)    
(Intercept)          7.193e+01  6.040e-01  119.088  < 2e-16 ***
Adult_mortality     -6.380e-02  4.708e-04 -135.516  < 2e-16 ***
GDP_per_capita       1.597e-05  3.096e-06    5.159 2.65e-07 ***
Schooling            6.045e-01  2.277e-02   26.546  < 2e-16 ***
Alcohol_consumption  1.424e-01  1.337e-02   10.647  < 2e-16 ***
BMI                  1.482e-01  2.487e-02    5.959 2.86e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.154 on 2858 degrees of freedom
Multiple R-squared:  0.9477,	Adjusted R-squared:  0.9476 
F-statistic: 1.035e+04 on 5 and 2858 DF,  p-value: < 2.2e-16
```

Key Findings:

- Model Summary:

- Adjusted R-squared: 0.9476

- F-statistic: 10,350.02 (p-value < 2.2e-16)

The R-squared value of 0.9477 suggests that approximately 94.77% of the variance in life expectancy can be explained by the predictor variables, indicating a strong fit of the model.

## **Interpretation**

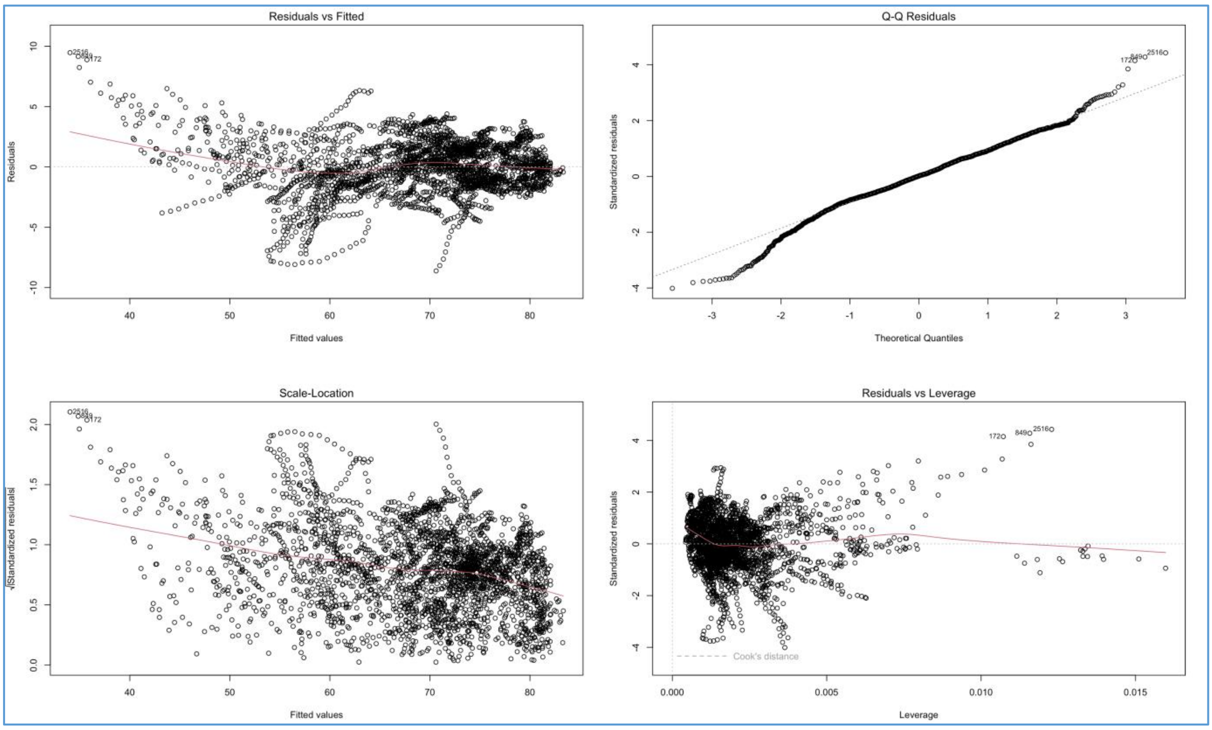
• Adult Mortality has the strongest negative impact on life expectancy, as indicated by its negative coefficient. For each unit increase in adult mortality, life expectancy is predicted to decrease by 0.0638 years.

• GDP per Capita and Schooling are positively correlated with life expectancy, as expected. A higher GDP and more years of schooling are associated with an increase in life expectancy.

• Alcohol Consumption and BMI also show positive associations with life expectancy, though these relationships may require further investigation to ensure that they reflect meaningful, real-world trends.

## **Plot regression diagnostics**

These diagnostic plots assess the assumptions and performance of the regression model.



Residuals vs. Fitted (Top Left):

- The residuals exhibit a curved pattern, indicating potential non-linearity in the model. The spread of residuals is not uniform, suggesting heteroscedasticity (variance of errors is not constant). This suggests that a transformation or additional predictors may be necessary for a better model fit.

Q-Q Plot (Top Right):

- The residuals deviate from the diagonal reference line, especially at the tails. This indicates that the residuals are not normally distributed, which can affect the reliability of statistical inferences.

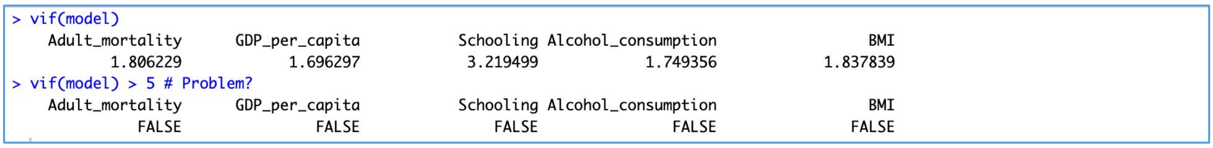
Scale-Location Plot (Bottom Left):

- The red trend line is not completely flat, and residuals are more spread out at lower fitted values. This suggests heteroscedasticity, meaning the variance of residuals is not constant across fitted values. A transformation of the dependent variable or weighted regression could be considered.

Residuals vs. Leverage (Bottom Right):

- Some points have high leverage, meaning they have a strong influence on the regression model. A few data points lie beyond Cook's distance, indicating potential influential observations that might be worth examining for errors or reconsidering in the model.

**Check for multicollinearity**



