Project 3

Wilber Delgado & Mason Delan

2023-12-03

```
library("readxl")
WHR = read_excel("D:/WHR2023.xls")
```

We are working with the World Happiness Report data set, which is a collection of surveys from around the world, asking peoples different judgement on a variety of things that could influence happiness, such as social support, freedom to make life choices, generosity, etc. To get a better set of data we are going to drop the country column, and drop any rows with missing data. Then we are going to take a random sample of 500 to get a normalized dataset.

Put some data to the side

```
## Warning: package 'caret' was built under R version 4.3.2
```

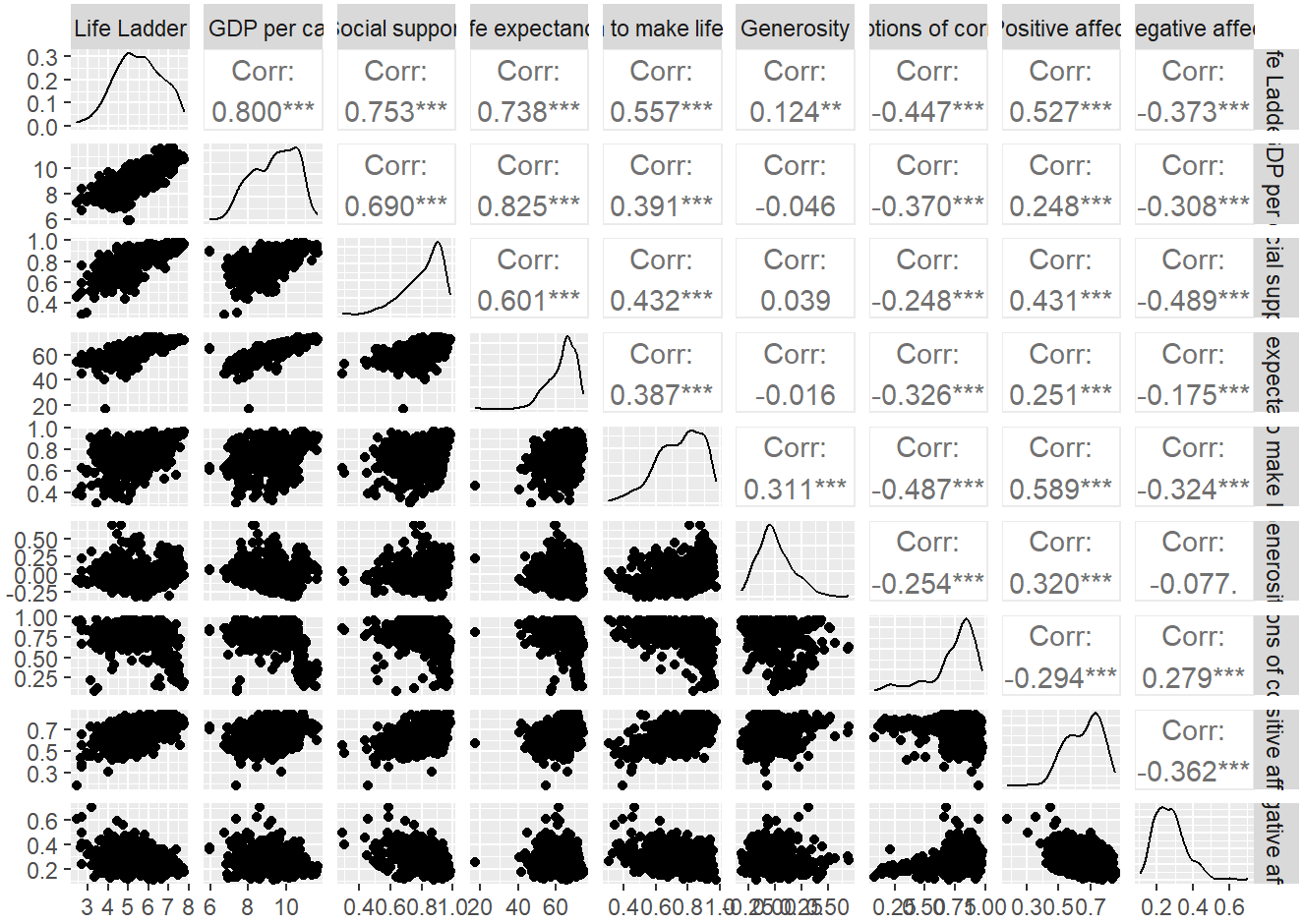
```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 4.3.2
```

```
## Loading required package: lattice
```

```
## Warning: package 'GGally' was built under R version 4.3.2
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```


Life Ladder
GDP per ca
ocial suppor
fe expectanc
to make life
Generosity
otions of cor
Positive affec
egative affe
Corr: 0.800***
Corr: 0.753***
Corr: 0.738***
Corr: 0.557***
Corr: 0.124**
Corr: -0.447***
Corr: 0.527***
Corr: -0.373***
Corr: 0.690***
Corr: 0.825***
Corr: 0.391***
Corr: -0.046
Corr: -0.370***
Corr: 0.248***
Corr: -0.308***
Corr: 0.601***
Corr: 0.432***
Corr: 0.039
Corr: -0.248***
Corr: 0.431***
Corr: -0.489***
Corr: 0.387***
Corr: -0.016
Corr: -0.326***
Corr: 0.251***
Corr: -0.175***
Corr: 0.311***
Corr: -0.487***
Corr: 0.589***
Corr: -0.324***
Corr: -0.254***
Corr: 0.320***
Corr: -0.077.
Corr: -0.294***
Corr: 0.279***
Corr: -0.362***

