Final Modeling Report

Chae Yeon Lee
Yashish Mohnot

Introduction

This modeling report aims to answer the research question, "What are the different factors that determine life expectancy?" In order to answer this question, we will be conducting observational and explanatory modeling to create a predictive model of factors that affect life expectancy. Our dataset contains information from 193 different countries, ranging from the year 2000 to 2015, focusing on factors that potentially affect average life expectancy.

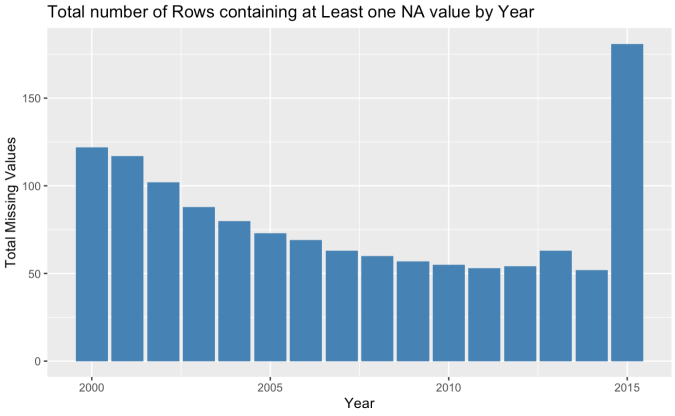
We will treat this research question from a purely predictive standpoint. We will try to see which models do a good job at predicting life expectancy. It is important to note that our dataset contains both categorical and continuous data. We will examine our data using various modeling techniques, like ANOVA, model selection, model diagnostics and time series regression.

One of the biggest challenges we faced during our analysis was dealing with the time series nature of the data. Since our data is collected over multiple years, this violates the assumption of independent observations, making it difficult to simply pool all of the data together. To address this challenge, we use time series methods to model the data. Overall, the findings from this modeling report could provide important real-world insights that can help identify and improve the factors that influence life expectancy, ultimately helping people live longer and healthier lives.

Data Pre-processing

Dealing with NA values.

After identifying 2400 missing values in our dataset, we explored different methods for handling them: removing rows with missing values, replacing missing values with 0, and approximating missing values with the mean. We evaluated each method's performance using 40-fold cross-validation on a linear model and found that removing rows with missing values had the lowest root mean squared error (RMSE) of 14.42, followed by approximating missing values (RMSE=17.81), and replacing missing values with 0 (RMSE=31.2). Therefore, we used the method of removing rows with missing values for the rest of the project, as it still left us with a substantial amount of data (1649 out of 2938 rows) and improved model performance.

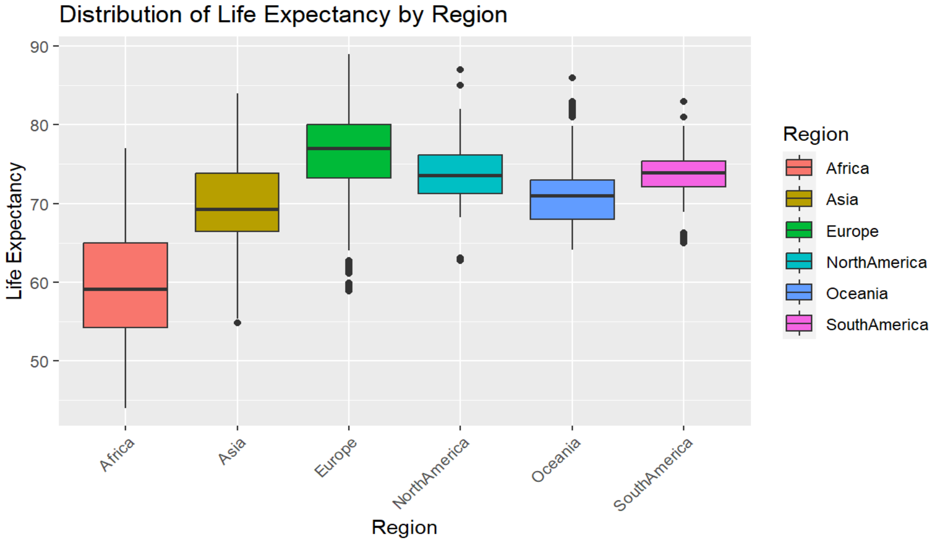


Additionally, we found that the missing data was evenly distributed over most time periods as seen above, and excluding the missing rows would not significantly impact the generalizability of our results as seen in the plot above. We decided to exclude the data from years 2001, 2002 and 2015 from our analysis due to abnormal counts of NA values.

Dealing with “Country” categorical variable

In our project, we faced a crucial decision regarding variable selection for our model. With 193 different countries in our dataset, we recognized that including each country as a categorical variable could potentially lead to overfitting and decreased model interpretability. Therefore, we made the decision to replace the country covariate with a "continent" covariate instead. This allowed us to group countries by their continent and reduce the number of categorical variables in our model.

To further justify our decision, we created a box plot that plotted the life expectancy for each continent. The box plot clearly showed a meaningful trend with the continent categories. Some continents had higher life expectancies on average, while others had lower life expectancies on average as seen below. This supported our decision to include continent as a covariate in our model since it was an essential factor in explaining the variation in life expectancy.



Modeling Tool Motivations and Assumptions

Model Diagnostics and Influential Observations

Motivation**:** In the beginning of our analysis, we performed model diagnostics using a full linear model that included all covariates. Our primary objective was to identify and remove any influential outliers from the dataset. Additionally, we assessed the validity of our modeling assumptions, including linearity, homoscedasticity, and normality of errors, to justify the use of linear Gaussian models.

Assumptions**:** Model diagnostics is a tool to validate assumptions, thus doesn't have any inbuilt assumptions itself.

ANOVA

Motivation**:** ANOVA methods are useful when analyzing the relationship between categorical variables and the response variable. Plotting the distribution of life expectancies for different regions, we noticed that there is a meaningful difference in life expectancy for regions. Furthermore, our EDA suggested that life expectancy has a positive correlation with GDP, a percentage expenditure on health as percentage of GDP, and schooling and a negative correlation with population. Since GDP, expenditure, schooling, and population have strong relationship with a country's status of development, we deemed that it is necessary to study the relationship between status of development and life expectancy. Hence, in this modeling report, we implemented an ANOVA model to investigate whether there are differences in life expectancy between different regions and development status.

Assumptions**:** Independence, normality, and homogeneity of variances. As mentioned in the introduction, since our data is collected over multiple years, this makes the observations not independent of each other. In order to address this issue, we selected one year, 2009, to perform ANOVA. Furthermore, selecting one year allows us to remove variances in life expectancy sourcing from years rather than from the features of interest – region and status. Hence, we have selected the year 2009 because it is a year with many data points of 126 rows. We validate the Normality assumption through model diagnostics. Note that after finishing our analysis for the year 2009, we also ran all our analysis tools for all years between 2003-2014 and achieved similar results.

Model Selection - Forward/Backward Selection using AIC/BIC

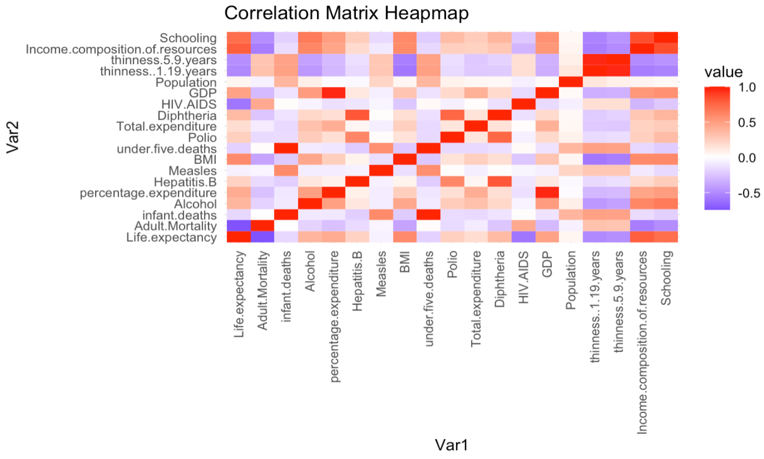
Motivation**:** Model selection is a critical step in any data analysis project because it ensures that the model accurately represents the relationships between the variables in the data. A well-chosen model can help in making accurate predictions and can lead to a better understanding of the underlying processes. On the other hand, an ill-chosen model can lead to poor predictions, biased parameter estimates, and unreliable conclusions. Considering the huge number of covariates in our full model, we need to perform model selection to achieve a good model. In our analysis, we performed model selection using the AIC and BIC criterion to choose the best subset of covariates that could explain the variation in our response variable while avoiding overfitting. By selecting the most important covariates, we were able to reduce the complexity of our model, improve its predictive power and increase its interpretability.