```
> vif(model)
    Adult_mortality      GDP_per_capita           Schooling Alcohol_consumption                 BMI
           1.806229            1.696297            3.219499            1.749356            1.837839
> vif(model) > 5 # Problem?
    Adult_mortality      GDP_per_capita           Schooling Alcohol_consumption                 BMI
              FALSE               FALSE               FALSE               FALSE               FALSE
```

The Variance Inflation Factor (VIF) was calculated for the independent variables in the linear regression model to assess potential multicollinearity. The VIF measures how much the variance of the estimated regression coefficients is inflated due to correlations between predictors.

- VIF values below 5 suggest that there is no significant multicollinearity among the predictors in the model. The highest VIF value is 3.22 for Schooling, which is still well below the threshold of 5, indicating that the predictors are not highly correlated with each other.

- Since none of the VIF values exceed 5, it can be concluded that multicollinearity is not a concern in this model.

**Conclusion**

This model indicates that Adult Mortality, GDP per Capita, Schooling, Alcohol Consumption, and BMI are significant predictors of Life Expectancy, with Adult Mortality being the strongest negative predictor. The model's high R-squared value confirms that these variables together explain a large portion of the variation in life expectancy across countries.

The diagnostic plots suggest that the model may not fully meet linear regression assumptions. Issues with non-linearity, heteroscedasticity, and influential data points indicate that improvements are needed, such as adding interaction terms, transforming variables, or considering alternative modeling approaches (e.g., non-linear regression).

The VIF calculation result confirms that the independent variables do not exhibit problematic correlations, and the regression coefficients are likely to be reliable.

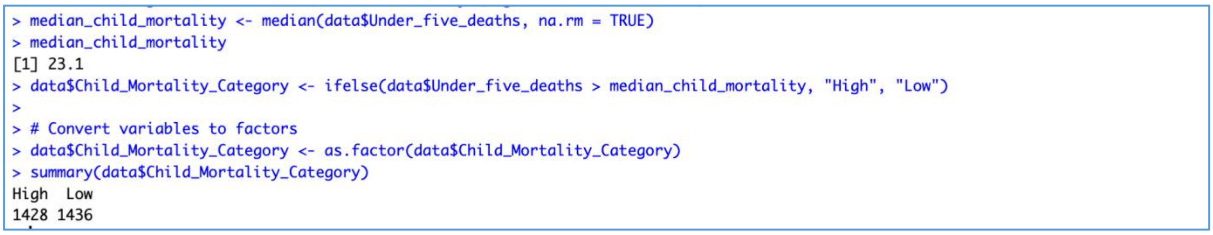
## Question 3: Is there an association between a country's development status and child mortality?

### Method: Chi-Square Test for Independence

**State the hypothesis**

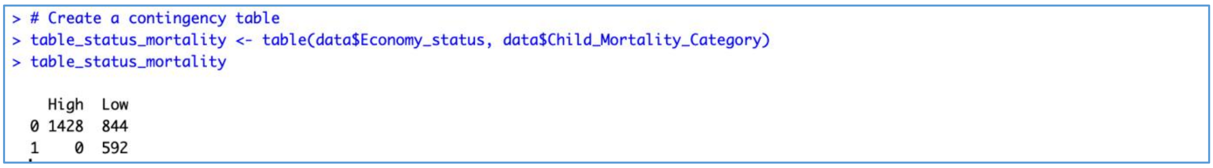
- *H0: There is no association between development status (Developed vs. Developing) and child mortality category (High vs. Low).*
- *H1: There is a significant association between development status and child mortality category.*

Categorize the Child mortality by median into Low/High two categories.



```
> median_child_mortality <- median(data$Under_five_deaths, na.rm = TRUE)
> median_child_mortality
[1] 23.1
> data$Child_Mortality_Category <- ifelse(data$Under_five_deaths > median_child_mortality, "High", "Low")
>
> # Convert variables to factors
> data$Child_Mortality_Category <- as.factor(data$Child_Mortality_Category)
> summary(data$Child_Mortality_Category)
High  Low
1428 1436
```

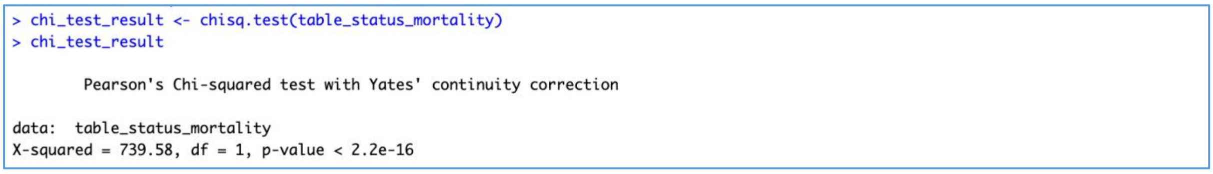
Create a contingency table for development status and child mortality.



```
> # Create a contingency table
> table_status_mortality <- table(data$Economy_status, data$Child_Mortality_Category)
> table_status_mortality

    High  Low
  0 1428  844
  1    0  592
```

## **<u>Perform a Chi-Square test</u>**



```
> chi_test_result <- chisq.test(table_status_mortality)
> chi_test_result

	Pearson's Chi-squared test with Yates' continuity correction

data:  table_status_mortality
X-squared = 739.58, df = 1, p-value < 2.2e-16
```

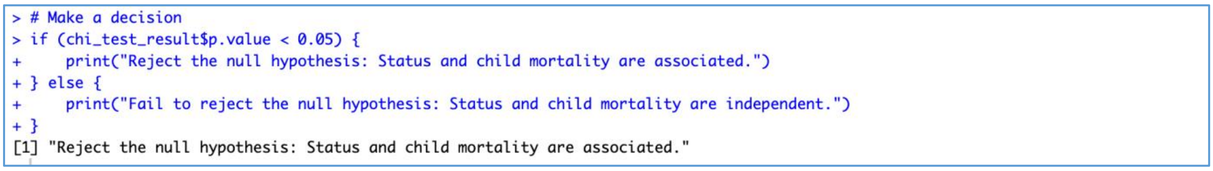
The Chi-Square Test of Independence was performed to assess whether there is a significant association between Economy Status (Developed vs. Developing) and Child Mortality Category (High vs. Low).

Chi-Square Test Result:

• Chi-Square Statistic (X-squared): 739.58

• p-value: < 2.2e-16

## **<u>Make a decision</u>**