```
##                                Life Ladder Log GDP per capita Social support
## Life Ladder                      1.0000000         0.79972358     0.75269728
## Log GDP per capita               0.7997236         1.00000000     0.68986520
## Social support                   0.7526973         0.68986520     1.00000000
## Healthy life expectancy at birth 0.7375541         0.82532776     0.60136407
## Freedom to make life choices     0.5565947         0.39095730     0.43183337
## Generosity                       0.1237212        -0.04602347     0.03886257
## Perceptions of corruption       -0.4466865        -0.37044156    -0.24841879
## Positive affect                  0.5274543         0.24795000     0.43128607
## Negative affect                 -0.3733800        -0.30764879    -0.48913341
##                                Healthy life expectancy at birth
## Life Ladder                                           0.7375541
## Log GDP per capita                                    0.8253278
## Social support                                        0.6013641
## Healthy life expectancy at birth                      1.0000000
## Freedom to make life choices                          0.3873445
## Generosity                                           -0.0163528
## Perceptions of corruption                            -0.3263934
## Positive affect                                       0.2513053
## Negative affect                                      -0.1749062
##                                Freedom to make life choices  Generosity
## Life Ladder                                       0.5565947  0.12372120
## Log GDP per capita                                0.3909573 -0.04602347
## Social support                                    0.4318334  0.03886257
## Healthy life expectancy at birth                  0.3873445 -0.01635280
## Freedom to make life choices                      1.0000000  0.31106699
## Generosity                                        0.3110670  1.00000000
## Perceptions of corruption                        -0.4870299 -0.25419026
## Positive affect                                   0.5893305  0.32014928
## Negative affect                                  -0.3243702 -0.07718074
##                                Perceptions of corruption Positive affect
## Life Ladder                                   -0.4466865       0.5274543
## Log GDP per capita                            -0.3704416       0.2479500
## Social support                                -0.2484188       0.4312861
## Healthy life expectancy at birth              -0.3263934       0.2513053
## Freedom to make life choices                  -0.4870299       0.5893305
## Generosity                                    -0.2541903       0.3201493
## Perceptions of corruption                      1.0000000      -0.2941284
## Positive affect                               -0.2941284       1.0000000
## Negative affect                                0.2786628      -0.3619913
##                                Negative affect
## Life Ladder                        -0.37337997
## Log GDP per capita                 -0.30764879
## Social support                     -0.48913341
## Healthy life expectancy at birth   -0.17490617
## Freedom to make life choices       -0.32437016
## Generosity                         -0.07718074
## Perceptions of corruption           0.27866281
## Positive affect                    -0.36199128
## Negative affect                     1.00000000
```

Based on the ggpairs plots we can see that our response variable ‘Healthy Life Expectancy at Birth’ is highly correlated to ‘Life Ladder’ and ‘GDP Per Capita’.

The correlation between “Life Ladder” and “Freedom to make life choices” is 0.5566, indicating a moderate positive correlation. Depending on the nature of the relationship, you might explore the inclusion of a quadratic term to capture potential non-linearities.

The correlation between “Life Ladder” and “Perceptions of corruption” is -0.4467. This negative correlation suggests that as perceptions of corruption decrease, life satisfaction tends to increase. In some cases, a log transformation might be considered if the relationship is better captured on a log scale.

There might be potential interactions between variables that could enhance the model. For example, interactions between “Freedom to make life choices” and other variables could be explored, given its moderate correlation with “Life Ladder.”

As far as feature engineering we will be taking out any rows that have missing data. Also, we are not going to use any of the interaction variables.

