MDA 9159A: Data analysis project

Dataset:

Life Expectancy (WHO)

Statistical Analysis on Factors Influencing Life Expectancy

Muzhi Lyu

Zirong Liu

Introduction

The Kaggle data originates from two significant sources: the Global Health Observatory (GHO) data repository, a division of the World Health Organization (WHO), and economic data from the United Nations. The data is concentrated on the years 2000 to 2015, a pivotal period for healthcare development. In total, it compiles information on 193 countries, resulting in a dataset with 22 columns and 2938 rows. These columns encompass various aspects, such as life expectancy, health factors, and economic indicators, categorized under headings like immunization, mortality, economy, and social factors.

The dataset incorporates data on different countries, years, and development statuses (developing or developed), encompassing diverse factors like life expectancy, adult mortality, infant mortality, alcohol consumption, expenditure percentage, hepatitis B, measles, body mass index (BMI), under-five mortality rate, polio, total expenditure, diphtheria, HIV/AIDS, gross domestic product (GDP), population, leanness in age groups 1-19 and 5-9, resources, income composition, and school attendance.

This dataset holds significance as it allows each country to identify factors influencing life expectancy, thereby enabling targeted interventions. The analysis is driven by the aim to provide valuable insights for developing policies and initiatives, with the goal of contributing to improvements in global public health and overall well-being.

In our exploration of this dataset, we built our study upon existing research, drawing insights from recent studies such as Rohan et al. (2022) and Bali et al. (2021). Rohan's investigation into health metrics encompasses a comprehensive outlook, indicating that High gross national income (GNI), moderate body mass index (BMI) and increased public health expenditure positively influence average life expectancy. Bali's work extends exploration, enhancing the predictive accuracy of life expectancy models by shedding light on the nuanced connections between education, health, economics, and social welfare. Our research assumption builds upon these critical studies, aiming to address gaps in the current understanding of life expectancy determinants.

Analysis

Data Cleaning:

In this analysis, we initiate the process by loading and exploring the life expectancy dataset. The initial examination includes checking the data's structure, removing empty rows, and addressing missing values. Data cleaning and manipulation are facilitated using the 'tidyr' and 'dplyr' packages.

Missing values in columns such as GDP, Population, Total Expenditure, and Hepatitis B are handled by replacing them with the respective median values. This approach aims to maintain the central tendency of the distribution, effectively filling gaps in the dataset. Additionally, the 'Status' column, indicating a country's development status, is transformed into a binary variable, assigning a value of 1 to "Developed" countries and 0 to "Developing". Subsequently, any remaining rows with missing values and duplicate rows are removed to ensure a complete and information-rich dataset.

As suggested by the requirements, we randomly select 1000 observations, creating a more manageable dataset for visualization and analysis. The resulting dataset's structure and summary statistics are displayed, offering insights into its characteristics.

General Understanding of the data: (Graph 0.1)

We employed the 'summary' and 'head' functions for a brief overview of the dataset. Specifically, these functions displayed minimum, median, quartiles, and maximum values for each numerical column.

In streamlining our analysis, we chose a simplified approach by visualizing scatter plots of life expectancy against various predictors. Noteworthy relationships emerged, particularly with "Income Composition of Resources" and "Schooling," where higher values correlated positively with life expectancy. Additionally, the scatter plots of "thinness 5-9 years" and "thinness 1-19 years" exhibited a strong resemblance, aligning with common sense. Conversely, "Total Expenditure" displayed a more random distribution, suggesting a less defined relationship. Remarkably, "Infant Deaths" remained consistently low across life expectancies. These visual insights provide a preliminary understanding of predictor relationships, paving the way for subsequent quantitative analyses to further explore and validate these initial observations.

Variable Selection: (Graph 0.2)

Additionally, we delved into the interplay between these selected predictors by plotting their correlations. These visualizations revealed correlation coefficients ranging from 1 to -0.8, providing valuable insights into the relationships among variables. To streamline our analysis and focus on the most influential predictors, we set a cutoff threshold of 0.4. Consequently, we retained approximately half of the predictors that exhibited stronger correlations with life expectancy. Subsequently, we pruned our dataset by removing the columns deemed less influential, aiming to simplify the subsequent stages of our analysis.

Estimation and Training:

The dataset was split into training dataset and testing datasets, with the former (80%) used for model training and optimization.

Model Evaluation:

To streamline the evaluation process and enhance the readability of our code, we implemented two helpful functions: "**evaluation model**" and "**evaluation test**". The former provides a comprehensive overview of the model by presenting a summary, ANOVA test results, and conducting a four-step residual evaluation, including residual tests, QQ plots, Breusch-Pagan tests, and Shapiro-Wilk tests. Additionally, it reports crucial metrics such as MSE and R2. The latter function evaluates the model's performance on the test set, calculating SSE, SST, and SSR to derive R2 and MSE. A scatter plot of observed versus predicted values is generated, where the ideal scenario is represented by points closely following the red line y=x, indicating a precise alignment between observed and predicted values.

Checking Model Assumptions:

Diagnostic plots were generated to assess linear regression assumptions, including homoscedasticity, heteroscedasticity (Breusch-Pagan test), and normality of residuals (Normal Q-Q plot, Shapiro-Wilk test). These checks ensure the validity of the regression model.

Model Interaction:

The introduction of a multiple linear regression model that incorporates interaction terms between various predictors and the 'Status' variable, designed to capture potential differential effects based on development status, is presented below.

Summary

Output and Results of Each Model:

Full Model: (Graph 1.1 1.2 1.3)

The initial model, "model_base", is a comprehensive multiple linear regression model encompassing various predictors: "Status," "Adult.Mortality," "BMI," "GDP," "thinness.5-9 years," "Schooling," "Life expectancy," "percentage expenditure," "HIV.AIDS," "thinness.1.19.years," and "Income.composition.of.resources."

The model exhibits robust performance with a high adjusted R-squared of 0.7914, suggesting a good fit to the data. However, several issues warrant attention. First, the Breusch-Pagan test points to heteroscedasticity in residuals, and the Shapiro-Wilk test indicates non-normality. The QQ Plot reveals deviations from the ideal line, with points above and below as quantiles vary. Second, P-values for GDP, thinness 1-19 years, and thinness 5-9 years are higher than 0.1, questioning their significance. The ANOVA table raises concerns about "thinness 5-9 years," with a p-value of 0.6, contradicting its high correlation.

Identifying issues with predictors through the ANOVA test prompted a subsequent examination using the Variance Inflation Factor (VIF) test. The results revealed generally acceptable levels of multicollinearity (VIF < 5) for most predictor variables, indicating their reliable inclusion in the regression model. However, 'Percentage Expenditure,' 'GDP,' 'Thinness.1.19. Years,' and 'Thinness.59'. Years exhibited moderate to high multicollinearity, requiring further investigation and consideration for potential impacts on coefficient stability and interpretability. Options such as variable selection or data transformation were suggested to address these multicollinearity issues.

To further investigate the observed multicollinearity between "thinness.1.19. years" and "thinness.5-9 years," a specific model, colinear_model, was employed. The obtained p-value of 2e-16 provides robust evidence against the null hypothesis, affirming the presence of collinearity.

Residual Plots: The spread of the residuals increases as the fitted value gets larger, providing evidence against the equal variance assumption.

Normal QQ plot: The values at the left and right ends deviate significantly from the linear line, suggesting a violation of the normality assumptions.

VIF Reduced Model: (Graph 2.1 2.2 2.3)

Building on these insights, a VIF-reduced model was fitted, excluding 'Thinness.5.9. Years' and retaining other predictors. The VIF test results indicated a reasonably low level of collinearity among these variables, with values ranging from 1.2569 to 2.6625.

The VIF-reduced model exhibits promising performance, as indicated by key metrics. The residual standard error stands at 3.805 on 792 degrees of freedom, suggesting an average prediction error of this magnitude. The Train Set Mean Squared Error (MSE) is 14.33215, and the R-squared value is 0.7918, signifying that approximately 79% of the variability in life expectancy is explained by the chosen predictors, like the previous model. The F-statistic of 430.2 with a p-value less than 2.2e-16 reinforces the model's overall significance.