Assumptions: The forward/backward selection method with AIC/BIC assumes that the relationship between the response variable and covariates is linear and additive, and that the errors are normally distributed with constant variance (homoscedasticity). We validate all these assumptions in the model diagnostic phase of our analysis.

Shrinkage Methods - Lasso, Ridge

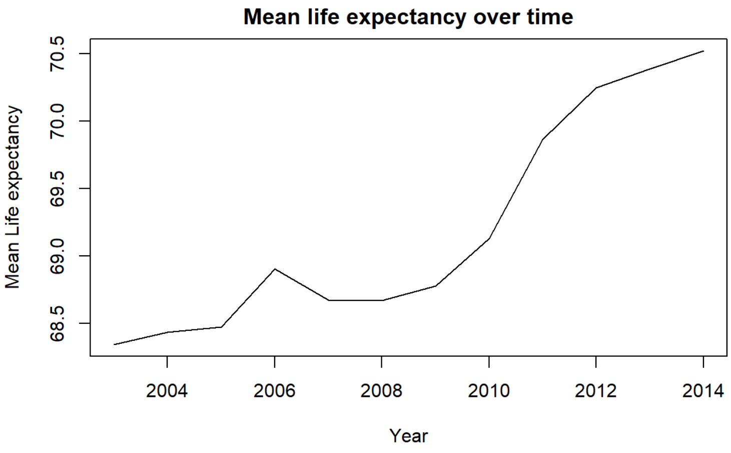
Motivation**:** In our analysis, we employed a combination of model selection and shrinkage techniques to identify the optimal model for our dataset. Along with selection through AIC and BIC, we also ran Lasso and Ridge Regression to select the most predictive covariates. To determine the most effective approach, we conducted cross-validation (on a held out dataset) and compared the results of all the model selection methods. By doing so, we aimed to identify the most suitable model that yields the best performance for our dataset.

Assumptions**:** Shrinkage methods assume linearity between the dependent variable and the independent variables, independence in observations of data, homoscedasticity, and normality. These conditions are validated in the model diagnostic of our analysis. Shrinkage methods, in addition to model selection techniques, aim to mitigate the issue of multicollinearity between covariates. We assessed the presence of multicollinearity by examining both the correlation matrix. According to the Correlation Matrix Heatmap, there is a strong correlation between multiple features. For example, the pair of variables with high correlation are (GDP and percentage expenditure), (Hepatitis B and Diphtheria), and (Schooling and Income composition of resources). Hence, we can observe that there exists multicollinearity and we anticipate that Ridge and Lasso would resolve such issues.



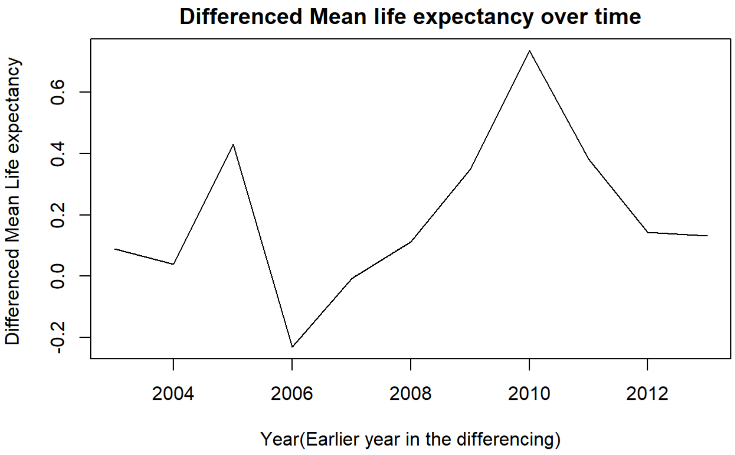
Time Series Regression

Motivation: On analyzing the data, we notice that the average life expectancy (averaged over all countries), had a linear relation with time as seen below.



We realized that this linear trend violated one of the critical assumptions of linear modeling, namely the independence of observations (y_i). Therefore, to address this issue for all the above techniques, we restricted our analysis to data spanning a single year. Nonetheless, since we needed to incorporate the time variable into our regression model, we decided to employ time-series regression. This approach allowed us to capture the temporal dependence of the life expectancy variable in our data.

Assumptions**:** The assumptions held for time-series regression is that the data must exhibit stationarity, homoscedasticity, and the error terms should be normally distributed. We verify homoscedasticity, and normality of errors through model diagnostics. Finally, we can achieve stationarity by applying differencing techniques that remove the linear trend, resulting in a stationary series as seen below.

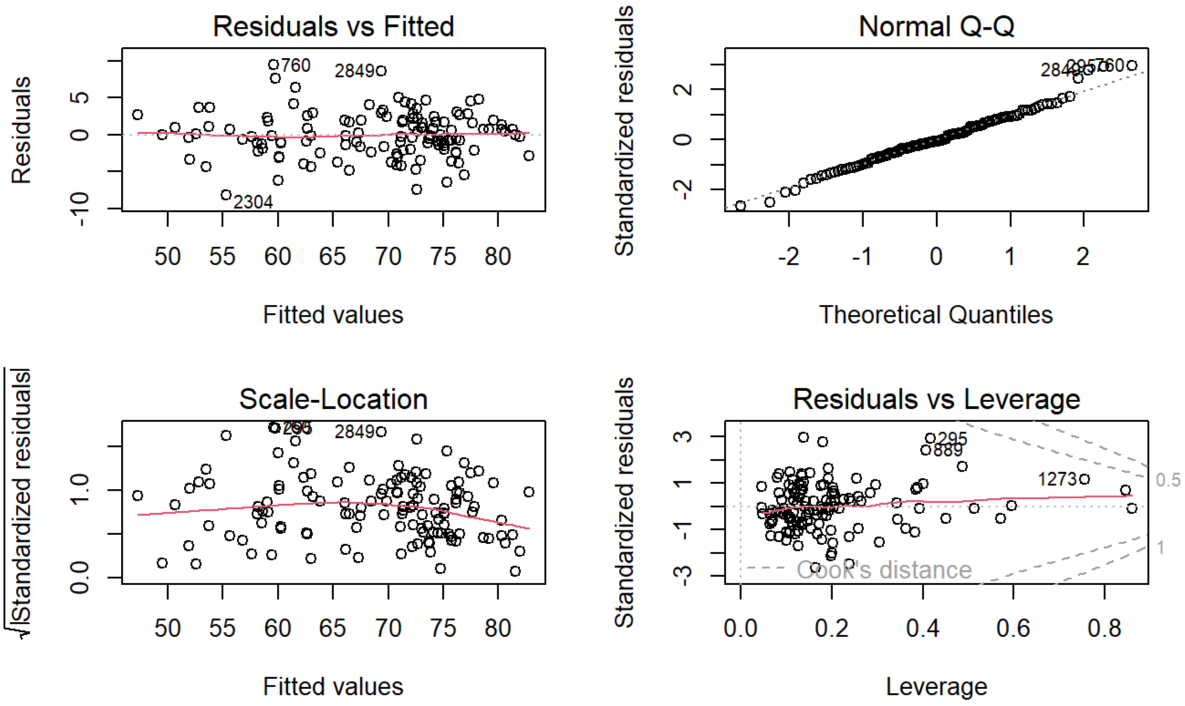


By satisfying these assumptions, we can build a robust time series regression model that accurately captures the underlying relationship between life expectancy over time.