```
##
## Call:
## lm(formula = `Healthy life expectancy at birth` ~ . + `Life Ladder` *
##     `Freedom to make life choices`, data = training_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.615  -1.432   0.374   1.974  12.904
##
## Coefficients:
##                                           Estimate Std. Error t value
## (Intercept)                             -288.53101   92.27266  -3.127
## year                                       0.14491    0.04582   3.162
## `Life Ladder`                              4.13934    1.08064   3.830
## `Log GDP per capita`                       3.61674    0.32088  11.271
## `Social support`                           3.73413    2.64704   1.411
## `Freedom to make life choices`            18.37125    6.83247   2.689
## Generosity                                -0.23814    1.30826  -0.182
## `Perceptions of corruption`               -1.32628    1.38149  -0.960
## `Positive affect`                         -1.50049    2.61817  -0.573
## `Negative affect`                         10.11501    2.64160   3.829
## `Life Ladder`:`Freedom to make life choices`   -3.38529    1.30976  -2.585
##                                           Pr(>|t|)
## (Intercept)                               0.001899 **
## year                                      0.001687 **
## `Life Ladder`                             0.000149 ***
## `Log GDP per capita`                       < 2e-16 ***
## `Social support`                          0.159137
## `Freedom to make life choices`            0.007478 **
## Generosity                                0.855657
## `Perceptions of corruption`               0.337629
## `Positive affect`                         0.566904
## `Negative affect`                         0.000150 ***
## `Life Ladder`:`Freedom to make life choices` 0.010110 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.93 on 390 degrees of freedom
## Multiple R-squared:  0.7046, Adjusted R-squared:  0.697
## F-statistic: 93.01 on 10 and 390 DF,  p-value: < 2.2e-16
```

```
## Warning: package 'randomForest' was built under R version 4.3.2
```

```
## randomForest 4.7-1.1
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
##
## Call:
##  randomForest(formula = Healthy.life.expectancy.at.birth ~ .,      data = training_data)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 3
##
##           Mean of squared residuals: 13.15291
##                     % Var explained: 74.14
```

We see that our linear model explains about 69.7% of the variability in the response variable. Key significant predictors include Life Ladder, Log GDP per capita, and the interaction term between Life Ladder and Freedom to make life choices. The Random Forest model explains approximately 74.14% of the variability in the response variable. Both models seem to perform reasonably well.

```
## Warning: Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
```

```
## Linear Regression
##
## 401 samples
##   9 predictor
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 321, 321, 321, 321, 320
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   3.894543  0.7061887  2.643731
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
## Warning: Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
```

```
## Random Forest
##
## 401 samples
##   9 predictor
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 321, 320, 321, 321, 321
## Resampling results across tuning parameters:
##
##   mtry  RMSE      Rsquared   MAE
##   2     3.574180  0.7638669  2.396839
##   5     3.522125  0.7670493  2.303159
##   9     3.620048  0.7565244  2.330448
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was mtry = 5.
```

```
## RMSE for Multiple Linear Regression: 3.894543
```

```
## RMSE for Random Forest: 3.522125
```

Based on our results, the Random Forest model seems to have a lower RMSE which indicates it is likely a better model for predictive performance.

