AMAT 565

APPLIED STATISTICS

PROJECT REPORT

on

**Forecasting Life Expectancy:**

**Predictive Analysis Using Historical Time Series Data**

by

**Team Infinity**

**Abstract:**

This project delves into predicting life expectancy by analyzing historical time series data. Utilizing datasets from diverse countries, the study employed Auto ARIMA, Random Forest Regression, and Linear Regression models. The analysis focused on variables such as population size, region, and income group to ascertain their influence on life expectancy prediction. The findings offer insights into the potential predictability of life expectancy based on these key factors.

**Introduction to the Problem Statement:**

In the quest to understand and anticipate life expectancy trends, this project delves into the realm of predictive analytics, leveraging historical time series data. By scrutinizing diverse datasets encompassing population statistics and life expectancy records, this study aims to unravel the intricate relationships between socio-demographic factors and longevity.

Life expectancy, a pivotal measure reflecting the overall health and well-being of populations, is influenced by multifaceted determinants. Exploring the potential predictability of this vital metric using data-driven methodologies is crucial in informing policy decisions and healthcare interventions.

Through the utilization of various statistical and machine learning models, including Auto ARIMA, Random Forest Regression, and Linear Regression, this study seeks to elucidate the influence of variables such as population size, income group, and geographic regions on forecasting life expectancy trajectories. The analysis aims to discern not only the predictability of life expectancy but also the underlying factors contributing significantly to its fluctuations over time.

Understanding the factors that shape life expectancy can offer valuable insights into societal well-being, public health strategies, and resource allocation. This investigation aims to contribute to the burgeoning field of predictive healthcare analytics, potentially providing actionable insights for policymakers and health practitioners worldwide.

In summary, the question/ problem statement of our project is:

**“Can we predict life expectancy using historical time series data?”**

**Details of the datasets:**

The datasets utilized in this project are sourced from the World Bank's repository of global indicators. The first dataset pertains to population statistics and is available through the link:

[**https://data.worldbank.org/indicator/SP.POP.TOTL**](https://data.worldbank.org/indicator/SP.POP.TOTL)

The population dataset encompasses comprehensive records of total population counts across various countries over multiple years. The second dataset concerns life expectancy and can be accessed via the link:

[**https://data.worldbank.org/indicator/SP.DYN.LE00.IN**](https://data.worldbank.org/indicator/SP.DYN.LE00.IN)

The life expectancy dataset comprises data reflecting life expectancy at birth for different countries, providing insights into changes in life expectancy over time.

Both datasets serve as invaluable resources for understanding demographic trends and patterns across nations. They offer a longitudinal perspective, spanning numerous years and multiple regions, enabling a comprehensive analysis of population dynamics and trends in life expectancy.

Utilizing these datasets, the project aims to leverage historical time series data to investigate and predict life expectancy trends based on diverse variables such as population size, income group, and regional categorizations. The rich and expansive nature of these datasets from the World Bank allows for a robust exploration into the factors influencing life expectancy globally.

**Data Preprocessing and Exploratory Data Analysis:**

The preprocessing of data is a crucial step in predictive modeling, ensuring that the datasets are structured and cleaned for accurate analysis. In this section, we delve into the intricate process of preparing the datasets for predicting life expectancy based on historical time series data.

**a. Loading Datasets:** The initial phase involved loading datasets containing life expectancy and population records across various countries. The read.csv function was utilized to import these datasets into the R environment. Preliminary exploration using head() displayed the initial records to understand the dataset's structure and contents.

**b. Pivoting and Transforming Dataframes:** Both the life expectancy and population datasets required transformation to facilitate analysis. The gather() function was used to pivot the datasets, converting years as columns into key-value pairs. This transformation helped convert the data structure from wide to long format, enabling easier analysis across time periods.

**c. Handling Empty Values:** An essential aspect of data preprocessing involves handling missing or empty values. A systematic approach was employed to address potential missing data issues. Specifically, the na.omit() function was used to drop rows with missing values. Additionally, the empty strings were detected and replaced with NA to ensure proper handling of null values.