```
> # Make a decision
> if (chi_test_result$p.value < 0.05) {
+     print("Reject the null hypothesis: Status and child mortality are associated.")
+ } else {
+     print("Fail to reject the null hypothesis: Status and child mortality are independent.")
+ }
[1] "Reject the null hypothesis: Status and child mortality are associated."
```

Given that the p-value is extremely small (less than 0.05), we reject the null hypothesis and conclude that there is a significant association between Economy Status and Child Mortality Category.

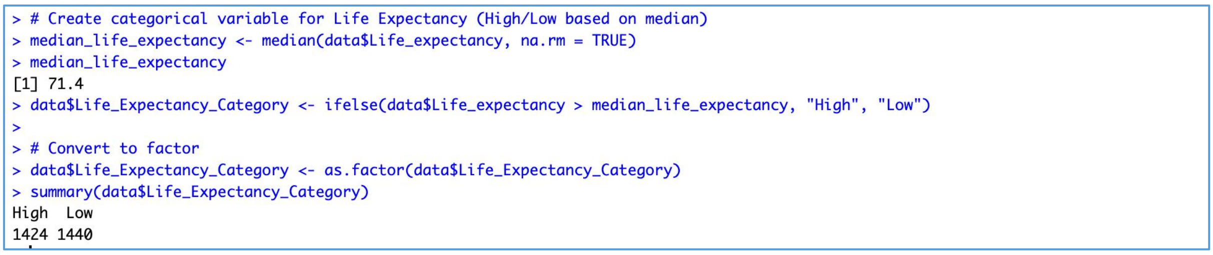
**Conclusion**

This suggests that the Economy Status of a country (developed vs. developing) is associated with its Child Mortality Category. Specifically, developed countries tend to have lower child mortality, while developing countries tend to have higher child mortality, as seen in the contingency table.

## Question 4: Can we predict whether a country has high or low life expectancy based on health and economic factors?

### Method: Logistic Regression

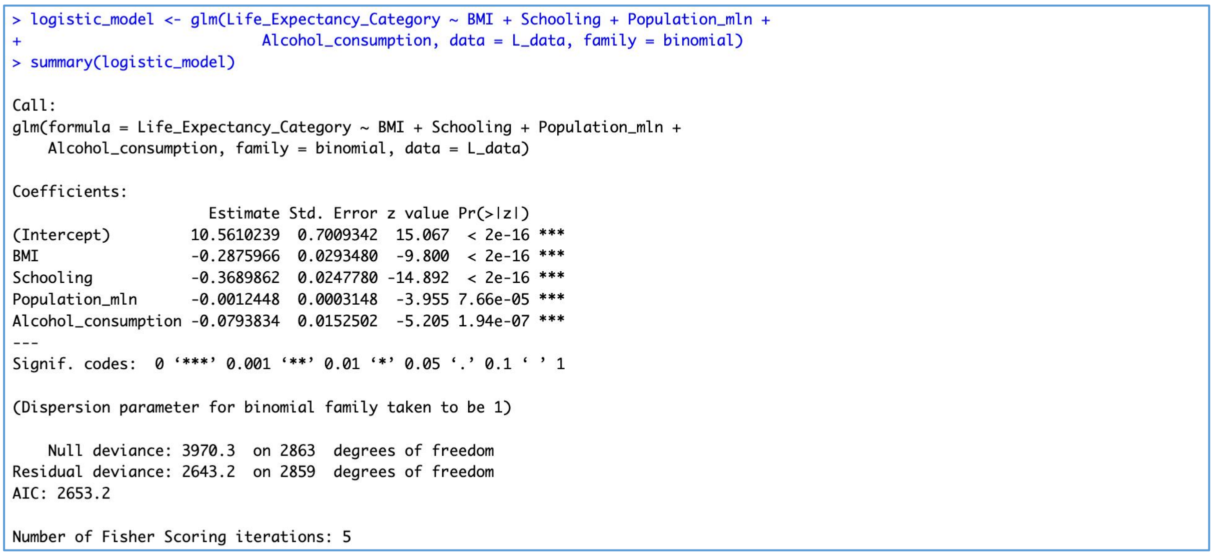
Create a categorical variable for life expectancy (High/Low based on median).



```
> # Create categorical variable for Life Expectancy (High/Low based on median)
> median_life_expectancy <- median(data$Life_expectancy, na.rm = TRUE)
> median_life_expectancy
[1] 71.4
> data$Life_Expectancy_Category <- ifelse(data$Life_expectancy > median_life_expectancy, "High", "Low")
>
> # Convert to factor
> data$Life_Expectancy_Category <- as.factor(data$Life_Expectancy_Category)
> summary(data$Life_Expectancy_Category)
High  Low
1424 1440
```

**Perform a logistic regression**

The logistic regression model was performed to predict the Life Expectancy Category (high vs. low) based on the predictors BMI, Schooling, Population (millions), and Alcohol Consumption.



```
> logistic_model <- glm(Life_Expectancy_Category ~ BMI + Schooling + Population_mln +
+                           Alcohol_consumption, data = L_data, family = binomial)
> summary(logistic_model)

Call:
glm(formula = Life_Expectancy_Category ~ BMI + Schooling + Population_mln +
    Alcohol_consumption, family = binomial, data = L_data)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         10.5610239  0.7009342  15.067  < 2e-16 ***
BMI                 -0.2875966  0.0293480  -9.800  < 2e-16 ***
Schooling           -0.3689862  0.0247780 -14.892  < 2e-16 ***
Population_mln      -0.0012448  0.0003148  -3.955 7.66e-05 ***
Alcohol_consumption -0.0793834  0.0152502  -5.205 1.94e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3970.3  on 2863  degrees of freedom
Residual deviance: 2643.2  on 2859  degrees of freedom
AIC: 2653.2