However, the studentized Breusch-Pagan test detected heteroscedasticity, suggesting the presence of non-constant variance in the residuals, and the Shapiro-Wilk normality test yielded a low p-value of 2.334e-11, indicating a departure from normality in the residuals.

To enhance the model's performance on residual tests, the Box-Cox transformation method was employed. This technique involves applying a power transformation to the response variable, aiming to stabilize variance and make the residuals more normally distributed. The optimal lambda value for the transformation was determined to be 1.3.

Residual Plots: The spread of the residuals increases as the fitted value gets larger, providing evidence against the equal variance assumption.

Normal QQ plot: The values at the left and right ends deviate significantly from the linear line, suggesting a violation of the normality assumptions.

Y-Transformed Model (with 1.3): (Graph 3.1 3.2 3.3 3.4)

The model, which transforms life expectancy through a power shift, resulted in a residual standard error of 13.4 on 790 degrees of freedom, indicating the average prediction error. The Multiple R-squared of 0.7915 shows no significant improvement. The F-statistic is 333.2, emphasizing the overall significance of the model. However, the studentized Breusch-Pagan test detects heteroscedasticity, suggesting non-constant variance in residuals. The Shapiro-Wilk normality test reveals a departure from normality. Overall, the residual tests did not exhibit the expected improvement on the train set.

However, the high MSE and negative R-squared on the test set indicate that the model is not performing well when applied to new, unseen data. This aligns with concerns about potential overfitting, where the model may have learned the training data too well but fails to generalize to different samples. Further refinement of the model, including revisiting variable selection and quadratic forms, is recommended to improve its predictive ability on new observations.

Residual Plots: The spread of the residuals increases as the fitted value gets larger, providing evidence against the equal variance assumption.

Normal QQ plot: The values at the left and right ends deviate significantly from the linear line, suggesting a violation of the normality assumptions.

From the catastrophic result, it's evident that including a 1.3 shift is not a good idea. Here comes our next model, involving the multiplication of categorical variables.

Model with Interaction Terms: (Graph 4.1 4.2 4.3)

The multiple linear regression model incorporates interaction terms between various predictors and the 'Status' variable, designed to capture potential differential effects based on development status. The model exhibits a good fit, with a Residual Standard Error of 3.7 on 785 degrees of freedom, signifying the average prediction error. The Multiple R-squared of 0.8049 and Adjusted R-squared of 0.8014 underscore the model's ability to account for approximately 80% of the variability in life expectancy. The F-statistic of 231.3, accompanied by a very low p-value, signifies the overall significance of the model. However, the studentized Breusch-Pagan test identifies heteroscedasticity, indicating non-constant variance in residuals, and the Shapiro-Wilk normality test reveals a departure from normality. The Test Set MSE of 12.681 and Test R-squared of 0.8167 suggest reasonably good generalization to the test set, warranting further investigation into the warning and potential model refinement.

Attention should still be given to potential heteroscedasticity, as suggested by the studentized Breusch-Pagan test (BP = 100.11, p = 8.502e-14), and a departure from normality, as indicated by the Shapiro-Wilk normality test (W = 0.9826, p = 3.804e-08).

Residual Plots: The spread of the residuals increases as the fitted value gets larger, providing evidence against the equal variance assumption.

Normal QQ plot: The values at the left and right ends deviate significantly from the linear line, suggesting a violation of the normality assumptions.

ANOVA Table for the Interaction Model:

The multiple linear regression model reveals few significant interactive predictors. From the ANOVA test, the interaction terms 'Schooling * Status' and 'Thinness in 1-19 Years * Status' are highly significant (p < 0.001).

Schooling * Status:

The significant interaction between 'Schooling' and 'Status' suggests that the relationship between years of schooling and life expectancy is not uniform across developed and developing countries. This interaction may indicate differing educational impacts on life expectancy in these two contexts.

Thinness in 1-19 Years * Status:

The observed significance of the interaction term 'Thinness in 1-19 Years * Status' indicates that the influence of thinness in the 1-19 years age group on life expectancy is dependent on the country's development status. This could imply varying health implications of thinness across developed and developing nations.

To unveil nuanced patterns within the data, we introduced quadratic forms into our model, aiming to capture potential nonlinear relationships and enhance the model's performance in addressing residual evaluation concerns.

Quadratic Model (Predictors with High ANOVA): (Graph 5.1 5.2 5.3)

The regression model exhibits strong overall performance, as indicated by a low residual standard error of 3.265 and a high Multiple R-squared of 0.8486. The Adjusted R-squared, accounting for the number of predictors, remains robust at 0.8453. The F-statistic of 257.9 with a p-value less than 2.2e-16 signifies a significant fit of the model to the data. The studentized Breusch-Pagan test detects no evidence of heteroscedasticity (BP = 100.11, df = 17, $p < 8.502e-14$). The Shapiro-Wilk normality test suggests some departure from normality in the residuals (W = 0.9826, p = 3.804e-08). The training set Mean Squared Error (MSE) is 10.4185, indicating good model fit. The test set also performs well, with an MSE of 10.2184 and an R-squared of 0.8523, suggesting the model's generalizability to new data.

Residual Plots: The spread of the residuals increases as the fitted value gets larger, providing evidence against the equal variance assumption.

Normal QQ plot: The values at the left and right ends deviate significantly from the linear line, suggesting a violation of the normality assumptions.

In our pursuit of refining the predictive accuracy of our model, we employed various techniques on our quadratic model. This exploration involved implementing backward selection using AIC and BIC criteria, ridge regression with cross-validation, boosting, and bagging. Given the limitations of some models in supporting BP and SW tests, we opted for alternative evaluation metrics, specifically MSE, R-squared, and a visual representation through observed versus predicted plots. It's important to note that all these assessments were conducted on the test set, ensuring an evaluation of our model's generalization performance.

Quadratic Model & AIC and BIC Backward Selection: (Graph 6.1 6.2 6.3 6.4 6.5)

In the process of backward selection using both AIC and BIC criteria, the AIC-selected model comprises 14 predictors, while the BIC-selected model includes 11 predictors. The AIC model incorporates quadratic terms, interactions, and various predictors such as 'Income Composition of Resources,' 'Thinness in 1-19 Years,' and 'HIV.AIDS^2.' The model performs well, as evidenced by a residual standard error of 3.263, an R-squared of 0.848, and an F-statistic of 337.4, indicating a significant fit to the data. However, the Breusch-Pagan test suggests potential heteroscedasticity (BP = 95.688, $p < 2.2e-16$), and the Shapiro-Wilk normality test indicates some departure from normality (W = 0.98167, p = 1.818e-08). The training set Mean Squared Error (MSE) is 10.46072, and the R-squared is 0.8480226, indicating satisfactory model performance. The BIC-selected model with 11 predictors shows a similar pattern, and both models provide valuable insights into predictors' impact on life expectancy.

Residual Plots: In both AIC and BIC graphs, the spread of the residuals increases as the fitted value gets larger, providing evidence against the equal variance assumption.

Normal QQ plot: In both AIC and BIC graphs, the values at the left and right ends deviate significantly from the linear line, suggesting a violation of the normality assumptions.

Based on the results presented above, it appears that linear regression may not be the most suitable model for analysis. To gain a deeper understanding of the data, we have opted to explore alternative models.

Ridge Regression Model with Best Lambda from Cross Validation: (Graph 7.1 7.2)

To enhance the precision of predicting life expectancy, we employed a regularization process that yielded a sparse matrix of coefficients. The intercept stands at approximately 64.57, with various predictors showing positive and negative coefficients. Notably, 'Income Composition of Resources,' 'Thinness 1-19 Years,' and the interaction term 'Thinness in 1-19 Years * Status' hold substantial coefficients, indicating their significant influence. The regularization process likely contributed to variable selection and shrinkage, underscoring the importance of certain features while potentially reducing overfitting. These coefficients offer insights into the relative importance of predictors within the chosen regularization method, aiding in model interpretation and simplification.

In evaluating the model's performance on the test set, the Mean Squared Error (MSE) of 10.23504 reflects the average squared difference between observed life expectancy values and predictions. A lower MSE is desirable for improved accuracy, and the Test Set R-squared value of 0.8521 indicates that the model explains a substantial portion of the variability in life expectancy. Together, these metrics affirm the model's efficacy in accurate predictions and capturing underlying patterns in the data.