**d. Merging Datasets:** The merging process involved integrating the population and life expectancy datasets using common attributes like 'Country.Name', 'Country.Code', and 'Year'. This step was pivotal in creating a unified dataset that contained both population and life expectancy records, facilitating holistic analysis.

**e. Data Quality Checks:** Data quality checks were performed to ensure the efficacy of the preprocessing steps. It calculated and displayed missing values per column before and after handling missing values. The verification ensured that essential data points were not lost during the cleaning process.

**f. Selection of Countries for further analysis:** Post-preprocessing, three countries were selected based on population size. The selection of countries was a crucial step in understanding diverse population dynamics. The code employed specific criteria to identify and isolate these countries based on their population size and data availability across the years. Following are the sequence of steps involved:

1. The initial step involved filtering countries that had comprehensive data spanning from 1960 to 2021. It is essential for the analysis to have continuous records without missing years to ensure robust predictions. For this purpose, the filter function is used in the coding part.

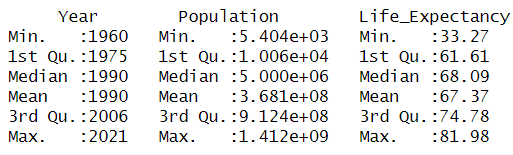
2. **Population-Centric Selection**: The selection strategy was population-based, aiming to include countries representing different population sizes.

3. **Statistical Indicators:** Statistical indicators were computed, such as the largest population, the population closest to the median, and the smallest population. These indicators assisted in categorizing countries based on their population sizes.

4. **Using these categories, three specific countries were filtered and isolated:** China, representing the largest population; Finland, representing a population close to the median; and Tuvalu, representing the smallest population.

Representative sampling and comparative insights were ensured by the strategic selection of China, Finland, and Tuvalu, laying the groundwork for exploring the relationship between population dynamics and life expectancy.

**Summary Statistics of the filtered data:**



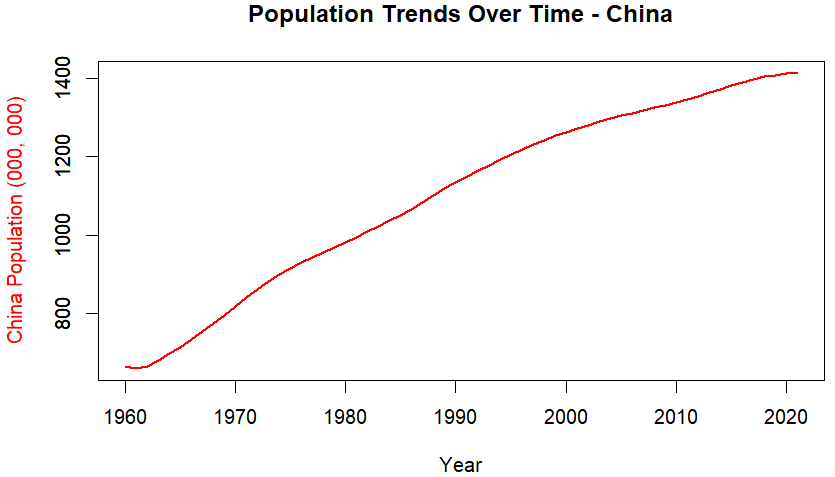
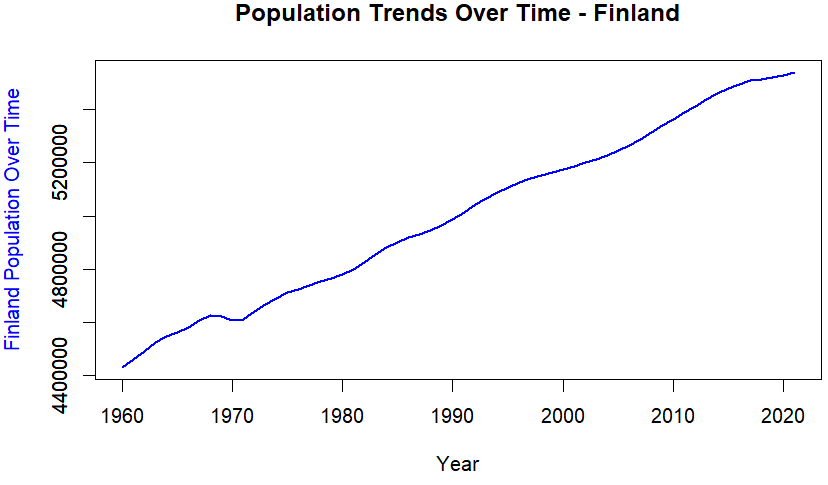
From the above summary statistics, following are the observations:

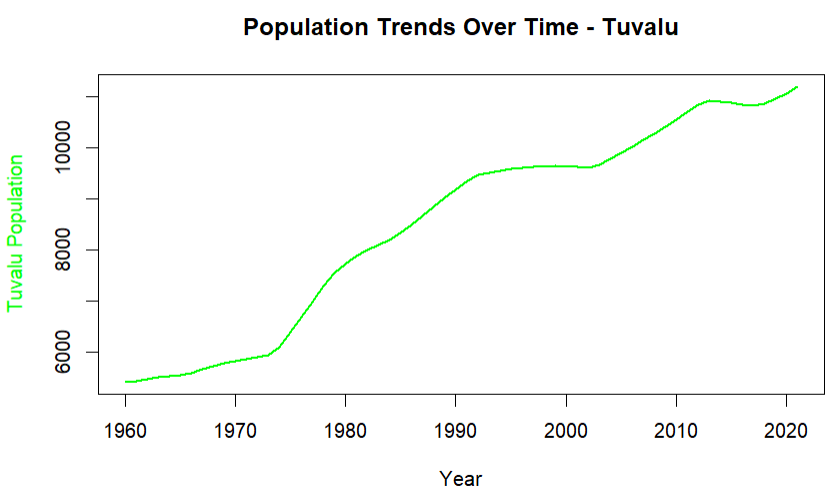
1. The earliest year is 1960, and the latest year is 2021.

2. The smallest population size is around 5,404, the median population is 5 million, and the largest population is 1.1412 billion.

3. The minimum life expectancy is 33.27 years; median life expectancy is 68.09 years, and the maximum life expectancy is 81.98 years.

**Population Trends over Time:**

 .

The analysis of population trends across China, Finland, and Tuvalu from 1960 to 2021 revealed a consistent upward trajectory, displaying significant growth over the examined period, reflecting an increasing population trend.

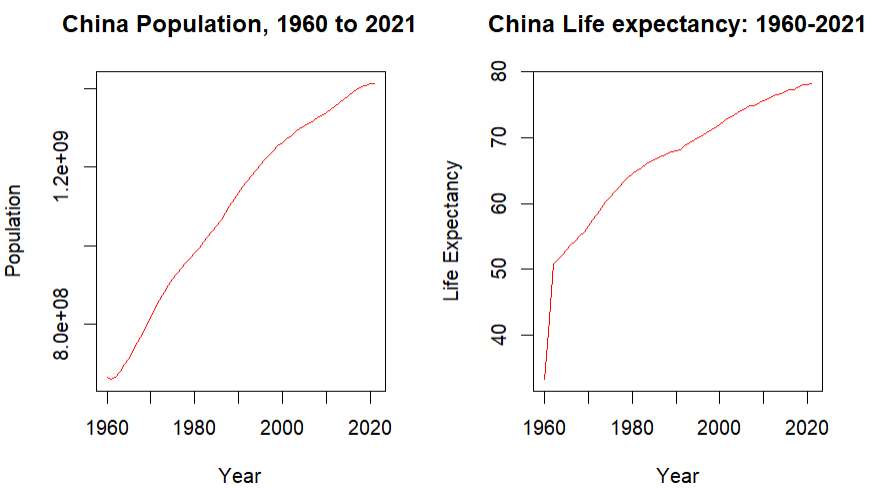
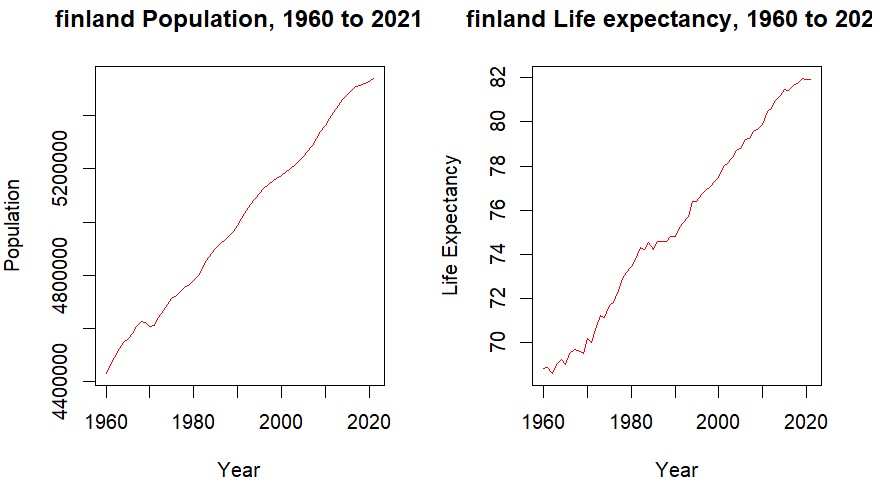
**BUILDING MODELS:**