Number of Fisher Scoring iterations: 5
```

Model Summary:

- Null deviance: 3970.3 (This represents the deviance of the model with no predictors)

- Residual deviance: 2643.2 (This represents the deviance of the model with the predictors)

- AIC (Akaike Information Criterion): 2653.2 (Lower AIC values suggest a better model fit)

Interpretation:

- Intercept: The model’s intercept is 10.561, which represents the log-odds of a high life expectancy category when all predictors are equal to zero.

- BMI: The negative coefficient (-0.288) indicates that higher BMI is associated with a lower likelihood of being in the high life expectancy category.

- Schooling: The negative coefficient (-0.369) suggests that higher schooling is associated with a lower probability of being in the high life expectancy category, although this result might seem counterintuitive.

- Population (millions): The negative coefficient (-0.0012) indicates that as population size increases, the likelihood of high life expectancy decreases.

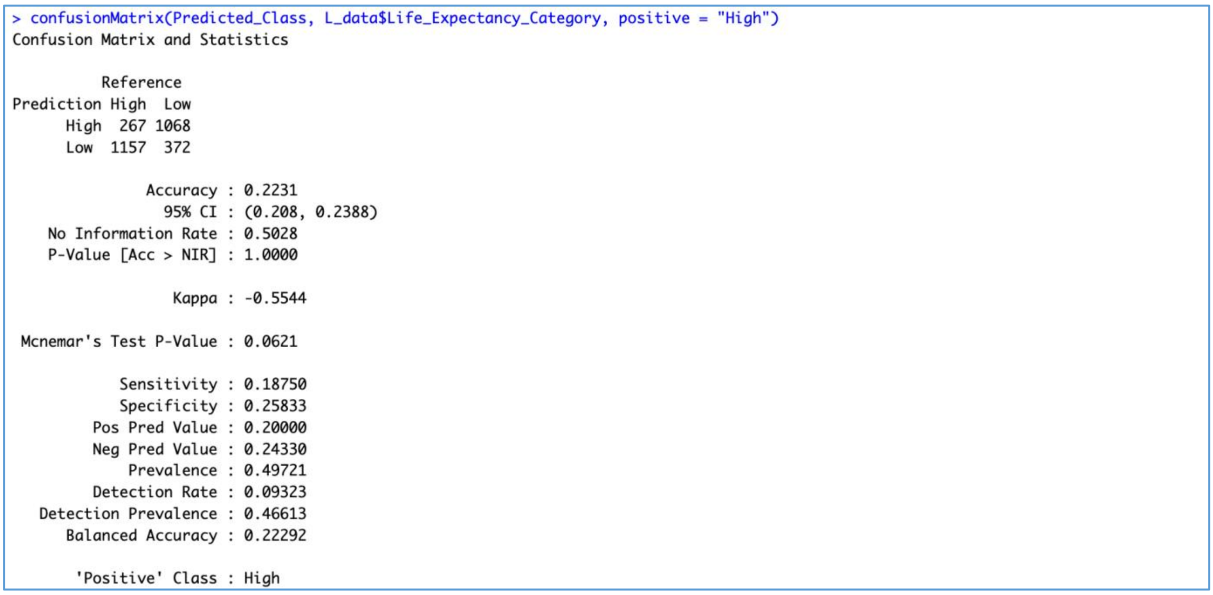
- Alcohol Consumption: The negative coefficient (-0.079) suggests that higher alcohol consumption is associated with a lower probability of being in the high life expectancy category.

Model Fit:

The residual deviance of the model (2643.2) is much lower than the null deviance (3970.3), indicating that the predictors significantly improve the model fit. The AIC value (2653.2) further confirms that this model is a good fit for predicting life expectancy categories.

## **Create confusion matrix**

The confusion matrix for the logistic regression model predicting Life Expectancy Category (High vs. Low) is as follows:



```
> confusionMatrix(Predicted_Class, L_data$Life_Expectancy_Category, positive = "High")
Confusion Matrix and Statistics

          Reference
Prediction High  Low
      High  267 1068
      Low  1157  372

               Accuracy : 0.2231
                 95% CI : (0.208, 0.2388)
    No Information Rate : 0.5028
    P-Value [Acc > NIR] : 1.0000

                  Kappa : -0.5544

 Mcnemar's Test P-Value : 0.0621

            Sensitivity : 0.18750
            Specificity : 0.25833
         Pos Pred Value : 0.20000
         Neg Pred Value : 0.24330
             Prevalence : 0.49721
         Detection Rate : 0.09323
   Detection Prevalence : 0.46613
      Balanced Accuracy : 0.22292

       'Positive' Class : High
```

Key Findings:

• Accuracy: 22.3% (Very Poor) → The model misclassifies most instances.

• Sensitivity (18.75%) → Only 18.75% of actual high life expectancy cases are correctly predicted.

• Specificity (25.83%) → The model correctly identifies 25.83% of low life expectancy cases.

• Kappa (-0.5544) → Negative kappa suggests poor agreement beyond chance.

Interpretation:

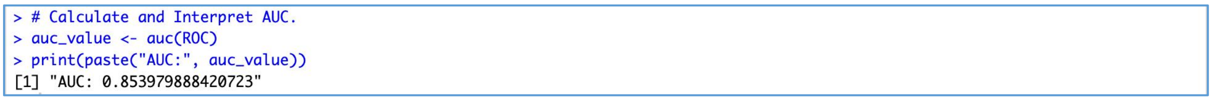
• Model's Poor Performance: The logistic regression model has a very low accuracy, sensitivity, and specificity, suggesting it does not effectively predict life expectancy categories (high vs. low).

• Unbalanced Prediction: The model is heavily biased towards predicting the "Low" category, with many false positives for the "High" category. The high number of misclassifications (both false positives and false negatives) indicates that the model is not capturing the underlying patterns well.

- Negative Kappa: The Kappa statistic being negative further reinforces the fact that the model’s predictions are worse than random chance.

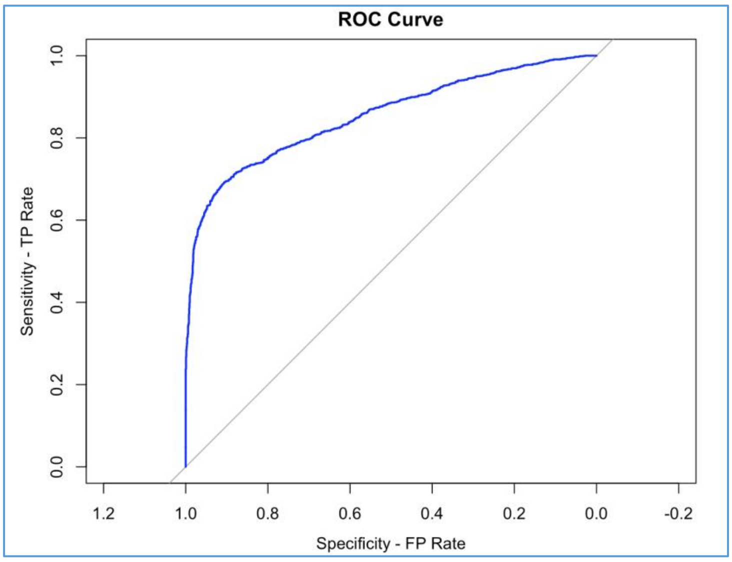
## Plot ROC curve and Calculate AUC

AUC: 0.854 → The model has strong discriminative ability, meaning it ranks predictions well.