Boosting & Generalized Boosted Regression Modeling (GBM): (Graph 8.1)

The boosting model (GBM) with 100 trees demonstrates strong predictive accuracy and pattern capture in the dataset, particularly compared to previous models. The Mean Squared Error (MSE) of 5.389733 signifies the average squared difference between observed and predicted values, with a lower MSE indicating improved accuracy. Additionally, the Boosting R-squared value of 0.9221 denotes the proportion of variance in the dependent variable effectively explained by the boosting model. This high R-squared value affirms the model's capability in accounting for a significant portion of the data's variability.

Bagging & Random Forest Model: (Graph 9.1)

Opting for the random forest model with 50 trees based on summary results and graphical analysis, we found it to outperform previous models, showcasing the strongest performance. The low Mean Squared Residuals (MSE) at 4.7295 indicates minimal squared differences between observed and predicted values on average. Moreover, the model explains a high percentage (93.13%) of the variance in the dependent variable, demonstrating its efficacy in capturing a substantial portion of the data's variability. The Bagging R-squared value of 0.9407 provides additional support for the model's ability to explain variance, while the low Bagging MSE of 4.1055 underscores its predictive accuracy. In summary, these metrics collectively affirm that the Random Forest model is effective in accurate predictions and adept at capturing the complex relationships within the dataset. Moreover, the visualization of OOB over number of trees shows that OOB decreased to a consistent value when we use more than 25 trees.

Discussion of the finding

Conclusion

The comprehensive analysis of multiple linear regression models, spanning from the full model to advanced techniques such as ridge regression, boosting, and random forests, unveiled intricate patterns in predicting life expectancy. Even though initial models exhibited high R-squared values, concerns related to heteroscedasticity and non-normality prompted a closer examination of variable selection and potential transformations. The exploration of VIF-reduced models provided insights into mitigating multicollinearity, and the introduction of interaction terms illuminated nuanced predictor effects, particularly concerning a country's development status. Although quadratic models improved the fit, residual issues endured. Ultimately, the bagging algorithm has been proved to deliver the best performance of our model.

Highlights

The importance of 'Income Composition of Resources', 'Schooling', and interaction terms influencing life expectancy was consistently emphasized across various models. Variable selection and shrinkage techniques, seen in ridge regression, provided valuable insights into predictor importance and potential overfitting mitigation. Random forest modeling emerged as the most effective, explaining 93.13% of variance and showcasing strong predictive accuracy.

Expected Results

Expected results included the identification of influential predictors, model refinement through variable selection and transformations, and improved predictive accuracy with advanced modeling techniques. The progression from linear regression to advanced models is expected to address inherent challenges and enhance model robustness.

Unexpected Results

Unexpectedly, despite advancements in model complexity, some challenges persisted, including heteroscedasticity and non-normality in residuals. From the articles we read, we expected BMI to be an influential predictor, but both BMI and BMI square were dropped in backward selection.

Learnings from the Analysis

In the initial stages of our research, our expectations, based on a preliminary understanding of the data, did not align with the actual findings as we delved deeper. When faced with situations where the data seemed to indicate something clearly, further investigation often revealed a more nuanced perspective.

To enhance the clarity and effectiveness of our work, we employed a method known as backward selection. This method systematically helped us eliminate less important factors one by one and allowed for a thorough comparison of different ideas.

In addition to linear regression, if the model performance did not achieve the best fit, we explored further additional regression methods. To better analyze the data, other than the investing variables, we explored the application for this dataset with bagging and boosting methods.

Association Between the Response and Predictor

The predictors for the response 'Life Expectancy' have many factors. The factors such as 'Income Composition of Resources' and 'Schooling,' consistently proved influential in various models. Considering a country's development added another layer to these relationships. In addition to the linear connections, we also explored non-linear connections between factors and life expectancy using quadratic terms and interactions. To improve and simplify our model, we utilized a method called backward selection (AIC & BIC). This allowed us to simply our model without reducing the accuracy.

Limitations

Despite the progress made in constructing and improving models, several limitations and potential challenges remain.

Handling Missing Data:

One limitation arises from the presence of numerous empty cells in several columns, such as GDP and Population. While these missing values were filled with median values to maintain data integrity, this approach may have limitations. The imputed values might become less representative, potentially leading to an underestimation of correlations. Additionally, the removal of rows during correlation checks may have further impacted the overall analysis.

Residual Tests:

The constructed models, while showing improved Mean Squared Error (MSE) and R-squared values on test sets, faced challenges in passing residual tests. Residual linearity, QQ plot, and Breusch-Pagan (BP) and Shapiro-Wilk (SW) tests revealed discrepancies. The source of these issues might stem from the initial variable selection process or the need for a more refined approach to identifying the best quadratic forms.

Further Questions

Overfitting Concerns:

The high R-squared value observed on the test set (94%) prompts concerns about potential overfitting. Further investigation into the model's generalization to the population is essential. This may involve assessing model performance on diverse datasets or considering regularization techniques to prevent overfitting.

Variable Selection Strategies:

An ongoing question involves the exploration of more effective variable selection strategies. The initial selection process might benefit from refinement, through the incorporation of advanced feature selection methods or algorithms designed to handle multicollinearity.
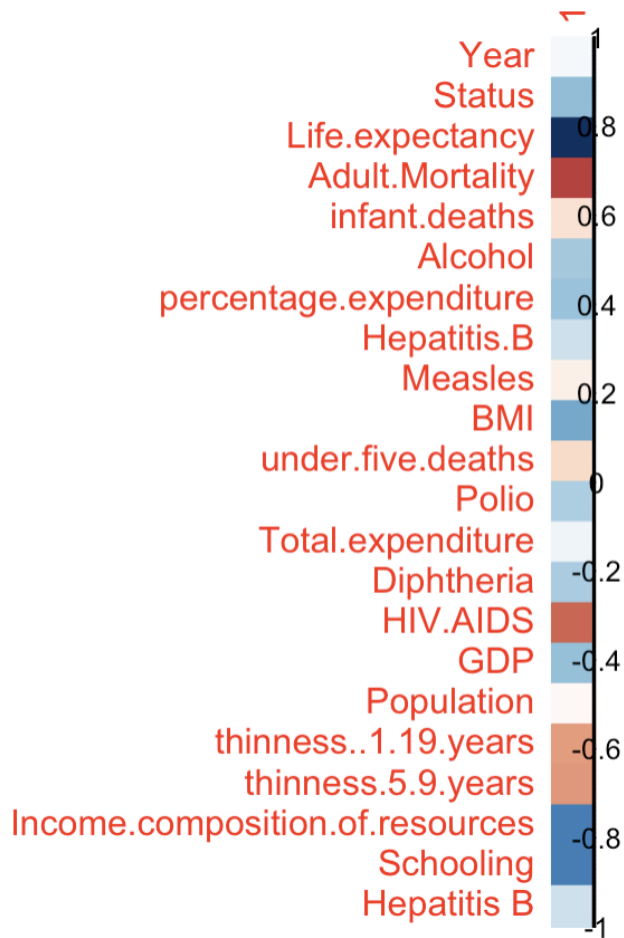
Appendix

Related Articles:

https://www.sciencedirect.com/science/article/pii/S2772442522000077#d1e1440

(Rohan et al, 2022)