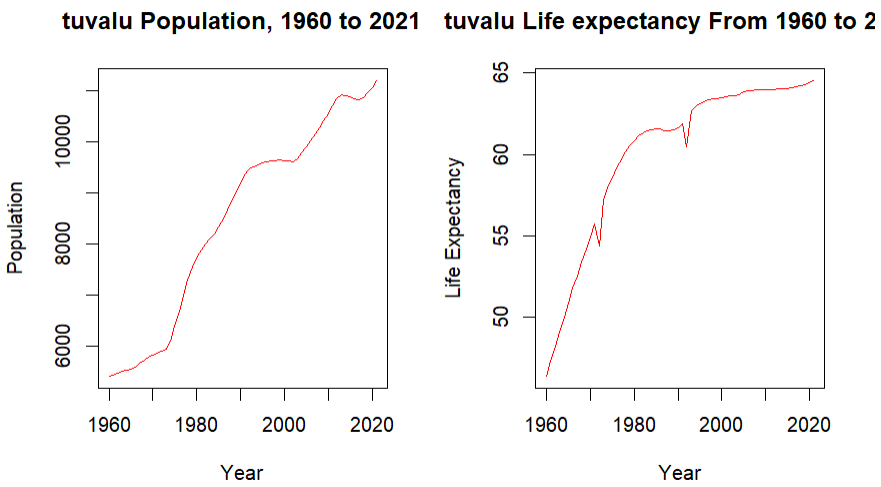
**a. ARIMA Model:**

**Time Series Analysis:**

The Time Series Analysis conducted in this project involved exploring and visualizing population and life expectancy trends across selected countries (China, Finland, and Tuvalu) from 1960 to 2021. It aimed to uncover patterns, trends, and fluctuations in these variables over time. The analysis included line plots illustrating the growth patterns and changes in population and life expectancy, highlighting consistent upward trajectories in population trends and variations in life expectancy rates over the examined period.