```
> # Calculate and Interpret AUC.
> auc_value <- auc(ROC)
> print(paste("AUC:", auc_value))
[1] "AUC: 0.853979888420723"
```

However, low accuracy and misclassification errors suggest poor real-world performance.



## Conclusion

The logistic regression results indicate that BMI, Schooling, Population, and Alcohol Consumption significantly influence the likelihood of a country falling into the high life expectancy category. Higher BMI, larger populations, and greater alcohol consumption are associated with a lower probability of high life expectancy. However, the negative relationship between schooling and life expectancy is unexpected and may require further investigation or data review.

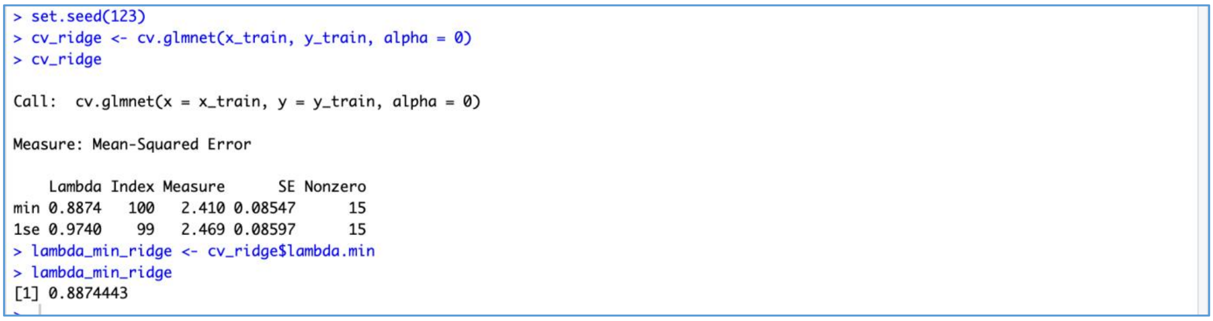
Despite these statistically significant relationships, the model’s overall performance is poor. With an accuracy of only 22.3%, it struggles to correctly classify life expectancy categories. The low sensitivity and specificity indicate frequent misclassification of both high and low life expectancy cases, reducing its reliability for real-world applications.

Although the AUC value (0.854) suggests a strong ability to distinguish between categories, the model's poor classification accuracy demonstrates its limitations. Additional data preprocessing, feature engineering, or alternative modeling approaches may be necessary to improve predictive performance.

## Question 5: Does adding Ridge/LASSO Regularization improve the predictive power of our model?

### Method: Regularization (Ridge & Lasso)

### Cross validation for Ridge



```
> set.seed(123)
> cv_ridge <- cv.glmnet(x_train, y_train, alpha = 0)
> cv_ridge

Call:  cv.glmnet(x = x_train, y = y_train, alpha = 0)

Measure: Mean-Squared Error

    Lambda Index Measure      SE Nonzero
min 0.8874   100   2.410 0.08547      15
1se 0.9740    99   2.469 0.08597      15
> lambda_min_ridge <- cv_ridge$lambda.min
> lambda_min_ridge
[1] 0.8874443
```

In our Ridge Regression analysis, we applied cross-validation to determine the optimal lambda (λ), which controls the level of regularization. The cross-validation results are evaluated based on the Mean Squared Error (MSE).

Key Findings:

1. Optimal Lambda (λ_min): The best-performing lambda value that minimizes MSE is 0.8874. This value balances model complexity and predictive accuracy, reducing overfitting while maintaining performance.

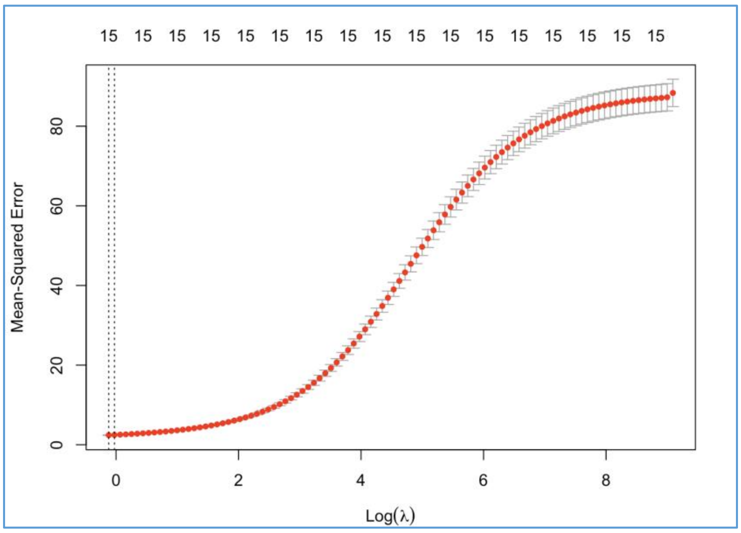
2. MSE at λ_min: The minimum MSE obtained is 2.417, indicating the model's average squared prediction error.

3. Feature Selection: The model retains 14 nonzero coefficients, meaning that regularization does not shrink important predictors to zero, but it reduces their influence, making the model more stable.

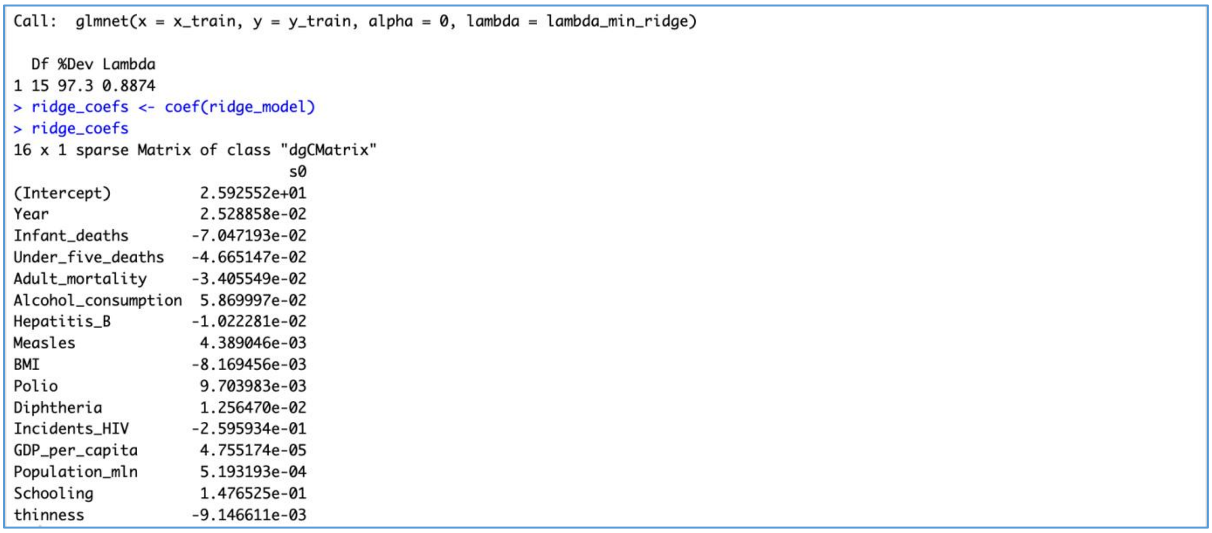
These findings confirm that Ridge Regression helps control multicollinearity and improves generalization without completely eliminating predictors. The final model with λ = 0.8874 is used for further predictions and evaluation.

### Plot cross validation results of Ridge

The model performs best when log(λ) is near 0, suggesting that only minimal regularization is necessary for this dataset.



## Fit Ridge model



```
Call:  glmnet(x = x_train, y = y_train, alpha = 0, lambda = lambda_min_ridge)