Presentation and interpretation of results

Model Diagnostics and Influential observations:

We perform model diagnostics over a full linear model considering just one year (year = 2009) because all of our models are selected based on the data corresponding to 2009. We plotted models for all years and got similar results. On plotting the respective graphs, we get the following results:



- Residuals vs Fitted Values plot: Ideally, the residuals should be randomly scattered around zero, with no visible pattern. As we can see there is no strong visible pattern, which suggests linearity of data. Also, the residuals roughly form a "horizontal band" around the 0 line, suggesting homoscedasticity.
- Scale Location: Ideally, the points should be randomly scattered around a horizontal line, with no clear pattern. Since the points are randomly scattered around the line y = 0.9 with no clear pattern, this indicates homoscedasticity.
- Normal Q-Q: The points should fall along a straight line to indicate that the residuals are normally distributed. Although the points slightly deviate from the straight line, indicating a large tailed distribution with more data at the extremes, the deviation is not significant, and none of these points are influential (next plot). Hence, this concentration of data at extremes will not have a significant effect.
- Residuals vs Leverage: Ideally, the points should be randomly scattered with no points having high leverage and high standardized residuals. As we can see there are no points with Cook's distance > 0.5, indicating that there are no influential outliers present.

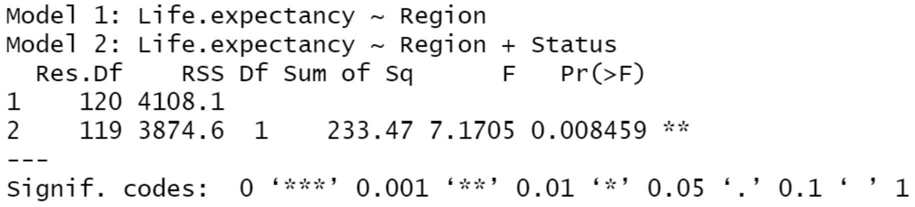
  In conclusion, our model diagnostic results indicate that the data met all the modeling assumptions we stated in the previous section over a span of any year. Therefore, if we select only one year of data, all the assumptions of the models are met.

  These diagnostic results were crucial to our analysis, as they validated all the assumptions we had to make for future modeling techniques. Specifically, these results provide evidence that the modeling tools we use are appropriate for our research question and will provide valid results.

ANOVA

When conducting ANOVA, our goals are twofold: to determine the effect of categorical variables on life expectancy and to determine if an interaction term between the variables should be included. In this section, we only consider the categorical variables without accounting for other covariates, as we explore the relationship between these variables and covariates in the model selection and shrinkage section.

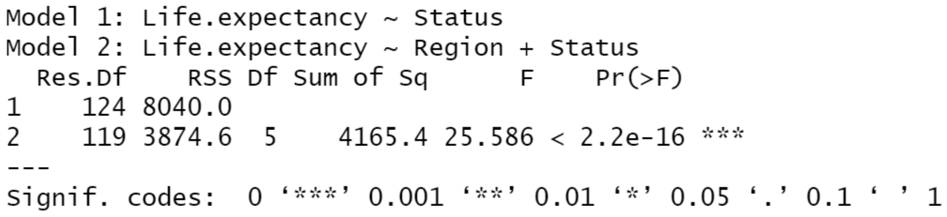
In our initial ANOVA test, we modeled life expectancy as a function of region. The null hypothesis (H_0) in this model is that there is no difference in mean life expectancy between regions, while the alternative hypothesis (H_a) states that the mean life expectancy of at least one region differs from the mean of the others.



```
Model 1: Life.expectancy ~ Region
Model 2: Life.expectancy ~ Region + Status
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1    120 4108.1
2    119 3874.6  1    233.47 7.1705 0.008459 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

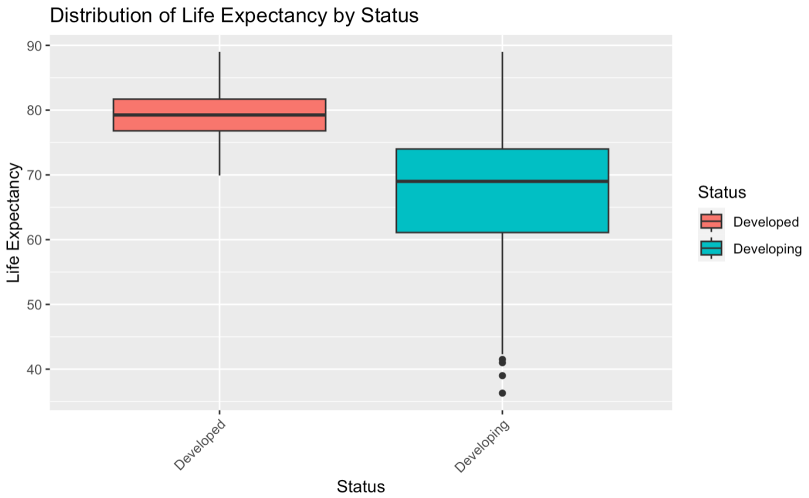
The region variable has a low p-value (0.008459), indicating that the region has a significant impact on life expectancy. This finding is consistent with the earlier graph, which displayed a strong correlation between life expectancy and region.

Additionally, we conducted a one-way ANOVA to model life expectancy as a function of a country's development status. In this case, the null hypothesis (H_0) states that there is no difference in mean life expectancy across different development statuses, while the alternative hypothesis (H_a) suggests that there is a significant difference in the mean life expectancy between various development statuses.

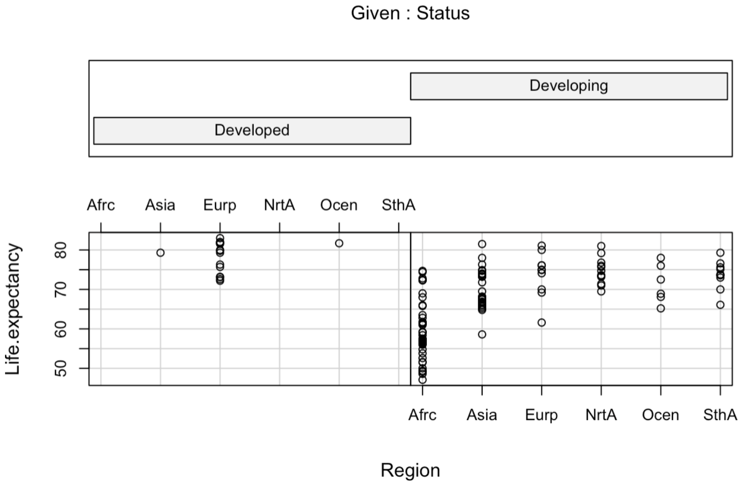


```
Model 1: Life.expectancy ~ Status
Model 2: Life.expectancy ~ Region + Status
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    124 8040.0
2    119 3874.6  5    4165.4 25.586 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The p-value of the Status variable is below the 0.05 cutoff, signifying that the development status has a significant impact on life expectancy. This result is consistent with the plot below, which indicates a substantial disparity in the life expectancy of developed countries compared to developing countries.



Moving forward, we examined the interaction term between status and region to determine if there was a significant association between them. To do this, we created a coplot that showed the relationship between life expectancy and region given status. We also listed all the developing countries from 2009 below.



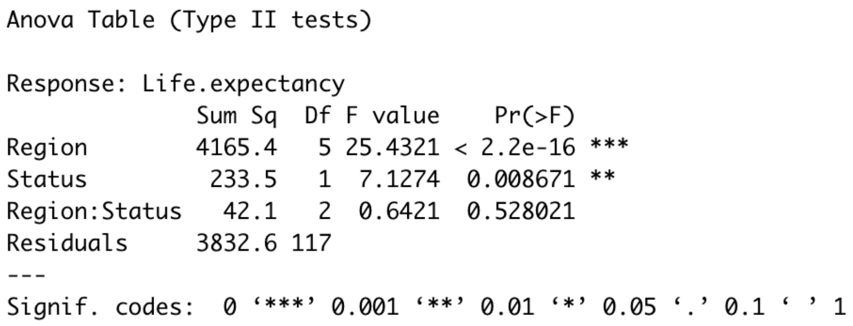
```
   Country Year    Status
 Australia 2009 Developed
   Austria 2009 Developed
   Belgium 2009 Developed
  Bulgaria 2009 Developed
   Croatia 2009 Developed
    Cyprus 2009 Developed
   Germany 2009 Developed
     Italy 2009 Developed
    Latvia 2009 Developed
 Lithuania 2009 Developed
Luxembourg 2009 Developed
     Malta 2009 Developed
    Poland 2009 Developed
  Portugal 2009 Developed
   Romania 2009 Developed
     Spain 2009 Developed
```

Surprisingly, we observed that in 2009, there were significantly fewer developed countries than developing countries, and notably, there were no developed countries in Africa, North America, and South America. Additionally, while all developed countries had a life expectancy of above 70, developing countries had a much wider range of life expectancy. We also noticed a constant shift in the means between the developed and developing categories, regardless of the region. This suggests that an interaction term might not be necessary. Nonetheless, we ran an ANOVA test to statistically explore the inclusion of an interaction term.