```
##
## Call:
## lm(formula = Healthy.life.expectancy.at.birth ~ year + Life.Ladder +
##     Log.GDP.per.capita + Negative.affect, data = training_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.037  -1.466   0.292   1.954  14.029
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -270.94242   88.05331  -3.077 0.002236 **
## year                  0.14262    0.04393   3.247 0.001266 **
## Life.Ladder           1.66570    0.29096   5.725 2.05e-08 ***
## Log.GDP.per.capita    3.80328    0.28589  13.303  < 2e-16 ***
## Negative.affect       8.10927    2.41066   3.364 0.000843 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.948 on 396 degrees of freedom
## Multiple R-squared:  0.6974, Adjusted R-squared:  0.6943
## F-statistic: 228.2 on 4 and 396 DF,  p-value: < 2.2e-16
```


```
##
## Call:
## lm(formula = Healthy.life.expectancy.at.birth ~ year + Life.Ladder +
##     Log.GDP.per.capita + Negative.affect, data = training_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.037  -1.466   0.292   1.954  14.029
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -270.94242   88.05331  -3.077 0.002236 **
## year                  0.14262    0.04393   3.247 0.001266 **
## Life.Ladder           1.66570    0.29096   5.725 2.05e-08 ***
## Log.GDP.per.capita    3.80328    0.28589  13.303  < 2e-16 ***
## Negative.affect       8.10927    2.41066   3.364 0.000843 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.948 on 396 degrees of freedom
## Multiple R-squared:  0.6974, Adjusted R-squared:  0.6943
## F-statistic: 228.2 on 4 and 396 DF,  p-value: < 2.2e-16
```


In selecting the best statistical model for predicting Healthy Life Expectancy at birth, a step wise variable selection approach was used, utilizing both forward and backward steps guided by AIC and BIC. This method balances model complexity and the accuracy of the fit. After considering various model specifications, the chosen model includes the predictors year, Life Ladder, Log GDP per capita, and Negative affect. The AIC and BIC, which penalize over fitting, indicated that this set of variables provides a good trade-off between explanatory power and

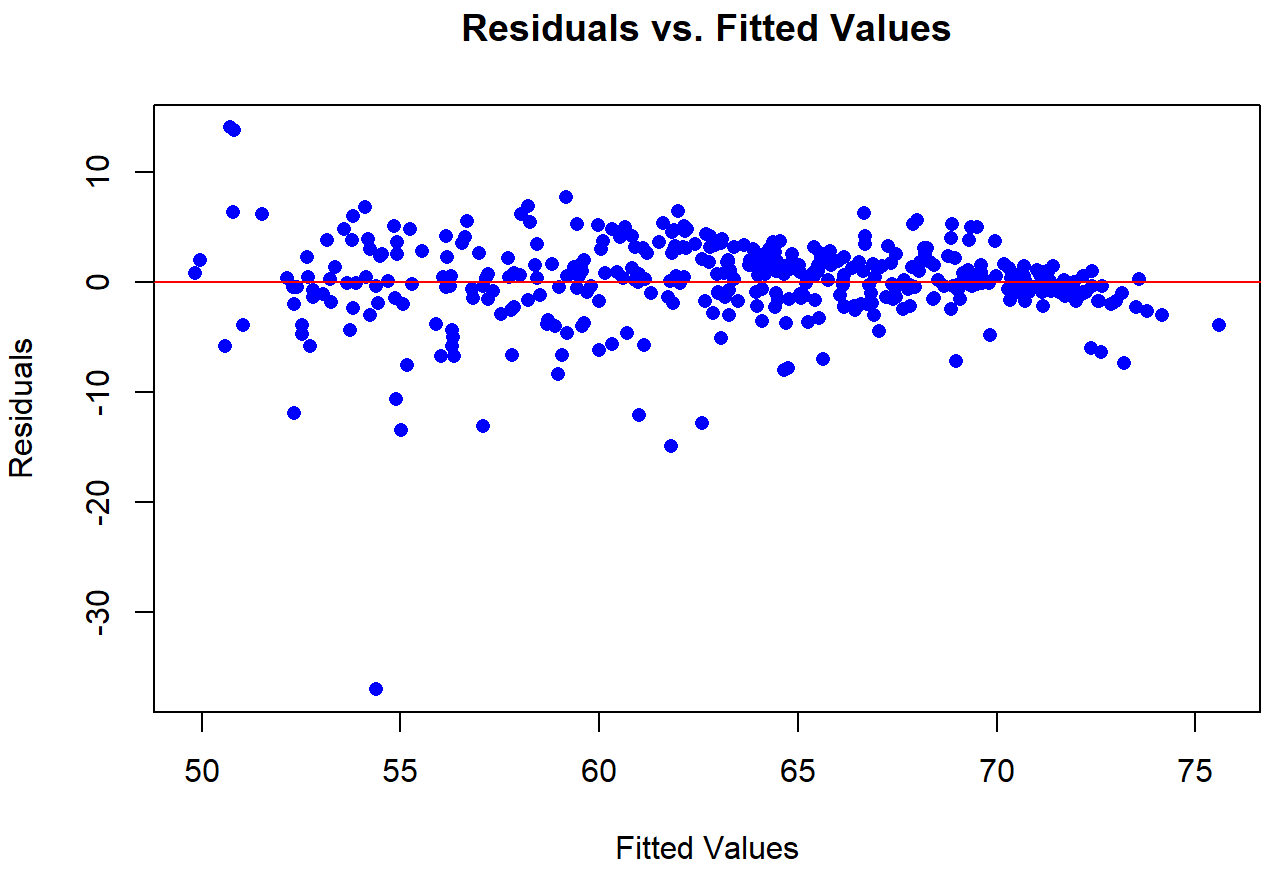
simplicity. The final model demonstrated a high adjusted R-squared value of 0.6943, indicating that approximately 69.43% of the variability in Healthy Life Expectancy at birth is explained by the selected predictors. Additionally, the F-statistic was highly significant, providing further evidence of the model's overall significance. These results suggest that the chosen model is a great representation of the relationship between the predictors and the response variable.