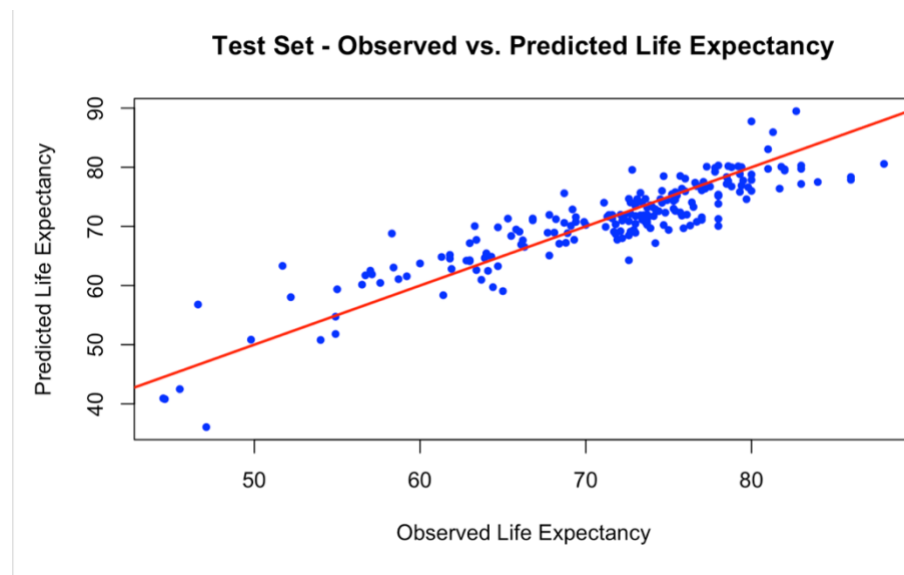https://www.researchgate.net/publication/356235215_Life_Expectancy_Prediction_Analysis_using_ML

(Bali et al, 2021)
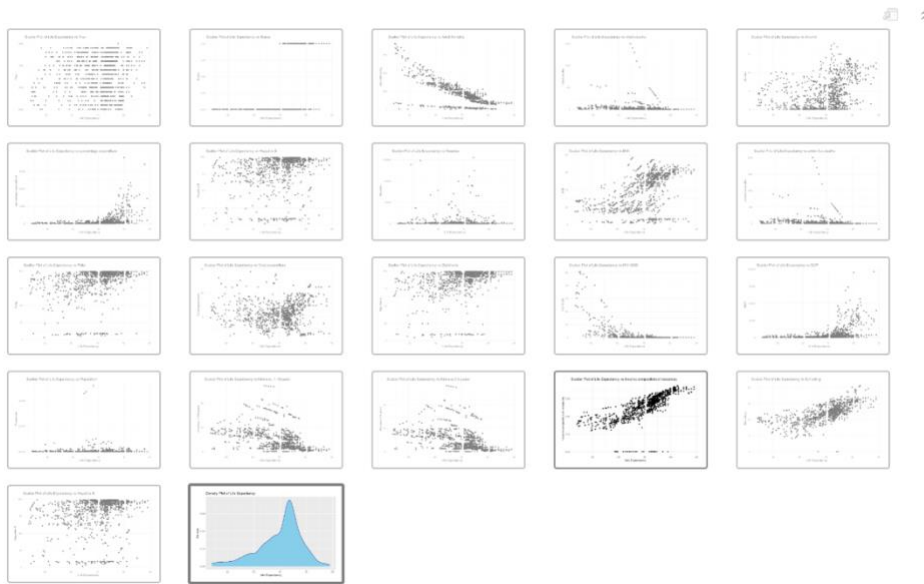
Graph 0.1

Graph 0.2

Graph 1.1
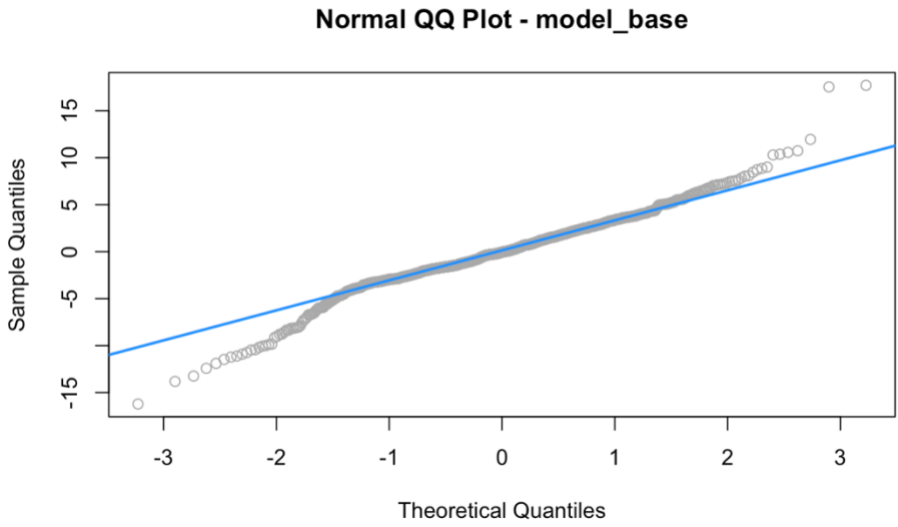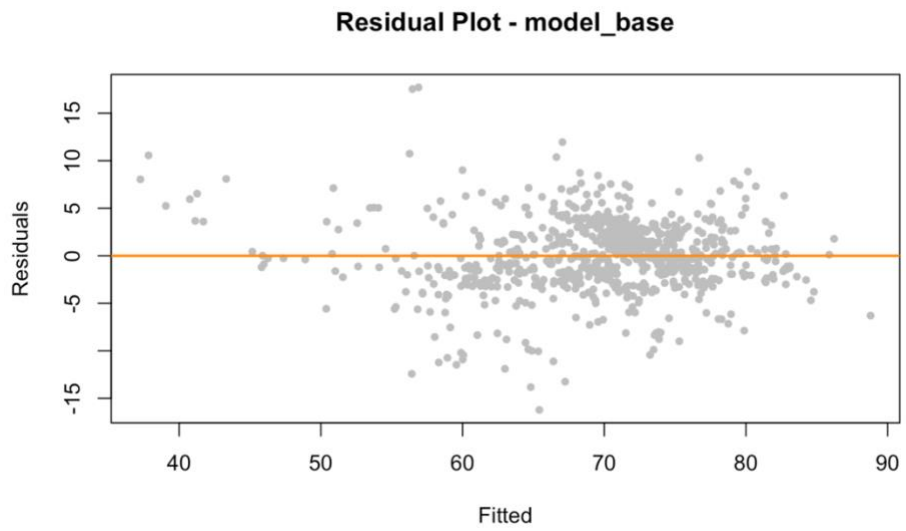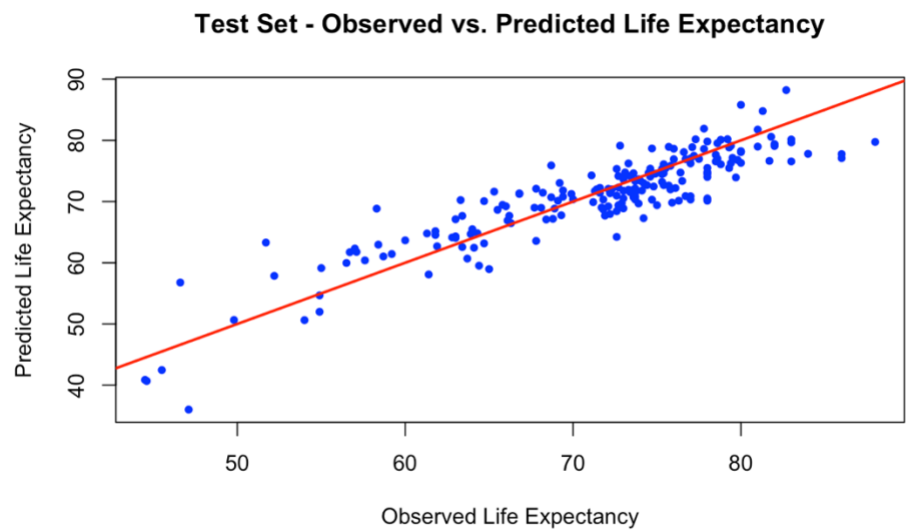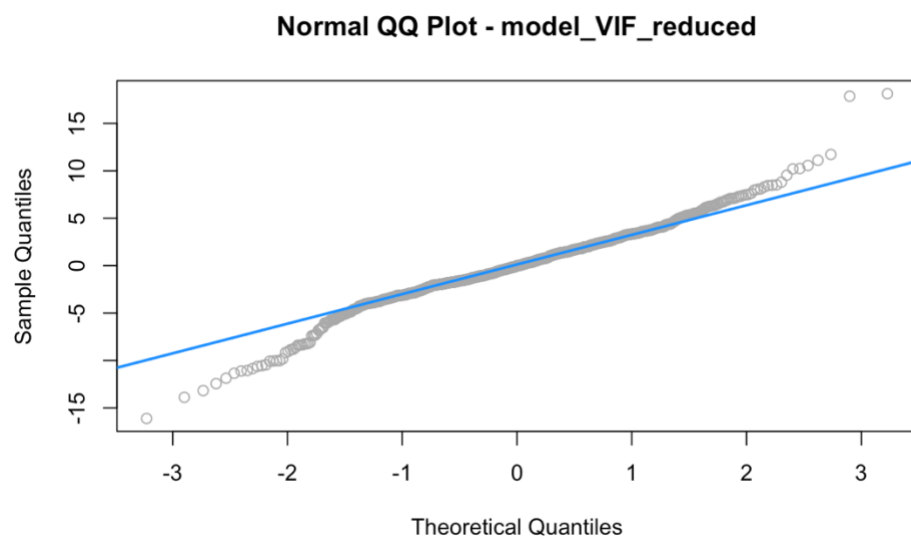
Normal QQ Plot - model_base
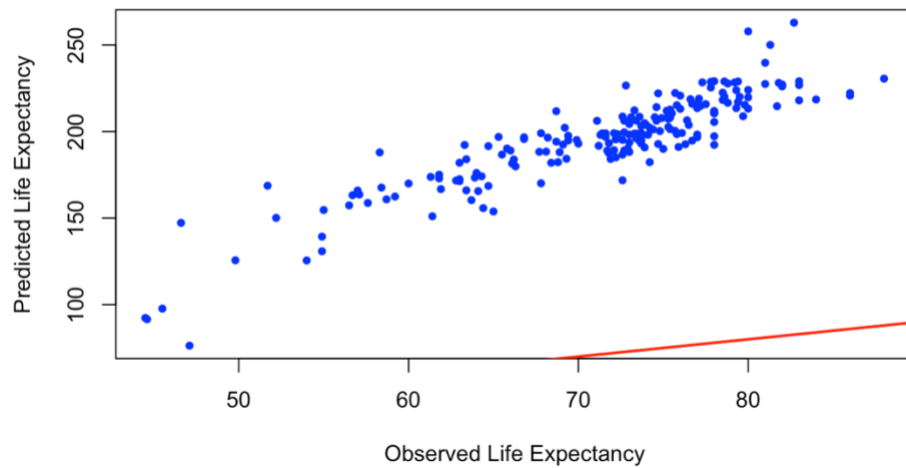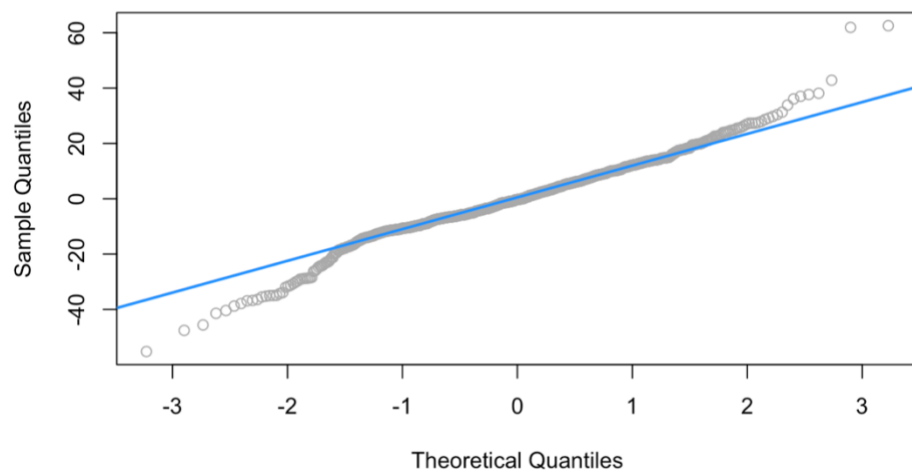