To follow the principle of marginality, we only include an interaction term if the main effects are significant in the model. As we have observed the main effects of both status and region on life expectancy, we have decided to create a model with interaction terms. Therefore, we have constructed four ANOVA models in this report:

Model	**Terms**
Full	Region, Status, Region x Status
No Interaction	Region, Status
Region	Region
Status	Status

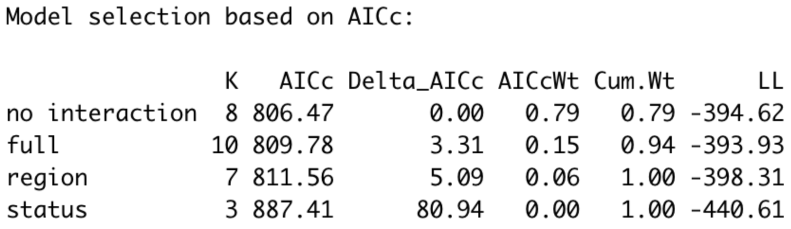
We used the ANOVA command to make one large ANOVA table summarizing the incremental F-tests for each individual variable.



```
Anova Table (Type II tests)

Response: Life.expectancy
              Sum Sq  Df F value    Pr(>F)    
Region        4165.4   5 25.4321 < 2.2e-16 ***
Status         233.5   1  7.1274  0.008671 ** 
Region:Status   42.1   2  0.6421  0.528021    
Residuals     3832.6 117                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

This shows us that each region and status variable are meaningful for predicting life expectancy but the interaction term between region and status is not statistically significant since its p-value is greater than 0.05.

In order to evaluate the explained variation, model complexity, and parsimony of each model, we calculated their AIC values. We have done this for all four ANOVA models constructed in this report.



```
Model selection based on AICc:

                K   AICc Delta_AICc AICcWt Cum.Wt      LL
no interaction  8 806.47       0.00   0.79   0.79 -394.62
full           10 809.78       3.31   0.15   0.94 -393.93
region          7 811.56       5.09   0.06   1.00 -398.31
status          3 887.41      80.94   0.00   1.00 -440.61
```

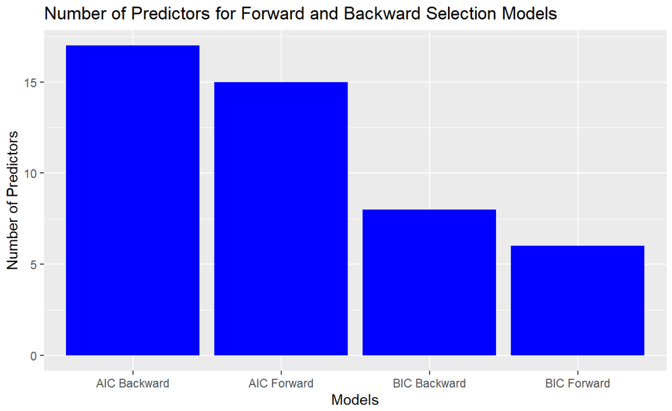
The model with no interaction term has the smallest AIC value of 806.47 compared to all other models. This result aligns with our earlier observations that the interaction term between region and status does not help with predicting life expectancy but both features – region and status – have meaningful relationships with the dependent variable.

This section is a critical component of our modeling approach, as it allows us to make two important conclusions.

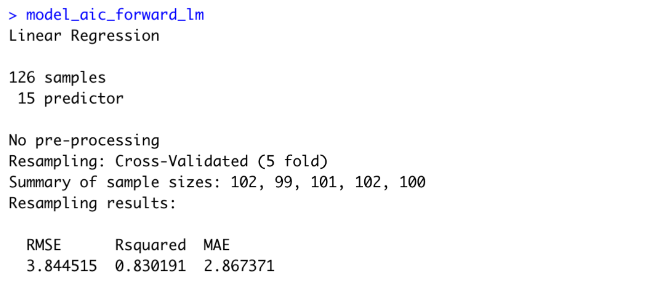
- We can confidently state that both the region and status categorical variables have a significant impact on life expectancy, when other covariates are not taken into account.
- Our analysis indicates that including an interaction term between these variables is not statistically significant, suggesting that it is not necessary to incorporate it as a covariate in our model.

Model Selection

Similar to ANOVA model, we have used data from 2009 to select features using forward and backward selection. We used four algorithms – Forward selection using AIC, Forward selection using BIC, Backward selection using AIC, and Backward selection using BIC – to create four models. To evaluate the performance of each model, 5-fold cross-validation was performed for each of the four models. The result is summarized in the table below. It shows that the RMSE of the AIC backward model, which selected 17 predictors, is the smallest by 3.630893 compared to all other three models when cross-validated on the data corresponding to year 2009.

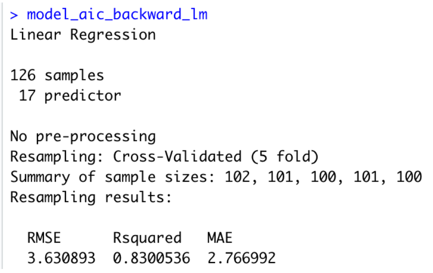


AIC Forward



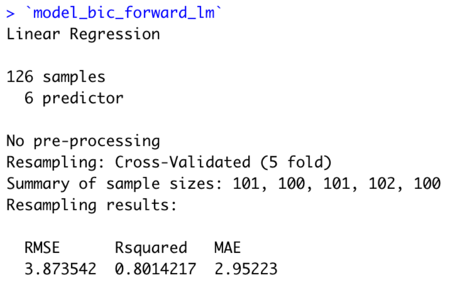
```
> model_aic_forward_lm
Linear Regression

126 samples
 15 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 102, 99, 101, 102, 100
Resampling results:

  RMSE      Rsquared  MAE
  3.844515  0.830191  2.867371
```

AIC Backward



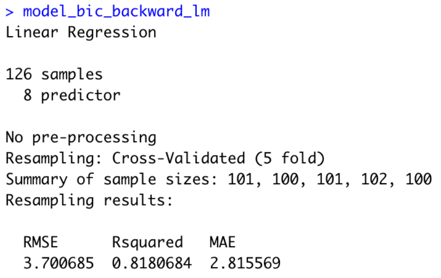
```
> model_aic_backward_lm
Linear Regression

126 samples
 17 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 102, 101, 100, 101, 100
Resampling results:

  RMSE      Rsquared   MAE
  3.630893  0.8300536  2.766992
```

BIC Forward



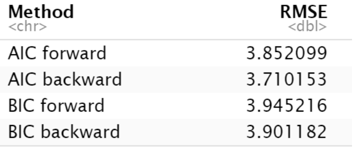
```
> `model_bic_forward_lm`
Linear Regression

126 samples
  6 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 101, 100, 101, 102, 100
Resampling results:

  RMSE      Rsquared   MAE
  3.873542  0.8014217  2.95223
```

BIC Backward



```
> model_bic_backward_lm
Linear Regression

126 samples
  8 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 101, 100, 101, 102, 100
Resampling results:

  RMSE      Rsquared   MAE
  3.700685  0.8180684  2.815569
```

Afterwards, in order to further validate that we were not overfitting to a single year, we used data from 2010 and cross validation with k = 5 folds for each model on the validation data, as seen on the right.

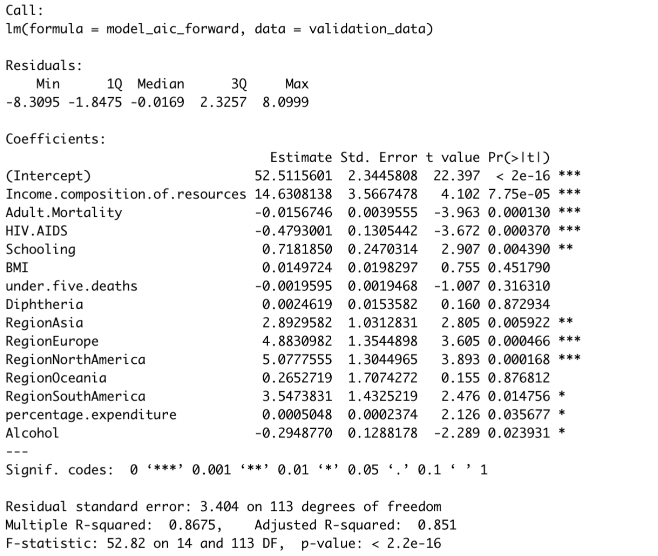


Method <chr>	RMSE <dbl>
AIC forward	3.852099
AIC backward	3.710153
BIC forward	3.945216
BIC backward	3.901182

We can observe that the RMSE of AIC backward is the smallest at 3.71 when we apply the model to 2010 data, which aligns with the result from 2009 data. Also, the BIC backward RMSE of the BIC backward is the largest at 3.9, which also aligns with the result from 2009 data that the BIC forward is the worst model of the four.

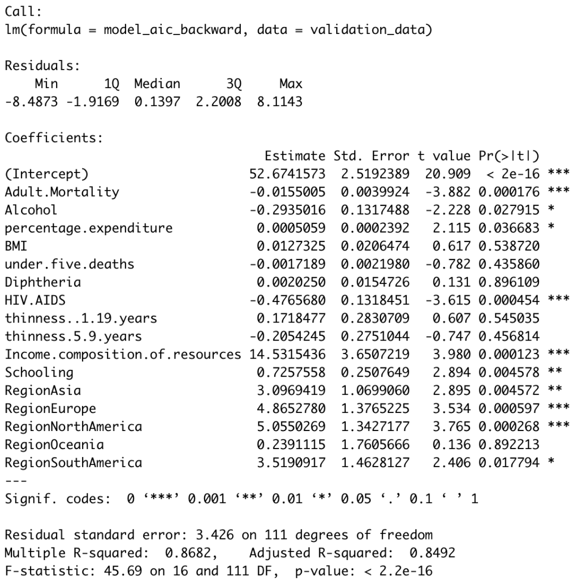
We also analyzed the linear models outputted by training on the 2010 data, and the results are presented below:

AIC Forward



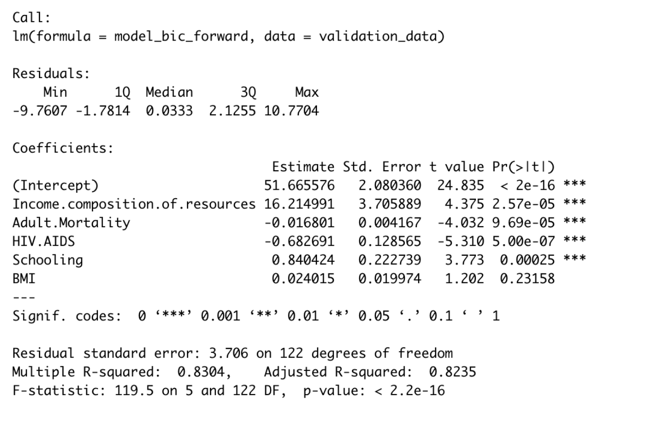
```
Call:
lm(formula = model_aic_forward, data = validation_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.3095 -1.8475 -0.0169  2.3257  8.0999 

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)                    52.5115601  2.3445808  22.397  < 2e-16 ***
Income.composition.of.resources 14.6308138  3.5667478   4.102 7.75e-05 ***
Adult.Mortality                 -0.0156746  0.0039555  -3.963 0.000130 ***
HIV.AIDS                        -0.4793001  0.1305442  -3.672 0.000370 ***
Schooling                        0.7181850  0.2470314   2.907 0.004390 ** 
BMI                              0.0149724  0.0198297   0.755 0.451790    
under.five.deaths               -0.0019595  0.0019468  -1.007 0.316310    
Diphtheria                       0.0024619  0.0153582   0.160 0.872934    
RegionAsia                       2.8929582  1.0312831   2.805 0.005922 ** 
RegionEurope                     4.8830982  1.3544898   3.605 0.000466 ***
RegionNorthAmerica               5.0777555  1.3044965   3.893 0.000168 ***
RegionOceania                    0.2652719  1.7074272   0.155 0.876812    
RegionSouthAmerica               3.5473831  1.4325219   2.476 0.014756 *  
percentage.expenditure           0.0005048  0.0002374   2.126 0.035677 *  
Alcohol                         -0.2948770  0.1288178  -2.289 0.023931 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.404 on 113 degrees of freedom
Multiple R-squared:  0.8675,    Adjusted R-squared:  0.851 
F-statistic: 52.82 on 14 and 113 DF,  p-value: < 2.2e-16
```

AIC Backward



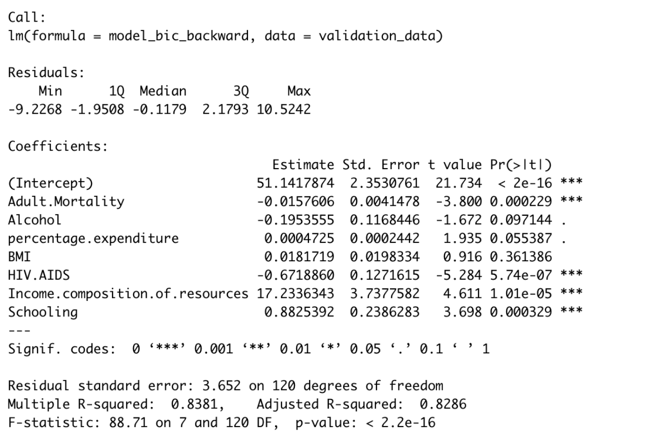
```
Call:
lm(formula = model_aic_backward, data = validation_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.4873 -1.9169  0.1397  2.2008  8.1143 

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)                    52.6741573  2.5192389  20.909  < 2e-16 ***
Adult.Mortality                 -0.0155005  0.0039924  -3.882 0.000176 ***
Alcohol                         -0.2935016  0.1317488  -2.228 0.027915 *  
percentage.expenditure           0.0005059  0.0002392   2.115 0.036683 *  
BMI                              0.0127325  0.0206474   0.617 0.538720    
under.five.deaths               -0.0017189  0.0021980  -0.782 0.435860    
Diphtheria                       0.0020250  0.0154726   0.131 0.896109    
HIV.AIDS                        -0.4765680  0.1318451  -3.615 0.000454 ***
thinness..1.19.years             0.1718477  0.2830709   0.607 0.545035    
thinness.5.9.years              -0.2054245  0.2751044  -0.747 0.456814    
Income.composition.of.resources 14.5315436  3.6507219   3.980 0.000123 ***
Schooling                        0.7257558  0.2507649   2.894 0.004578 ** 
RegionAsia                       3.0969419  1.0699060   2.895 0.004572 ** 
RegionEurope                     4.8652780  1.3765225   3.534 0.000597 ***
RegionNorthAmerica               5.0550269  1.3427177   3.765 0.000268 ***
RegionOceania                    0.2391115  1.7605666   0.136 0.892213    
RegionSouthAmerica               3.5190917  1.4628127   2.406 0.017794 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.426 on 111 degrees of freedom
Multiple R-squared:  0.8682,    Adjusted R-squared:  0.8492 
F-statistic: 45.69 on 16 and 111 DF,  p-value: < 2.2e-16
```

BIC Forward



```
Call:
lm(formula = model_bic_forward, data = validation_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.7607 -1.7814  0.0333  2.1255 10.7704 