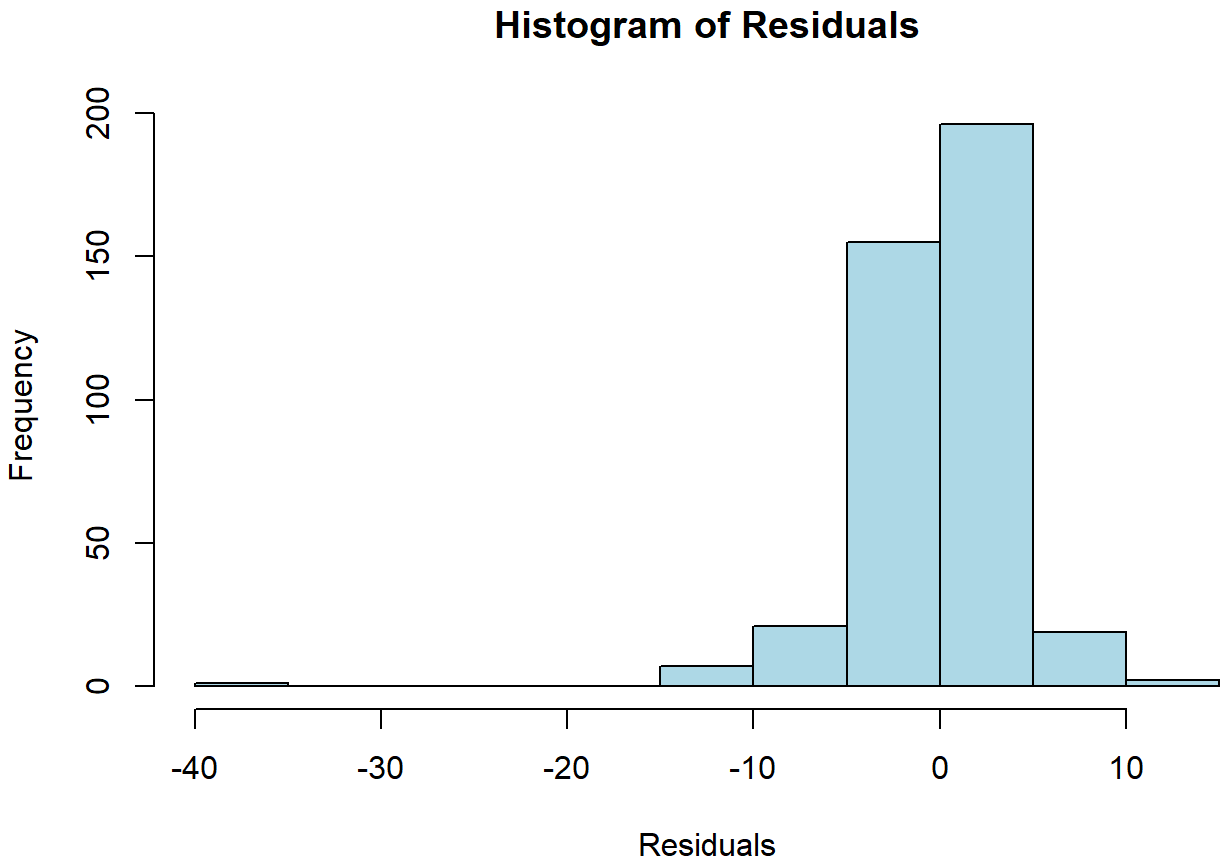
The intercept, -270.94, represents the estimated Healthy Life Expectancy at birth when all predictor variables are zero. The coefficient for year suggests that, on average, each additional year is associated with an increase of approximately 0.1426 units in Healthy Life Expectancy at birth. For each one-unit increase in the Life Ladder score, we expect Healthy Life Expectancy at birth to increase by approximately 1.6657 units. The coefficient implies that a one-unit increase in the logarithm of GDP per capita is associated with an estimated increase of about 3.8033 units in Healthy Life Expectancy at birth. Finally, a one-unit increase in Negative Affect is associated with an estimated increase of about 8.1093 units in Healthy Life Expectancy at birth.

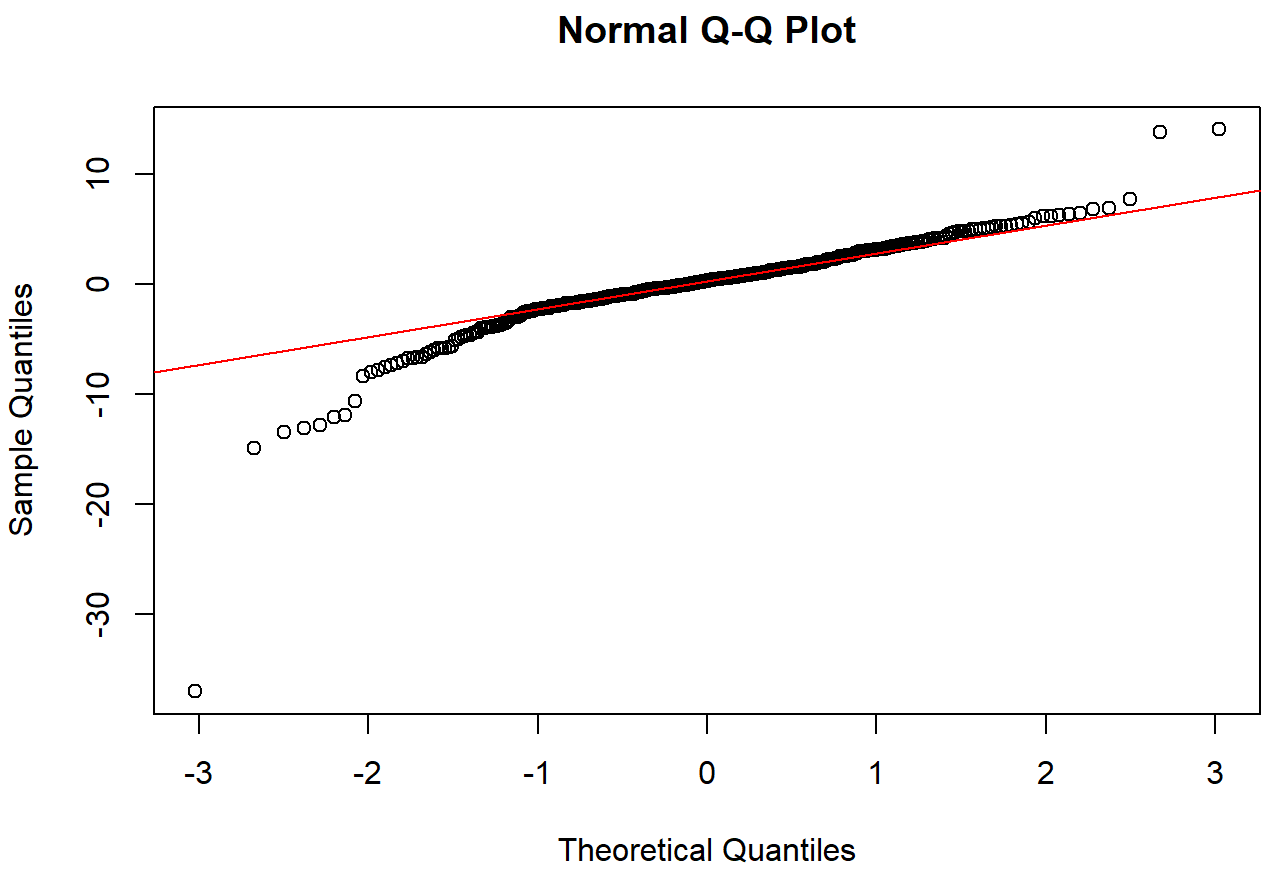
```
## Test R-squared: 0.7948604
```

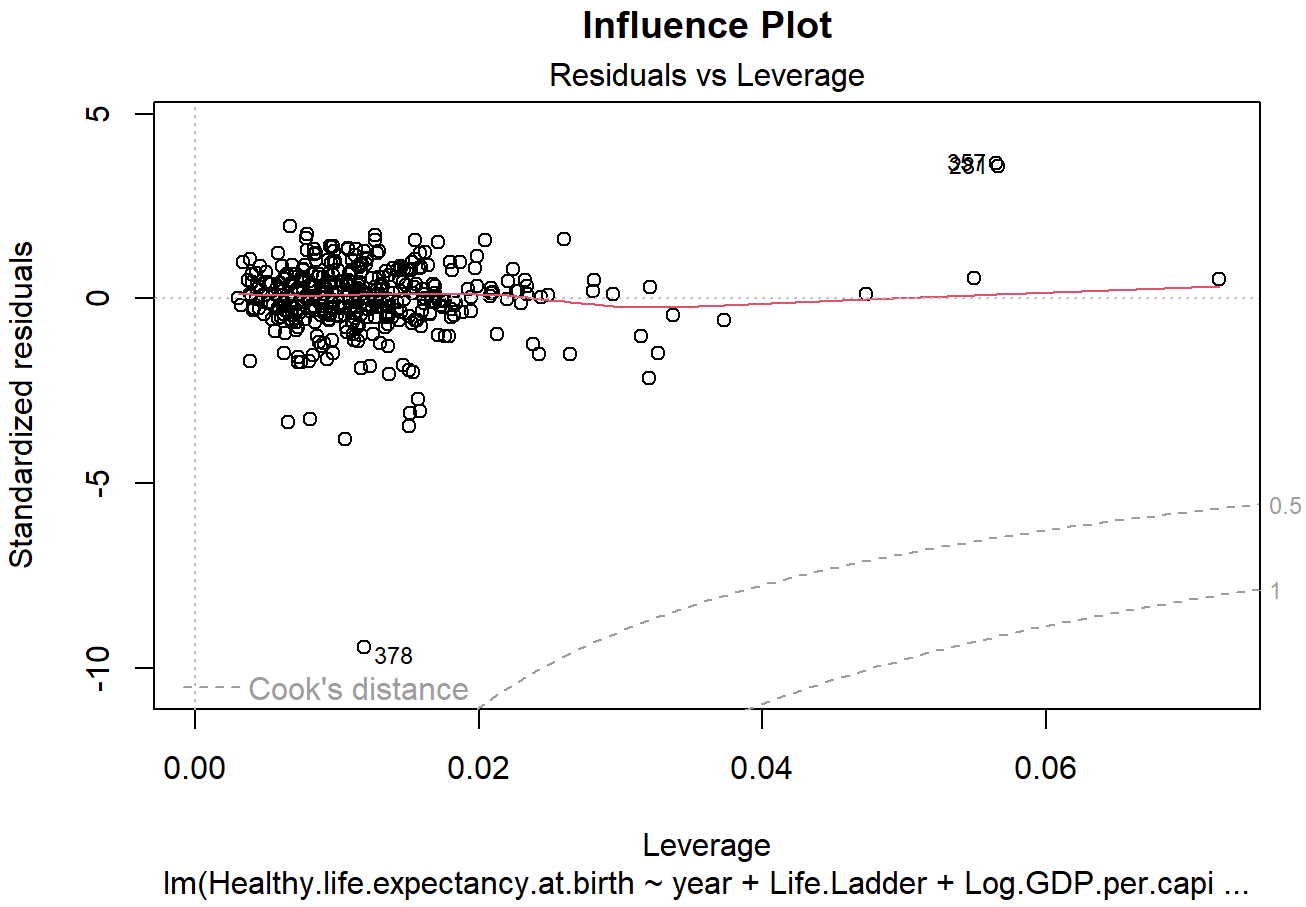
```
## Test adjusted R-squared: 0.7861311
```

The test R-squared value is 0.7948604, and the test adjusted R-squared value is 0.7861311. These are relatively high values, suggesting that the model performs well in explaining the variability in the Healthy life expectancy at birth. A high R-squared does not guarantee that the model will accurately describe the population. There might be other factors or unobserved variables that could influence the response variable.


Residuals vs. Fitted Values
Residuals
10
0
-10
-20
-30
50
55
60
65
70
75
Fitted Values


Histogram of Residuals
Frequency
200
150
100
50
0
-40
-30
-20
-10
0
10
Residuals


Normal Q-Q Plot
Sample Quantiles
10
0
-10
-20
-30
-3
-2
-1
0
1
2
3
Theoretical Quantiles


Influence Plot
Residuals vs Leverage
Standardized residuals
5
0
-5
-10
357
378
Cook's distance
0.5
1
0.00
0.02
0.04
0.06
Leverage
lm(Healthy.life.expectancy.at.birth ~ year + Life.Ladder + Log.GDP.per.capi ...

Based on the Residuals vs. Fitted Values, with the points around the line, we can assume that the model assumptions are met (such as heteroscedasticity, non-linearity, etc.). The concentration of residuals in the central range might indicate that the model is performing well for a majority of observations. The deviation from the line at the extremes could suggest departures from normality, possibly indicating the presence of outliers or non-normal features in the data. The majority of observations have low to moderate influence on the model, as indicated by their distance to the red line in the lower Cook's Distance range. A few points deviate from the red line, indicating higher Cook's Distance.