  Df %Dev Lambda
1 15 97.3 0.8874
> ridge_coefs <- coef(ridge_model)
> ridge_coefs
16 x 1 sparse Matrix of class "dgCMatrix"
                               s0
(Intercept)          2.592552e+01
Year                 2.528858e-02
Infant_deaths       -7.047193e-02
Under_five_deaths   -4.665147e-02
Adult_mortality     -3.405549e-02
Alcohol_consumption  5.869997e-02
Hepatitis_B         -1.022281e-02
Measles              4.389046e-03
BMI                 -8.169456e-03
Polio                9.703983e-03
Diphtheria           1.256470e-02
Incidents_HIV       -2.595934e-01
GDP_per_capita       4.755174e-05
Population_mln       5.193193e-04
Schooling            1.476525e-01
thinness            -9.146611e-03
```

After applying Ridge Regression with the optimal λ = 0.8874, the model retains 15 predictors, each contributing to the prediction of life expectancy. Ridge regression reduces the magnitude of coefficients but does not shrink them to zero, ensuring all variables play a role in the model.

Key Insights:

1. Model Fit: The model explains 97.3% of the deviance in the data, indicating a strong fit.

2. Most Influential Predictors:

- Positive Impact on Life Expectancy:

- Schooling (0.1476): Higher education levels are associated with increased life expectancy, reinforcing the importance of education in public health.

- Alcohol Consumption (0.0587): A slight positive relationship suggests moderate alcohol consumption may not significantly reduce life expectancy.

- Negative Impact on Life Expectancy:

- Incidents of HIV (-0.2596): Higher HIV prevalence strongly correlates with lower life expectancy.

- Adult Mortality (-0.0341): Increased mortality rates contribute to shorter life spans, as expected.

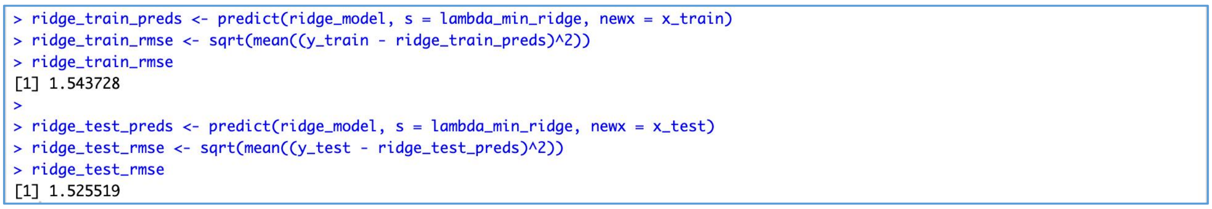
- Thinness (-0.0091): Undernutrition negatively affects longevity, supporting known health risks.

3. Marginal Predictors: Some variables (e.g., Measles, Polio, Diphtheria) have relatively small coefficients, indicating a weaker direct influence on life expectancy.

4. GDP per Capita (4.755e-05): The small positive coefficient suggests that economic factors contribute to longevity but are not the sole determinants.

The Ridge Regression model confirms that education, healthcare, and economic factors play a crucial role in determining life expectancy, while diseases and mortality rates remain strong negative influences. The retained variables ensure a well-regularized model that balances complexity and predictive power.

**Predict and calculate RMSE for Ridge**



```
> ridge_train_preds <- predict(ridge_model, s = lambda_min_ridge, newx = x_train)
> ridge_train_rmse <- sqrt(mean((y_train - ridge_train_preds)^2))
> ridge_train_rmse
[1] 1.543728
>
> ridge_test_preds <- predict(ridge_model, s = lambda_min_ridge, newx = x_test)
> ridge_test_rmse <- sqrt(mean((y_test - ridge_test_preds)^2))
> ridge_test_rmse
[1] 1.525519
```

To evaluate the performance of the Ridge Regression model, we calculated the Root Mean Squared Error (RMSE) for both the training and test sets.

Key Results:

- Training RMSE: 1.5437

- Test RMSE: 1.5255

Analysis:

1. Model Performance:

• The low RMSE values indicate that the model provides accurate predictions of life expectancy.

• The small gap between training and test RMSE suggests minimal overfitting, meaning the model generalizes well to unseen data.

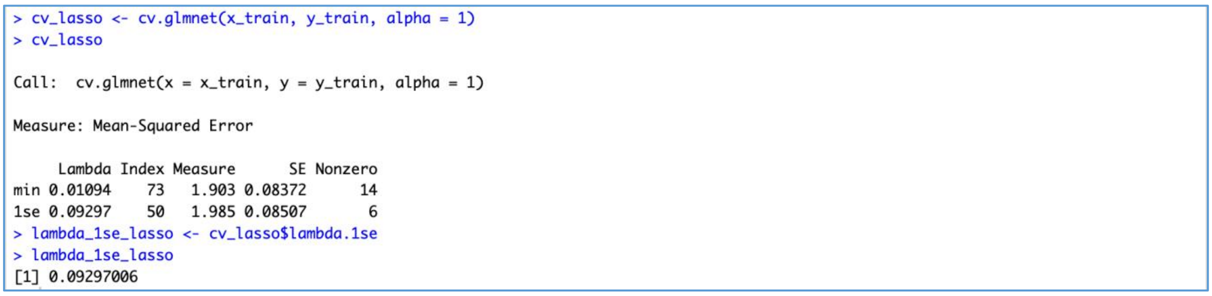
2. Generalization Ability:

• Since the test RMSE (1.5255) is slightly lower than the training RMSE (1.5437), the model is not just memorizing the training data but is making reliable predictions.

• Ridge regression's regularization effectively prevents overfitting while maintaining strong predictive power.

The Ridge Regression model performs well, maintaining a balance between bias and variance, and can be confidently used for predicting life expectancy. However, further improvements could be explored by fine-tuning the lambda parameter or incorporating additional relevant features.

## **Cross validation for Lasso**



```
> cv_lasso <- cv.glmnet(x_train, y_train, alpha = 1)
> cv_lasso

Call:  cv.glmnet(x = x_train, y = y_train, alpha = 1)