Graph 1.2

Residual Plot - model_base

Graph 1.3

Test Set - Observed vs. Predicted Life Expectancy

Graph 2.1

Normal QQ Plot - model_VIF_reduced

Graph 2.2

Residual Plot - model_VIF_reduced
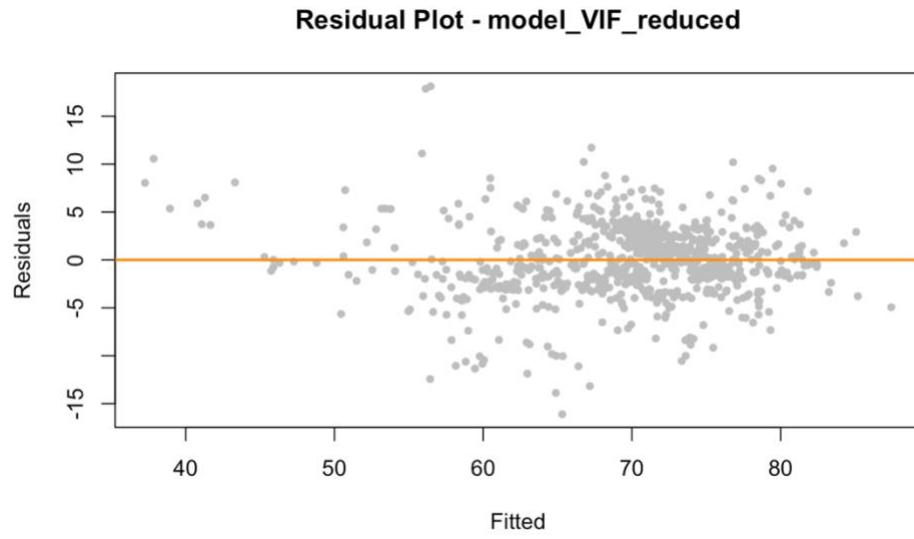
Graph 2.3
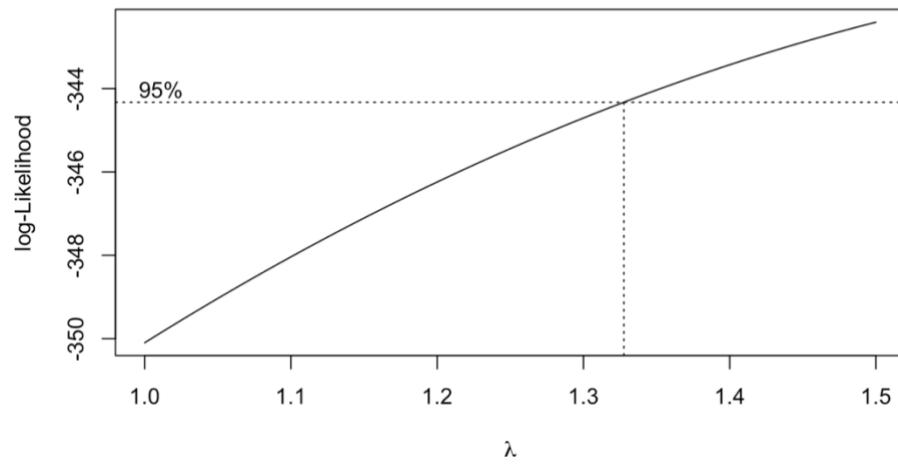
3.1