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)    
(Intercept)                    51.665576   2.080360  24.835  < 2e-16 ***
Income.composition.of.resources 16.214991   3.705889   4.375 2.57e-05 ***
Adult.Mortality                -0.016801   0.004167  -4.032 9.69e-05 ***
HIV.AIDS                       -0.682691   0.128565  -5.310 5.00e-07 ***
Schooling                       0.840424   0.222739   3.773  0.00025 ***
BMI                             0.024015   0.019974   1.202  0.23158    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.706 on 122 degrees of freedom
Multiple R-squared:  0.8304,	Adjusted R-squared:  0.8235 
F-statistic: 119.5 on 5 and 122 DF,  p-value: < 2.2e-16
```

BIC Backward



```
Call:
lm(formula = model_bic_backward, data = validation_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.2268 -1.9508 -0.1179  2.1793 10.5242 

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)                    51.1417874  2.3530761  21.734  < 2e-16 ***
Adult.Mortality                -0.0157606  0.0041478  -3.800 0.000229 ***
Alcohol                        -0.1953555  0.1168446  -1.672 0.097144 .  
percentage.expenditure          0.0004725  0.0002442   1.935 0.055387 .  
BMI                             0.0181719  0.0198334   0.916 0.361386    
HIV.AIDS                       -0.6718860  0.1271615  -5.284 5.74e-07 ***
Income.composition.of.resources 17.2336343  3.7377582   4.611 1.01e-05 ***
Schooling                       0.8825392  0.2386283   3.698 0.000329 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.652 on 120 degrees of freedom
Multiple R-squared:  0.8381,	Adjusted R-squared:  0.8286 
F-statistic: 88.71 on 7 and 120 DF,  p-value: < 2.2e-16
```

Some things we notice are as follows:

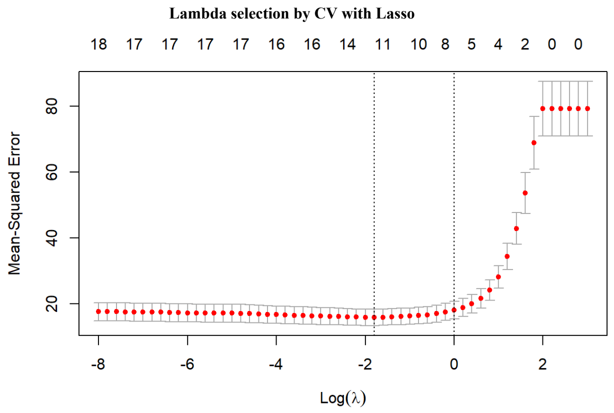
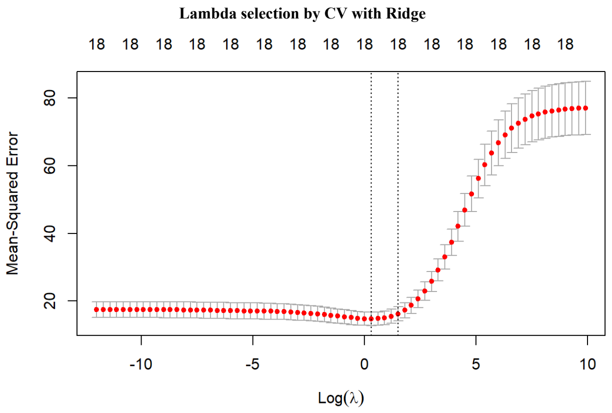
- The AIC models have a higher R squared value of 0.851 compared to the BIC models. This agrees with our cross validation RMSE on the data from 2010, which means that our parsimonious models outputted by the BIC selection techniques end up having too much omitted variable bias.
- We can also see that a lot of the covariates, especially in the AIC models have very high p-values, indicating that they have lower statistical significance. This underscores the importance of using a validation dataset from a different year to analyze trends that are consistent across multiple years and to identify variables that have a more stable relationship with the response variable. By doing so, we can build more robust models that are better suited for making accurate predictions and informed decisions.

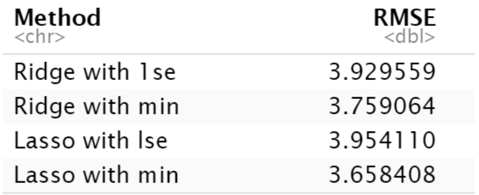
In conclusion, our model selection techniques, specifically forward and backward selection using AIC and BIC, play a crucial role in identifying the subset of variables that have the most significant impact on life expectancy. By using these techniques, we were able to build a more robust and accurate linear model that can predict life expectancy more reliably by reducing the variance inflation caused by collinearity. The AIC backward model, which selected 17 predictors, was found to be the most accurate for predicting the response variable in both the training and validation datasets. Moreover, our analysis emphasized the importance of using a validation dataset from a different year to validate the model and avoid overfitting to a specific year's data. By doing so, we were able to confirm the stability of our model and identify variables that have a more stable relationship with the response variable.

Shrinkage Methods

We now conduct both Lasso and Ridge regression. For these shrinkage methods, we have excluded the Region variable from analysis because when we performed shrinkage after one hot encoding categorical variable Region, Lasso model would select certain Regions (e.g. Asia, North America) as important covariate whereas some other regions as insignificant covariates. We viewed that such feature selections complicate the interpretation of the model. Therefore, we have considered continuous variables when building shrinkage models. Also, similar to forward and backward selection and ANOVA, we used data from 2009 to build these shrinkage models. Below, we use cv.glmnet, which performs k–fold cross validation, to find the optimal lambda for each model. lambda is the tuning parameter for the bias-variance tradeoff and we estimate it using cross validation.

We considered both models selected by ‘lambda.min’ criteria and the ‘lambda.lse’ criteria. For ridge regression, the lambda found for min and lse are 27.11264 and 66.68633, respectively. For Lasso regression, the lambda found for min and lse are 0.54881 and 1.49182, respectively.

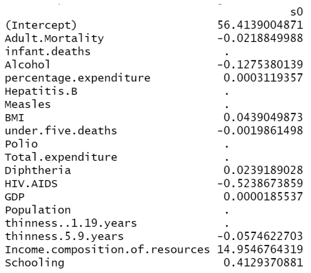
Lasso	**Ridge**
	
The value of lambda that minimizes MSE_{CV} is 0.1653. The corresponding number of variables is 12. The value for lambda that is 1 se above the minimum is 1. The corresponding number of variables is 7.	The value of lambda that minimizes MSE_{CV} is 1.34986. Ridge regression selects all features. The value for lambda that is 1 se above the minimum is 4.48169. Ridge regression still selects all features.



```
Method                        RMSE
<chr>                        <dbl>
Ridge with 1se            3.929559
Ridge with min            3.759064
Lasso with lse            3.954110
Lasso with min            3.658408
```

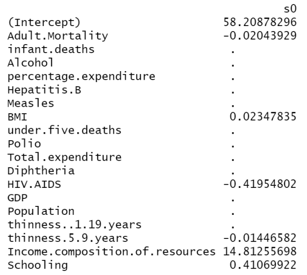
After finding the optimal lambda values for each of these models, we calculated their RMSE using the validation dataset from 2010. The Lasso model with lambda value of 0.1653 (selected using minimum criteria) achieved the smallest RMSE value, suggesting that the Lasso model is more effective than ridge and that the minimum criteria is better for selecting lambda value than the 1se criteria in our context . Ridge regression has notably higher RMSE value than the Lasso regression, and minimum criteria reduced RMSE for Ridge regression as well. The Lasso model suggests that eight features–adult mortality, BMI, Polio, HIV AIDS, GDP, thinness 5.9 years, income composition of resources, and schooling– are necessary for the prediction of life expectancy. The coefficients for each feature of both ridge and lasso models are summarized below:

Lasso (min)



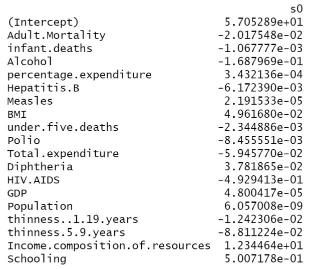
```
                                          s0
(Intercept)                    56.4139004871
Adult.Mortality                -0.0218849988
infant.deaths                   .           
Alcohol                        -0.1275380139
percentage.expenditure          0.0003119357
Hepatitis.B                     .           
Measles                         .           
BMI                             0.0439049873
under.five.deaths              -0.0019861498
Polio                           .           
Total.expenditure               .           
Diphtheria                      0.0239189028
HIV.AIDS                       -0.5238673859
GDP                             0.0000185537
Population                      .           
thinness..1.19.years            .           
thinness.5.9.years             -0.0574622703
Income.composition.of.resources 14.9546764319
Schooling                       0.4129370881
```

Lasso (1se)



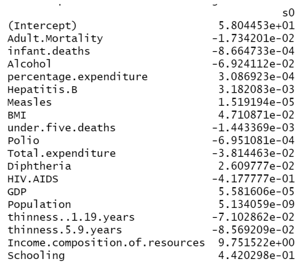
```
                                         s0
(Intercept)                     58.20878296
Adult.Mortality                 -0.02043929
infant.deaths                    .         
Alcohol                          .         
percentage.expenditure           .         
Hepatitis.B                      .         
Measles                          .         
BMI                              0.02347835
under.five.deaths                .         
Polio                            .         
Total.expenditure                .         
Diphtheria                       .         
HIV.AIDS                        -0.41954802
GDP                              .         
Population                       .         
thinness..1.19.years             .         
thinness.5.9.years              -0.01446582
Income.composition.of.resources 14.81255698
Schooling                        0.41069922
```

Ridge (min)