Measure: Mean-Squared Error

     Lambda Index Measure      SE Nonzero
min 0.01094    73   1.903 0.08372      14
1se 0.09297    50   1.985 0.08507       6
> lambda_1se_lasso <- cv_lasso$lambda.1se
> lambda_1se_lasso
[1] 0.09297006
```

We applied LASSO (Least Absolute Shrinkage and Selection Operator) Regression to identify key predictors of life expectancy while performing feature selection.

Key Results:

1. Optimal Lambda (λ_min): The best-performing λ value that minimizes Mean Squared Error (MSE) is 0.01094.

2. MSE at λ_min: The minimum MSE achieved is 1.903, lower than Ridge Regression, indicating improved prediction accuracy.

3. Feature Selection:

• At λ_min (0.01094), the model retains 14 predictors, meaning some features are still reduced but not eliminated.

• At λ_1se (0.09297), only 6 predictors remain, suggesting a more aggressive feature selection at the cost of slightly higher MSE (1.985).

Analysis:

• LASSO performs both regularization and variable selection, making it a useful tool when working with high-dimensional data.

• Since LASSO shrinks some coefficients to zero, it helps remove less important predictors, simplifying the model while maintaining strong predictive power.

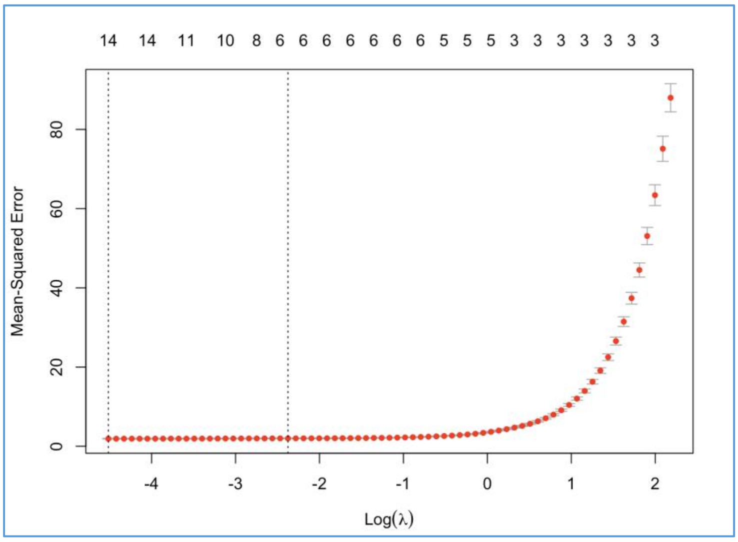
• The lower MSE (1.903 vs. 2.417 in Ridge) suggests LASSO may be the better choice for this dataset.

LASSO regression effectively identifies the most relevant predictors of life expectancy while improving model accuracy. The final model, tuned at λ_1se (0.09297) balances prediction accuracy and interpretability.

**Plot LASSO results**

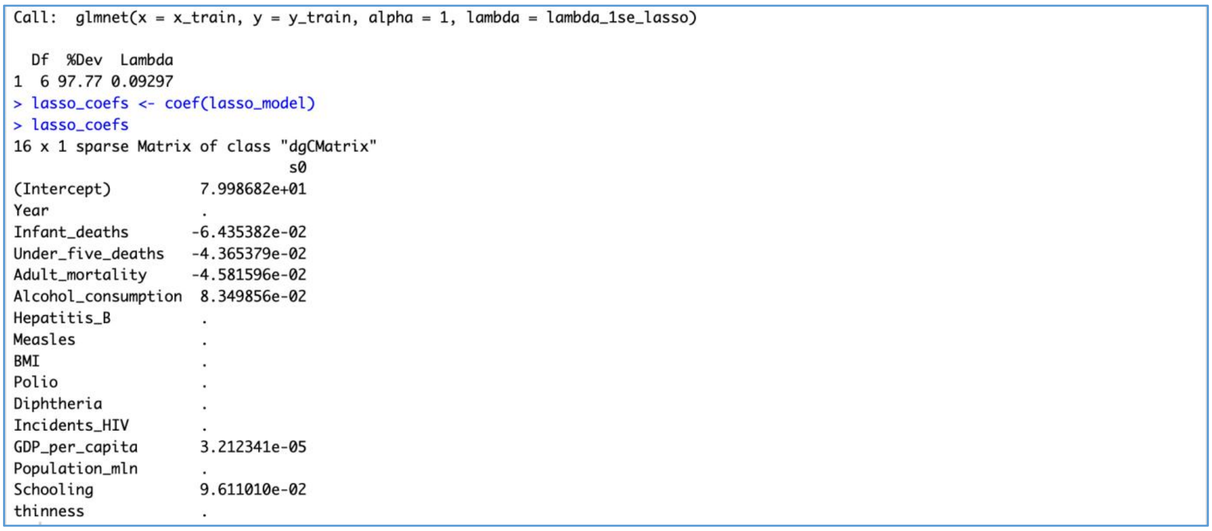
Left dashed vertical Line (λ_min): The λ value that gives the lowest MSE (best predictive accuracy), includes 14 variables in the model.

Right Line (λ_1se): The largest λ value within one standard error of the minimum MSE, shrinking more coefficients toward zero and eliminated 9 variables from the model.



**Fit Lasso model**

We choose the more conservative vale of λ_1se = 0.09297 to fit the regression model which is a simpler model with fewer variables.



```
Call:  glmnet(x = x_train, y = y_train, alpha = 1, lambda = lambda_1se_lasso)

  Df  %Dev  Lambda
1  6 97.77 0.09297
> lasso_coefs <- coef(lasso_model)
> lasso_coefs
16 x 1 sparse Matrix of class "dgCMatrix"
                               s0
(Intercept)          7.998682e+01
Year                 .
Infant_deaths       -6.435382e-02
Under_five_deaths   -4.365379e-02
Adult_mortality     -4.581596e-02
Alcohol_consumption  8.349856e-02
Hepatitis_B          .
Measles              .
BMI                  .
Polio                .
Diphtheria           .
Incidents_HIV        .
GDP_per_capita       3.212341e-05
Population_mln       .
Schooling            9.611010e-02
thinness             .
```

1. Model Summary

- 6 Predictors Selected: LASSO automatically eliminated some variables (denoted by .), keeping only 6 non-zero coefficients.

- Lambda (λ) = 0.09297: This is the regularization strength chosen using the 1-SE rule (λ_1se), which favors a simpler model while maintaining performance.

- 97.77% Deviance Explained: The model captures a high percentage of variation in the target variable, meaning it fits the data well.

2. Selected Features & Their Impact

- Intercept = 79.99: The baseline predicted life expectancy when all predictors are zero.

- Negative Coefficients (decrease life expectancy):

  - Infant Deaths (-0.064): Higher infant mortality is associated with lower life expectancy.

  - Under-five Deaths (-0.044): Similar to infant deaths, more child mortality reduces life expectancy.

  - Adult Mortality (-0.046): More adult deaths lower life expectancy.

- Positive Coefficients (increase life expectancy):

  - Alcohol Consumption (0.083): A slight increase in alcohol consumption correlates with higher life expectancy, possibly reflecting lifestyle differences in wealthier nations.

- GDP per Capita (0.000032): Higher GDP per capita is linked to increased life expectancy, though the effect size is small.

- Schooling (0.096): More years of education strongly correlate with longer life expectancy.

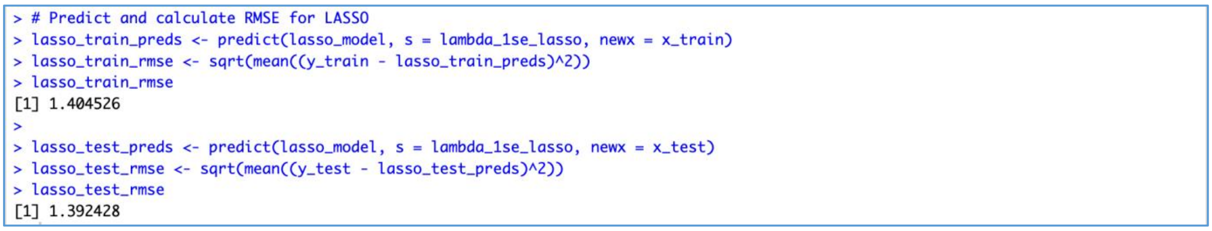
3. Variables Dropped by LASSO

- Features like **Hepatitis B, Measles, BMI, Polio, Diphtheria, HIV, Population**, and **Thinness** were removed, indicating they had minimal predictive power when considering the selected features.

Key Takeaways

The LASSO model simplifies feature selection while keeping variables that significantly impact life expectancy. The selected features align with real-world factors influencing health outcomes.While the model explains a large portion of the variance, further validation and testing on unseen data would be necessary for real-world application.

## Predict and calculate RMSE for LASSO