Test Set - Observed vs. Predicted Life Expectancy


3.2

Normal QQ Plot - model_y_shift

3.3

Residual Plot - model_y_shift

3.4

**Test Set - Observed vs. Predicted Life Expectancy**
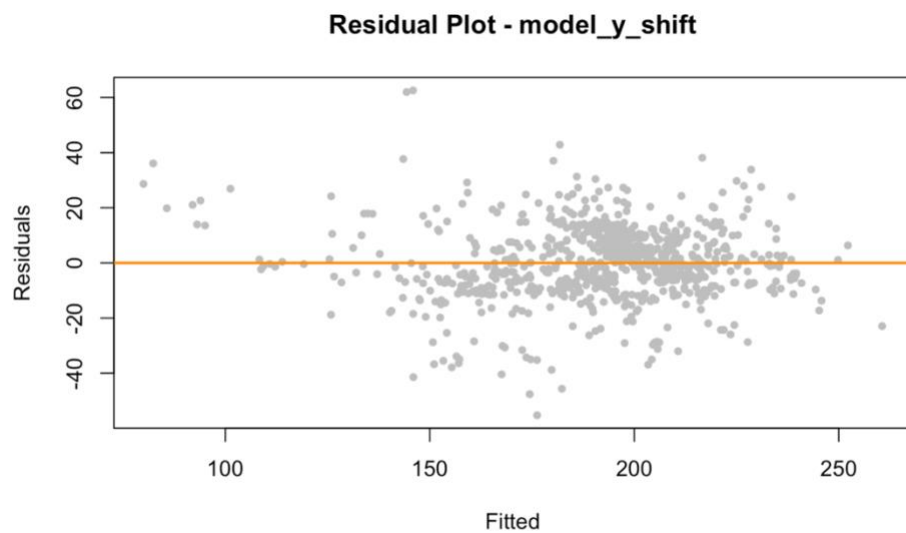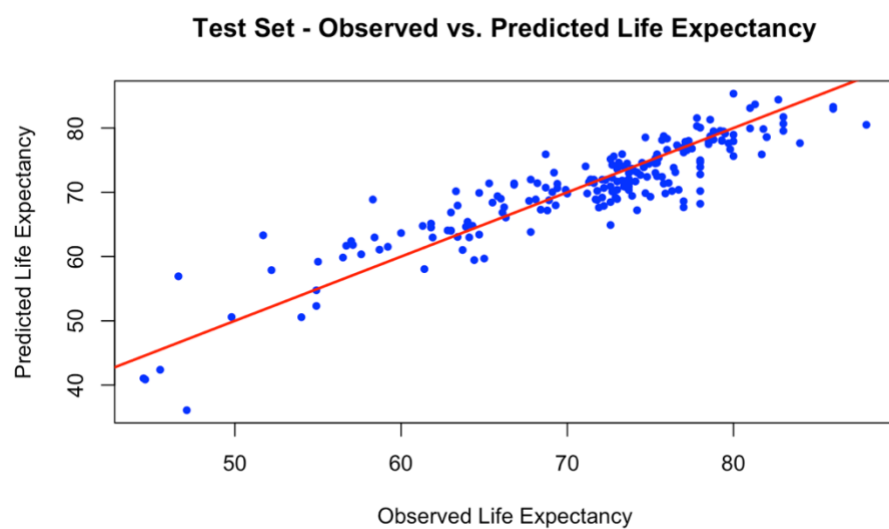
4.1

Normal QQ Plot - model_categorial
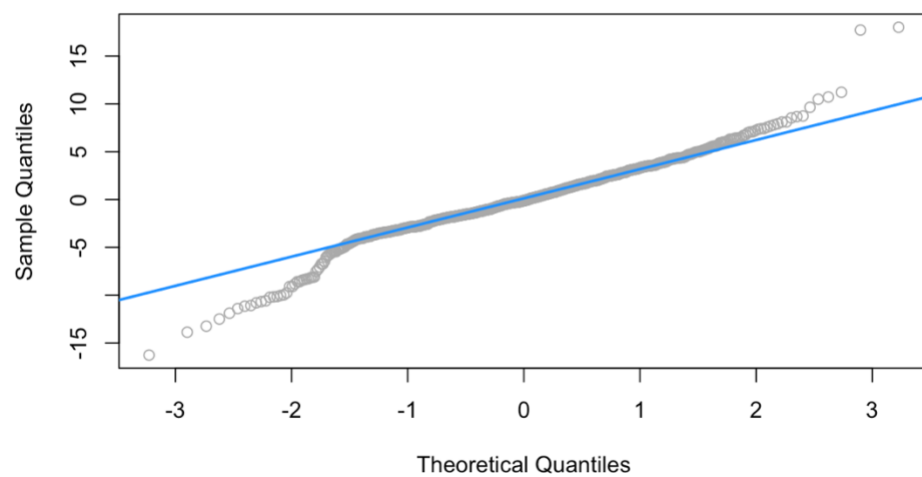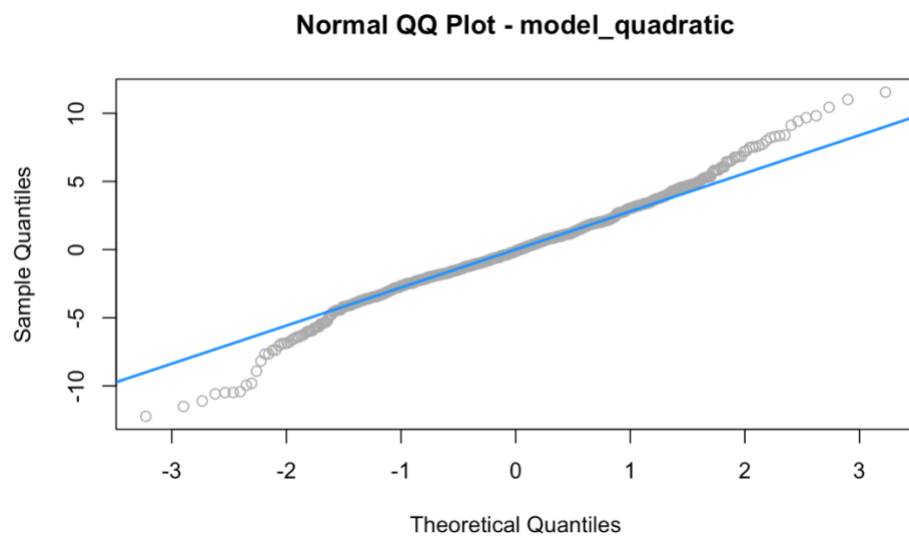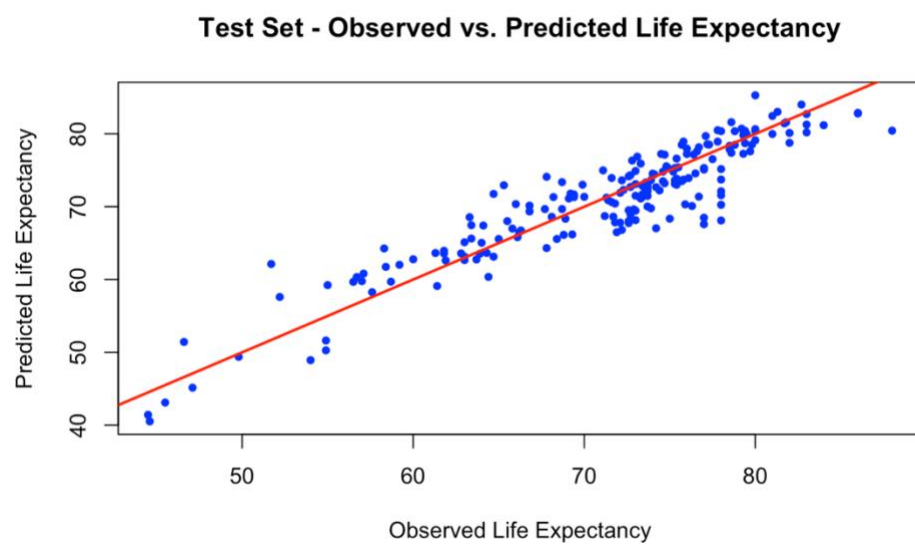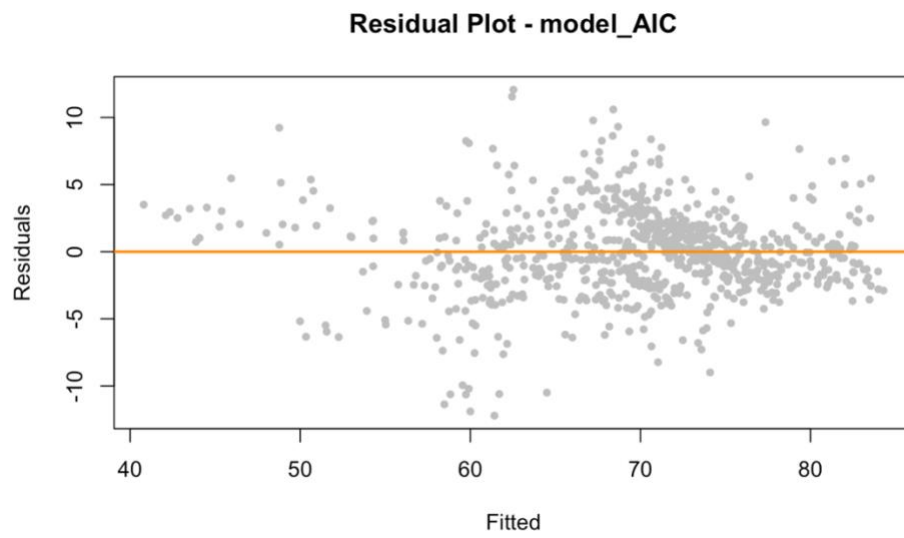