```
                                           s0
(Intercept)                      5.705289e+01
Adult.Mortality                 -2.017548e-02
infant.deaths                   -1.067777e-03
Alcohol                         -1.687969e-01
percentage.expenditure           3.432136e-04
Hepatitis.B                     -6.172390e-03
Measles                          2.191533e-05
BMI                              4.961680e-02
under.five.deaths               -2.344886e-03
Polio                           -8.455551e-03
Total.expenditure               -5.945770e-02
Diphtheria                       3.781865e-02
HIV.AIDS                        -4.929413e-01
GDP                              4.800417e-05
Population                       6.057008e-09
thinness..1.19.years            -1.242306e-02
thinness.5.9.years              -8.811224e-02
Income.composition.of.resources  1.234464e+01
Schooling                        5.007178e-01
```

Ridge (1se)

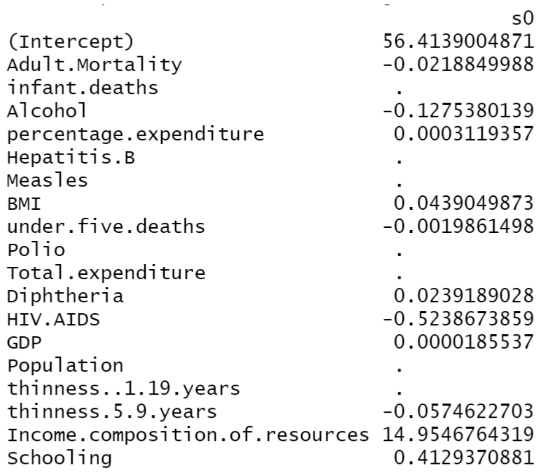


```
                                           s0
(Intercept)                      5.804453e+01
Adult.Mortality                 -1.734201e-02
infant.deaths                   -8.664733e-04
Alcohol                         -6.924112e-02
percentage.expenditure           3.086923e-04
Hepatitis.B                      3.182083e-03
Measles                          1.519194e-05
BMI                              4.710871e-02
under.five.deaths               -1.443369e-03
Polio                           -6.951081e-04
Total.expenditure               -3.814463e-02
Diphtheria                       2.609777e-02
HIV.AIDS                        -4.177777e-01
GDP                              5.581606e-05
Population                       5.134059e-09
thinness..1.19.years            -7.102862e-02
thinness.5.9.years              -8.569209e-02
Income.composition.of.resources  9.751522e+00
Schooling                        4.420298e-01
```

Some things we notice are as follows:

- Depending on the lambda criterion – minimum or 1se – the features selected for Lasso models are different. For example, thinness.5.9.years and polio are selected as significant for the Lasso model with minimum criteria for lambda, whereas these are excluded in the Lasso model with 1se criteria. In a similar vein, while the lasso model built with minimum criteria selected eight features, the Lasso model with lse criteria selected only five features.
- The Lasso model with the lambda that minimizes MSE_{CV} outperforms all the other models we trained including the ones in the previous section. Hence the optimal model we create (considering one year of data) has 11 covariates and the intercept.

The model has these coefficients for the year 2010:

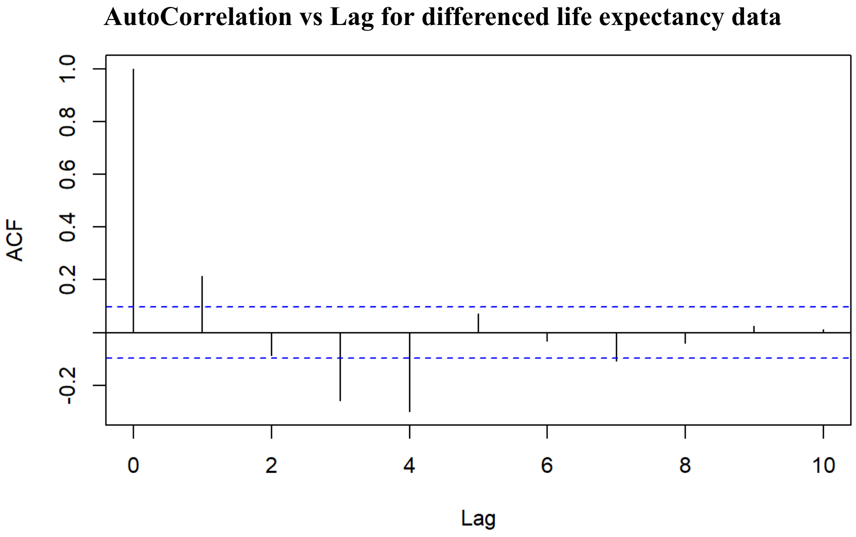