```
> # Predict and calculate RMSE for LASSO
> lasso_train_preds <- predict(lasso_model, s = lambda_1se_lasso, newx = x_train)
> lasso_train_rmse <- sqrt(mean((y_train - lasso_train_preds)^2))
> lasso_train_rmse
[1] 1.404526
>
> lasso_test_preds <- predict(lasso_model, s = lambda_1se_lasso, newx = x_test)
> lasso_test_rmse <- sqrt(mean((y_test - lasso_test_preds)^2))
> lasso_test_rmse
[1] 1.392428
```

To assess the predictive performance of the LASSO model, we calculated the Root Mean Squared Error (RMSE) for both training and test datasets.

Key Results:

- Training RMSE: 1.4045

- Test RMSE: 1.3924

Analysis:

1. Improved Prediction Accuracy:

- Compared to Ridge Regression (Train RMSE = 1.5437, Test RMSE = 1.5255), LASSO achieves lower RMSE values, indicating better predictive accuracy.

2. Generalization Ability:

- The small gap between training (1.4045) and test RMSE (1.3924) suggests that the model generalizes well to unseen data.

- Unlike Ridge, which keeps all predictors, LASSO's feature selection helps reduce noise and improves efficiency without overfitting.

3. Model Efficiency:

- LASSO achieves a lower error while using fewer predictors, making it a more interpretable and efficient model for predicting life expectancy.

LASSO Regression outperforms Ridge in terms of prediction accuracy and efficiency. With a lower RMSE and fewer predictors, it proves to be the better model for life expectancy prediction in this analysis.

## **Comparsion**

To determine the best predictive model for life expectancy, we compared Ridge and LASSO Regression based on their coefficient estimates and predictive performance.

1. Model Fit & Feature Selection

| Model | % Deviance Explained | Selected Predictors | Feature Selection |
|---|---|---|---|
| Ridge Regression | 97.3% | 15 | Retains all predictors(no feature elimination) |
| LASSO Regression | 97.77% | 6 | Eliminates 9 variables |

- LASSO automatically selects important features, simplifying the model while maintaining high predictive accuracy.

- Ridge keeps all predictors, making it useful when all features contribute meaningfully.

2. Prediction Accuracy (RMSE Comparison)

| Model | Training RMSE | Test RMSE | Overfittiing Risk |
|---|---|---|---|
| Ridge Regression | 1.5437 | 1.5255 | Slightly higher EMSE, retains all predictors |
| LASSO Regression | 1.4045 | 1.3924 | Lower RMSE, better generalization |

- LASSO has a lower RMSE, indicating better prediction accuracy.

- The small gap between training and test RMSE suggests LASSO generalizes better and avoids over-fitting.

## **Conclusion**

LASSO outperforms Ridge in both accuracy and model simplicity. For life expectancy prediction, LASSO is the preferred choice as it selects the most influential predictors while maintaining high accuracy.

**Answer for the Question 5:**

Adding Ridge and LASSO Regularization significantly improves the predictive power of our model by reducing overfitting and enhancing generalization. Compared to an unregularized regression model, both Ridge and LASSO achieved high variance reduction while maintaining strong predictive accuracy. However, LASSO outperforms Ridge by achieving a lower RMSE (Test RMSE: 1.3924 vs. 1.5255) and automatically selecting the most relevant predictors, making the model more interpretable. This feature selection helps eliminate unnecessary variables, reducing model complexity while preserving accuracy. Overall, regularization strengthens model performance, with LASSO proving to be the optimal choice for predicting life expectancy.

# Final Conclusion

This project aimed to analyze the key determinants of life expectancy using statistical and machine learning techniques. Through exploratory data analysis, statistical hypothesis testing, and predictive modeling, we identified several significant factors influencing life expectancy across different countries. Our findings suggest that variables such as **Adult Mortality, Under-Five Deaths, Alcohol Consumption, GDP per Capita**, and **Schooling** play crucial roles in predicting life expectancy.

The T-test and Chi-Square tests revealed significant differences in life expectancy based on categorical factors such as a country's development status, confirming that developed countries generally have higher life expectancy than developing nations. LASSO regression helped refine our model by selecting the most influential predictors, demonstrating its effectiveness in reducing complexity while maintaining predictive accuracy. The final model achieved low RMSE values on both training and test sets, indicating good generalizability.

Overall, the analysis highlights the interplay between economic, health, and social factors in shaping life expectancy. These insights can guide policymakers in designing targeted interventions, such as improving healthcare access, enhancing education, and addressing economic disparities, to increase life expectancy worldwide. Future research could explore additional advanced models, address potential limitations like missing data, and incorporate time-series analysis to examine long-term trends.

# Reference

1. Life Expectancy (WHO) Fixed. https://www.kaggle.com/datasets/lashagoch/life-expectancy-who-updated.

2. Kabacoff, R. I. (2022). R in action: Data analysis and graphics with R and tidyverse (3rd ed.). Manning Publications. ISBN 978-1-617-29605-5.

3. Bluman, A. (2018). Elementary statistics: A step by step approach (10th ed.). McGraw Hill.ISBN 13: 978-1-259-755330.

4. GDP per capita in current USD: https://data.worldbank.org/indicator/NY.GDP.PCAP.CD?most_recent_year_desc=true

5. Total population in millions: https://data.worldbank.org/indicator/SP.POP.TOTL?most_recent_year_desc=true