4.2

Residual Plot - model_categorial

4.3

Test Set - Observed vs. Predicted Life Expectancy

5.1

Residual Plot - model_quadratic

5.2

Normal QQ Plot - model_quadratic

5.3

Test Set - Observed vs. Predicted Life Expectancy

6.1

Residual Plot - model_AIC

6.2

Normal QQ Plot - model_AIC

6.3

Residual Plot - model_BIC

6.4

Normal QQ Plot - model_BIC
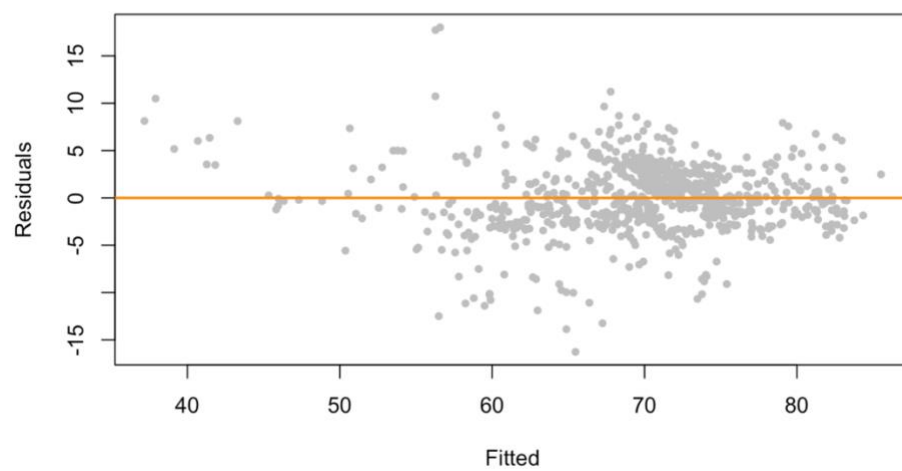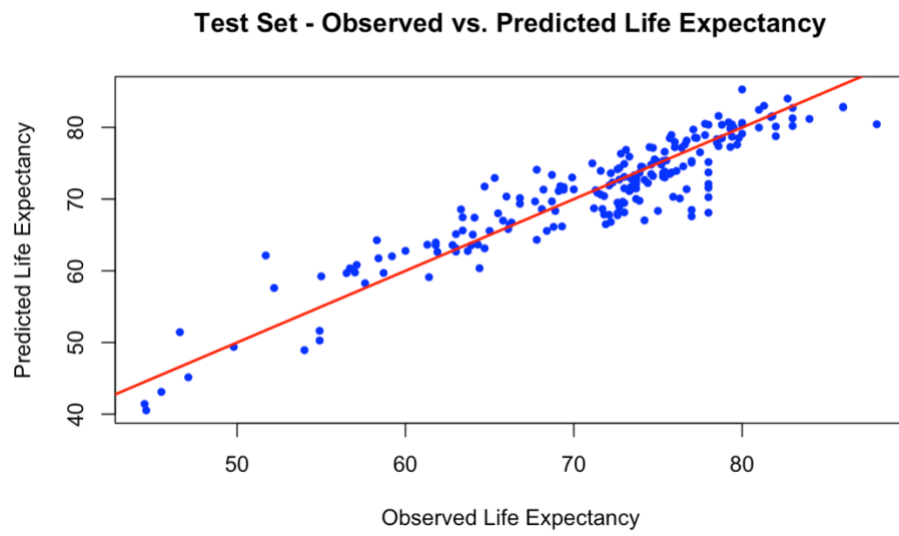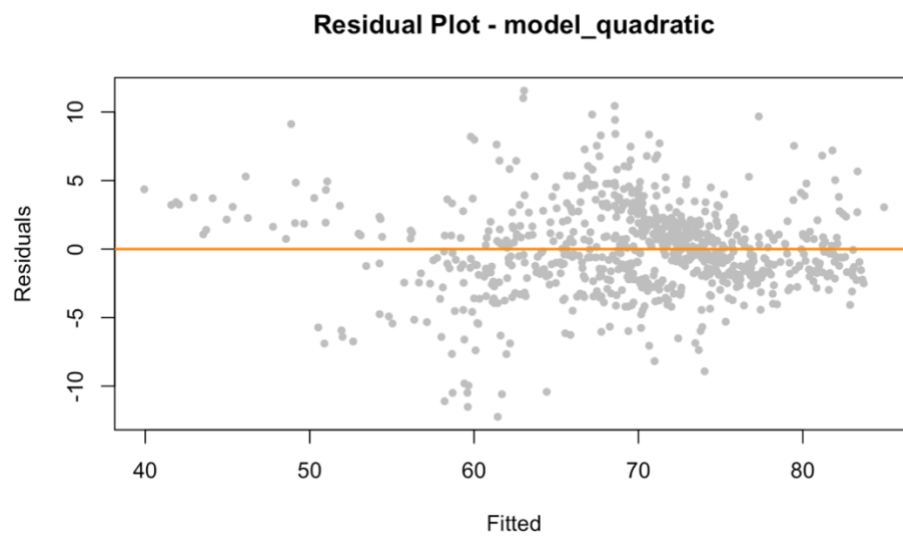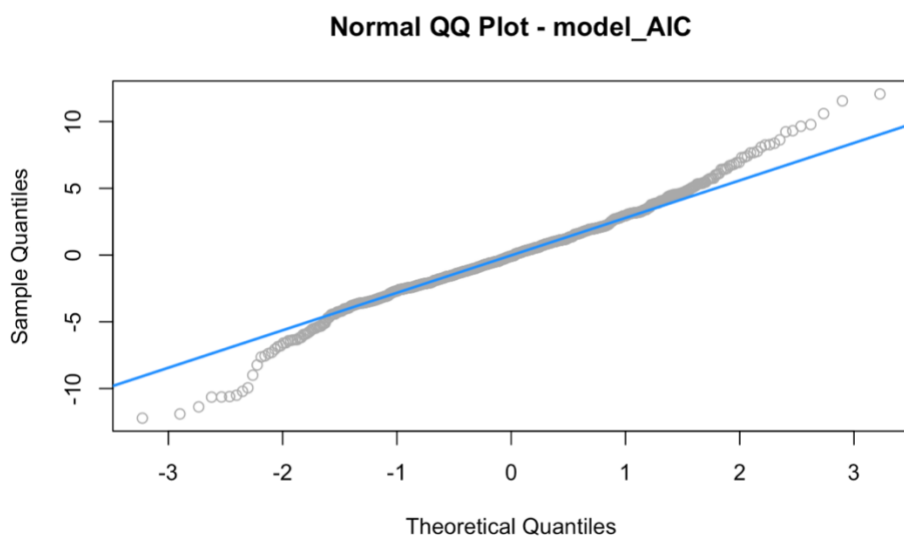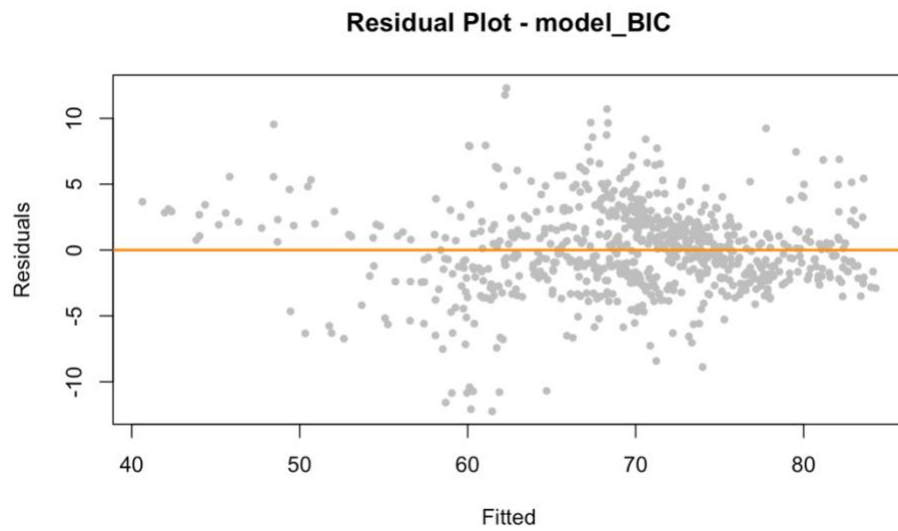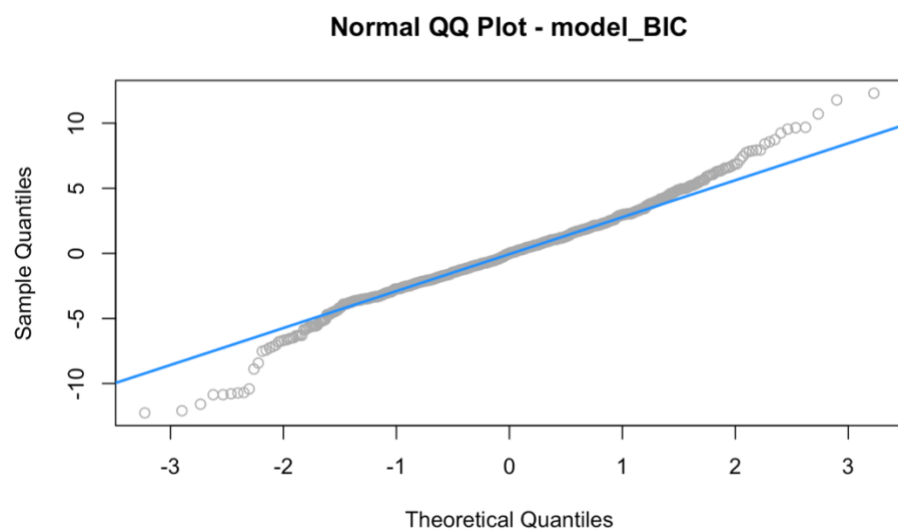