```
## [1] "Confidence Interval for Mean Predicted Value:"
```

```
##        fit      lwr      upr
## 1 78.19563 76.67606 79.71519
```

```
## [1] "Prediction Interval for Future Predicted Value:"
```

```
##        fit      lwr      upr
## 1 78.19563 70.28743 86.10382
```

The model predicts a mean Healthy Life Expectancy at Birth of approximately 78.20, with a 95% confidence interval ranging from 76.68 to 79.72. This implies that, based on the given set of predictors, we are reasonably confident that the true average Healthy Life Expectancy at Birth falls within this interval. Also, for an individual future observation, the model predicts a value of 78.20, and we are 95% confident that the actual value will fall within the broader prediction interval of 70.29 to 86.10. This wider range accounts for the variability in individual observations and highlights the importance of considering prediction intervals for a more comprehensive understanding of the model's predictive capability.

In conclusion, our analysis aimed to understand the factors influencing healthy life expectancy at birth. We initially explored feature engineering, model selection, and cross-validation to determine the most suitable predictive models. The random forest model and a multiple linear regression model were considered, with the latter incorporating interaction terms. Cross-validation results favored the random forest model, which exhibited a lower RMSE. However, for interpretability and simplicity, we opted for a multiple linear regression model. Stepwise selection, based on AIC and BIC, helped refine the model to include the 'year,' 'Life Ladder,' 'Log GDP per capita,' and 'Negative affect' as significant predictors.