```
                                          s0
(Intercept)                    56.4139004871
Adult.Mortality                -0.0218849988
infant.deaths                   .
Alcohol                        -0.1275380139
percentage.expenditure          0.0003119357
Hepatitis.B                     .
Measles                         .
BMI                             0.0439049873
under.five.deaths              -0.0019861498
Polio                           .
Total.expenditure               .
Diphtheria                      0.0239189028
HIV.AIDS                       -0.5238673859
GDP                             0.0000185537
Population                      .
thinness..1.19.years            .
thinness.5.9.years             -0.0574622703
Income.composition.of.resources 14.9546764319
Schooling                       0.4129370881
```

Time Series Regression

Although our previous analysis only considered data from 2009, we wanted to validate our models for each year to ensure their robustness. Upon rerunning our models for each year, we found that the results were consistent across all years. The main difference between the models was the intercept coefficient (β_0), indicating a relationship between the year covariate and life expectancy. Our earlier plot in the assumption section indicated a linear relationship. With this in mind, we decided to investigate the relationship between time and life expectancy further by using an ARMA model.

As we observed in the earlier section, the differenced data exhibited a stationary distribution. After analyzing the autocorrelations of the differenced data, we were able to plot the results and gain further insight into the relationship between time and life expectancy.



After analyzing the plot above, we observed that the ACF was significant at lag = 0, 1, 3, and 4. This led us to consider two possible ARIMA models: MA(4) or a seasonal ARMA$(0,0,1)x(0,0,1)_4$ model (also called a SARIMA model).

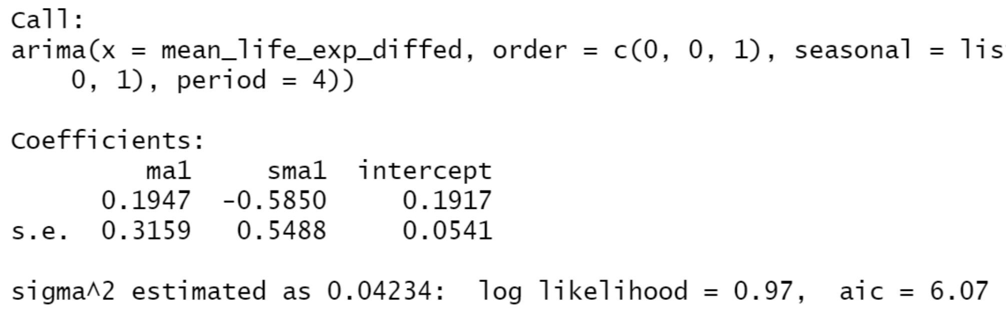
These are the formulas for the two ARIMA models we considered:

MA(4) $\Leftrightarrow X_t = \mu + W_t + \theta_1 W_{t-1} + \theta_2 W_{t-2} + \theta_3 W_{t-3} + \theta_4 W_{t-4}$, where $W_t \sim WN(0, \sigma^2)$ for all t

ARMA$(0,0,1)x(0,0,1)_4 \Leftrightarrow X_t = \mu + (1 + \theta_1 B)(1 + \theta_2 B)^4 W_t$, where $W_t \sim WN(0, \sigma^2)$ for all t, while B is the backshift operator.

The backshift operator is the operator which connects W_t to W_{t-i} such that $W_t B^i = W_{t-i}$ for all t, i.

After experimenting with both models, we found that the MA(4) model suffered from numerical instability, while the ARMA$(0,0,1)x(0,0,1)_2$ gave us the below fit.

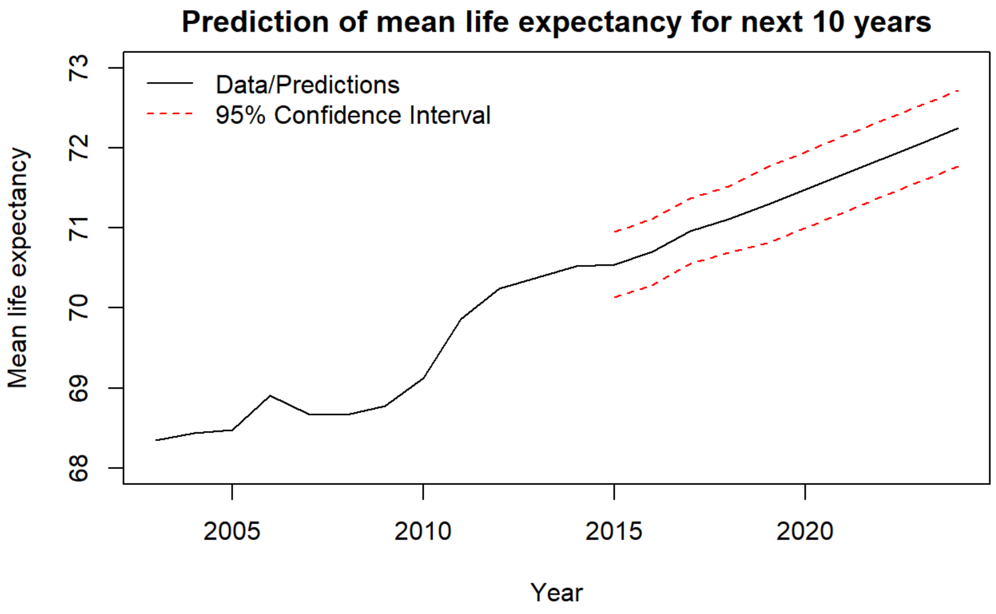


```
Call:
arima(x = mean_life_exp_diffed, order = c(0, 0, 1), seasonal = list(order = c(0,
    0, 1), period = 4))

Coefficients:
         ma1     sma1  intercept
      0.1947  -0.5850     0.1917
s.e.  0.3159   0.5488     0.0541

sigma^2 estimated as 0.04234:  log likelihood = 0.97,  aic = 6.07
```

Here we can see that if we use the ARMA$(0,0,1)x(0,0,1)_4 \Leftrightarrow X_t = \mu + (1 + \theta_1 B)(1 + \theta_2 B)^4 W_t$ model, we would call X_t as the differenced mean life expectancy at time t, and we achieve our coefficients as:
$\theta_1 = 0.1947$, SE$(\theta_1) = 0.3159$, $\theta_2 = -0.5850$, SE$(\theta_2) = 0.5488$, $\mu = 0.1917$, SE$(\mu) = 0.0541$

So our final model for the relationship between the mean life expectancy (L_y) and year (y) would be:
$L_y = \mu + L_{y-1} + (1 + \theta_1 B)(1 + \theta_2 B)^4 W_y$, where θ_1 and θ_2 are defined as above and W_y is White noise with mean = 0 and variance = 0.04234. The small variance also indicates that this ARMA model is very close to a linear model with a very small amount of White Noise added.

From looking at the above equation, it is evident that the life expectancy is growing per year at an expected rate of 0.1917 per year. On analysis of our data, we noticed that the intercept term (taking categorical variables into consideration) was also growing at a similar rate. This information when put together with the rest of our modeling allows us to generalize our results across multiple future years, by scaling the intercept to be
$\beta_y = \beta_{2010} + \Sigma^y_{i=2010}$ pred$_i$ where pred$_i$ is the prediction made by the SARIMA model.
When predicting the mean life expectancy for the next 10 years, the following predictions were obtained:



Overall, this modeling tool was vital to our analysis in many ways.

- It allowed us to generalize our results over multiple years and account for the fact that the observations were temporal and not independent across years. By taking into consideration the temporal nature of the data, we were able to generate predictions for future years and gain deeper insight into the relationship between time and life expectancy.
- Tied in with the model we have built in our earlier section, this model can predict life expectancy for any given time period and country.

Conclusion

Based on the analysis performed in this modeling report, we were able to create a model that considered various factors affecting life expectancy. Our analysis used a wide variety of statistical tools, including ANOVA, time series regression, model diagnostics, model selection, and shrinkage methods.

During our analysis, we uncovered some intriguing findings that shed light on various aspects of life expectancy:

- The ANOVA section revealed that the region covariate contains a wealth of valuable information that has a significant impact on life expectancy. On the other hand, we found that an interaction term considering both region and status would not be very useful in our analysis.
- The AIC/BIC and shrinkage sections of our study played a critical role in identifying the most useful covariates to include in our model. By using these techniques, we were able to evaluate the statistical significance and explanatory power of each covariate and select a subset of the most important predictors for our model. Furthermore, we assessed the robustness of our model by validating its performance on data from a different year than the one used for model selection. This approach allowed us to check whether the covariates selected in our model were still relevant and necessary for accurate predictions. This step helped to ensure that our model was not overfitting to a specific year's data, which could lead to biased predictions and erroneous conclusions.
- The time series regression section provided us with valuable insights into the long-term trends of life expectancy. We discovered that on average, life expectancy increased at a steady rate of 0.1917 years per year. Furthermore, our analysis revealed a very close to linear trend in the trajectory of life expectancy, suggesting that we can make reasonably accurate predictions about life expectancy in the future.

A few challenges we faced were dealing with NA values, the country categorical variable, and addressing the temporal aspect of the data. Through careful analysis, we were able to justifiably remove NA rows and replace the country category with region, which improved the model's accuracy. Using time series regression techniques, we fit a SARIMA model to examine how the life expectancy variable changes over time, which helped to account for the dependence between observations.

In the future, we would like to study a few more things:

- First, we would like to see how each model changes when trained on the data collected after 2015 because this dataset goes only up to 2015. There has been much more growth in industries, medical technologies in the past ten years, and therefore, the models that we have built with datasets before 2015 is not a good representation of the life expectancy model for 2023. Therefore, we would like to retrain a model with more recent data. Additionally, we are also interested in how global pandemic COVID 19 has changed the relationship between covariates and life expectancy. COVID 19 has brought many changes in healthcare industries, general public’s awareness of health, as well as policies. We would like to compare the models that are built before COVID with models trained on after COVID to see how coronavirus affected life expectancy.
- We would also like to combine our time series model and our model we achieved from variable selection and see if we can build a model which can predict the life expectancy for any value of country in any given year, even if it is not a year in our dataset.
- We would like to conduct an experiment to try to determine causal inference and find which factors cause changes in life expectancy.

Overall, the findings from this modeling report provide valuable insights that can help identify and improve the factors influencing life expectancy. The use of statistical techniques and careful analysis of the data allowed us to identify important factors that affect life expectancy. By understanding these factors, policymakers can make informed decisions that promote longer and healthier lives for people around the world.