6.5

```
Best Lambda: 0.01
18 x 1 sparse Matrix of class "dgCMatrix"
                                                  s1
(Intercept)                              6.456726e+01
Status                                   6.732716e+00
Adult.Mortality                         -9.740832e-03
BMI                                      4.045912e-02
HIV.AIDS                                -8.197396e-01
GDP                                     -1.314455e-05
thinness..1.19.years                    -2.215408e-01
Income.composition.of.resources         -2.550070e+01
Schooling                                7.047164e-01
I(Adult.Mortality^2)                    -6.175090e-06
I(BMI^2)                                -5.029508e-04
I(HIV.AIDS^2)                            1.071608e-02
I(GDP^2)                                 2.396413e-10
I(thinness..1.19.years^2)                7.521940e-03
I(Income.composition.of.resources^2)    4.379009e+01
I(Schooling^2)                          -2.100125e-02
I(thinness..1.19.years * Status)        -1.831525e+00
I(Schooling * Status)                   -2.996159e-01
```
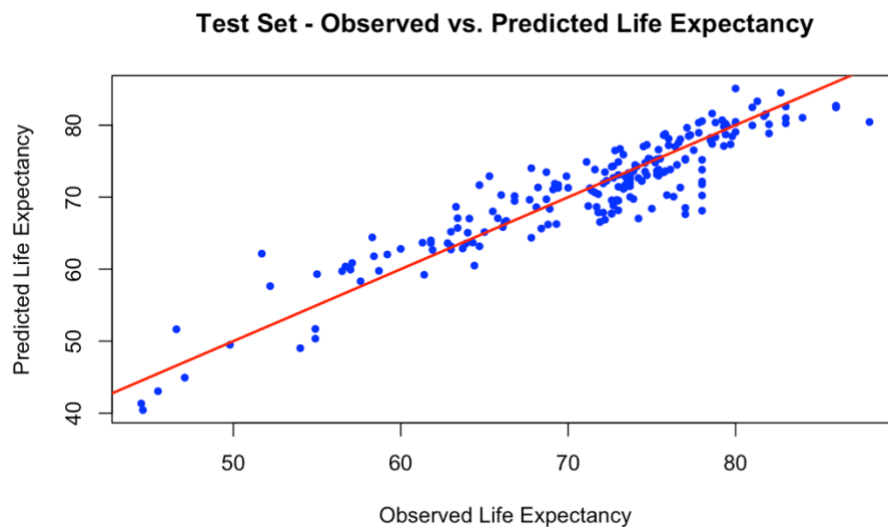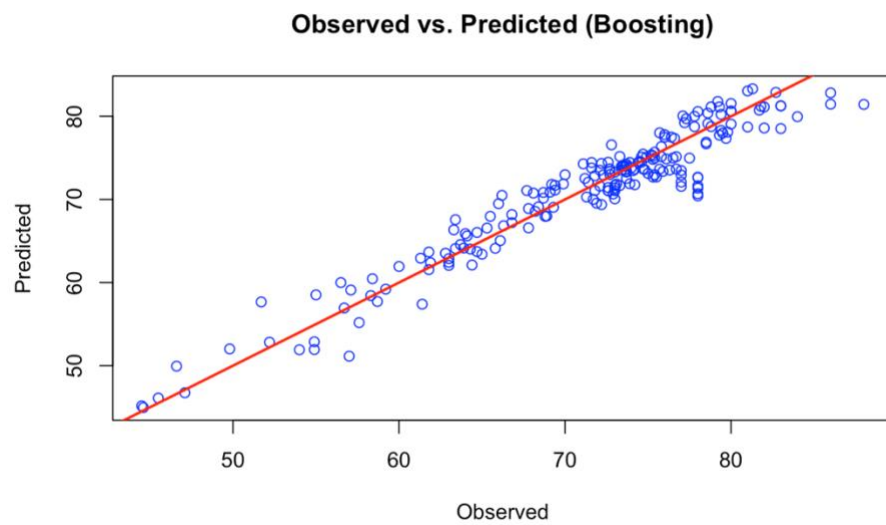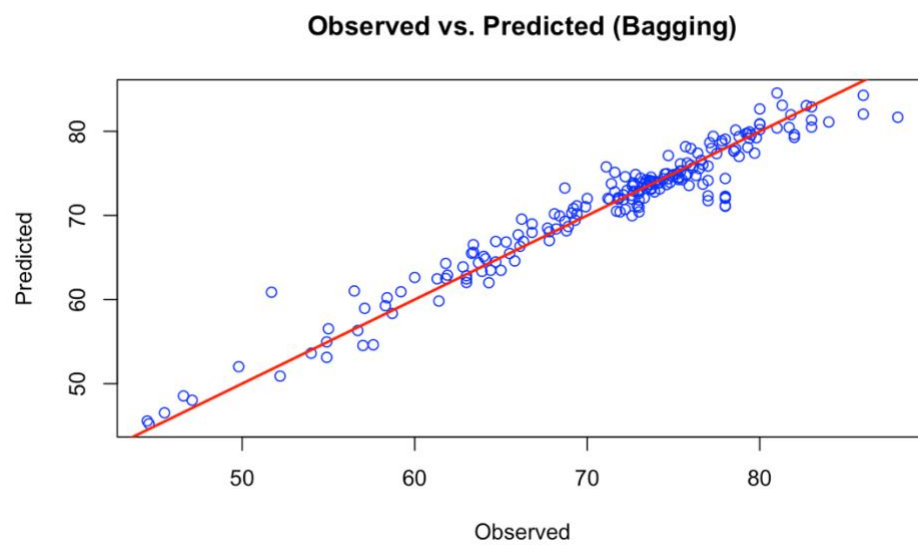
7.1

Test Set - Observed vs. Predicted Life Expectancy

7.2

Observed vs. Predicted (Boosting)

8.1

Observed vs. Predicted (Bagging)

9.1