**Time Series Plots:**

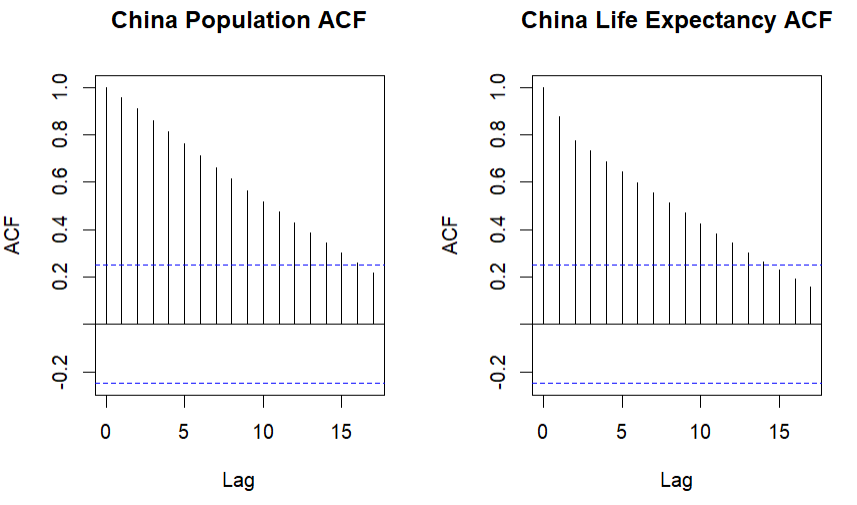
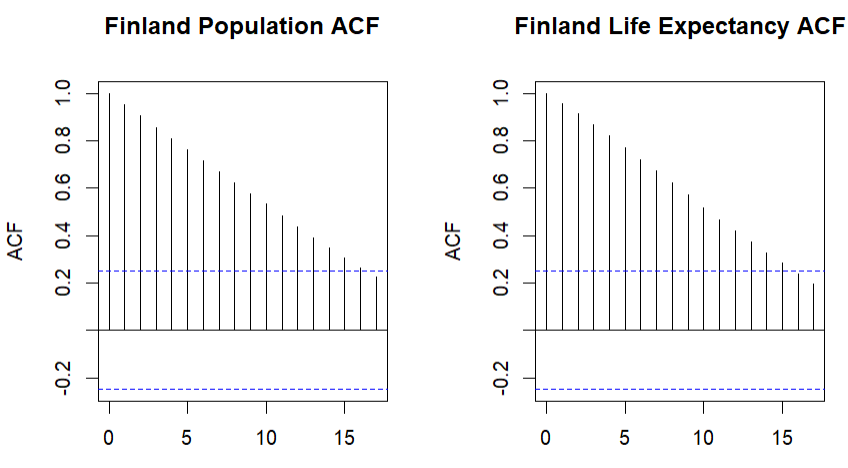


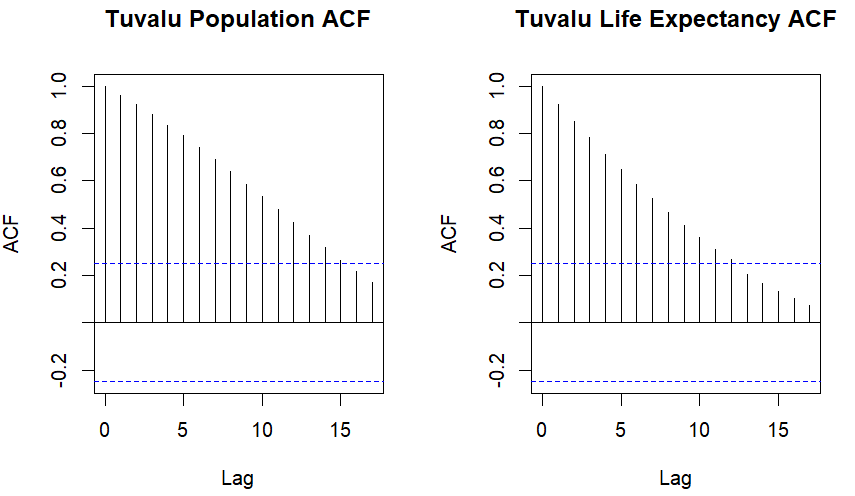
**ARIMA (Auto Regressive Integrated Moving Average) Modeling:**

ARIMA, a robust statistical method for time series forecasting, was applied in this project to model and predict life expectancy based on historical data. The ARIMA modeling process included several steps:

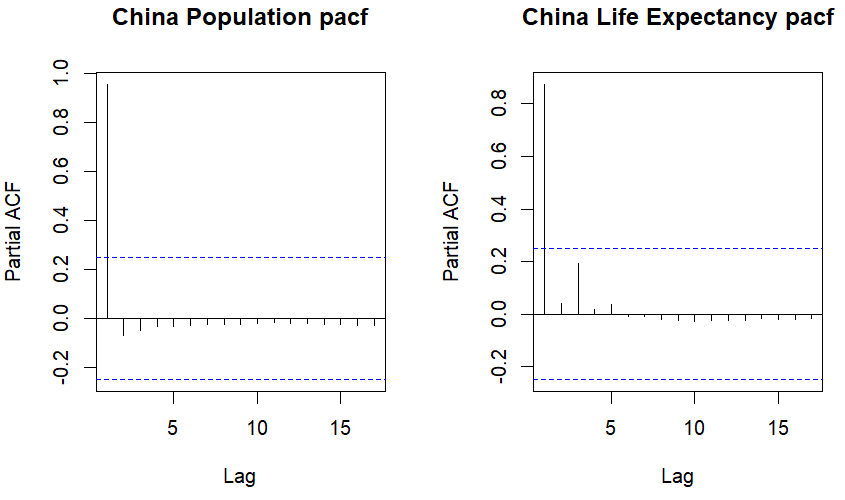
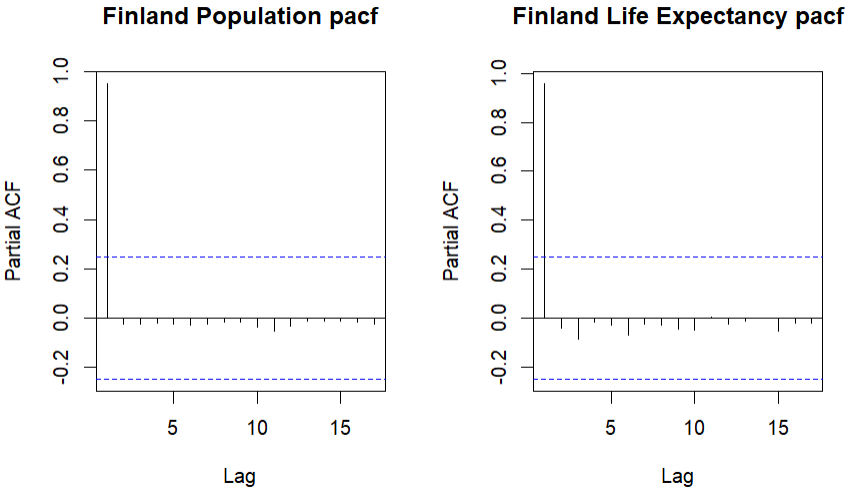
**Model Identification:** Determining the order of differencing (I), autoregressive (AR), and moving average (MA) terms through analysis of ACF (Autocorrelation Function) and PACF (Partial Autocorrelation Function) plots.

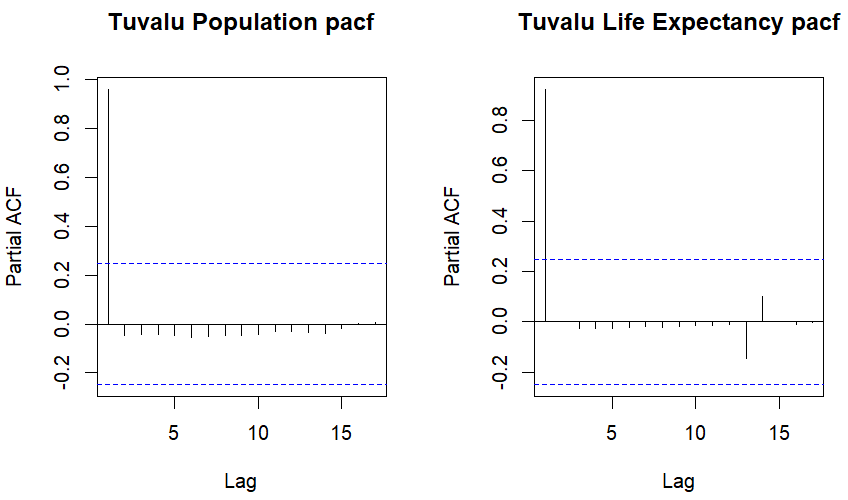
**ACF Plots:**



**PACF Plots:**

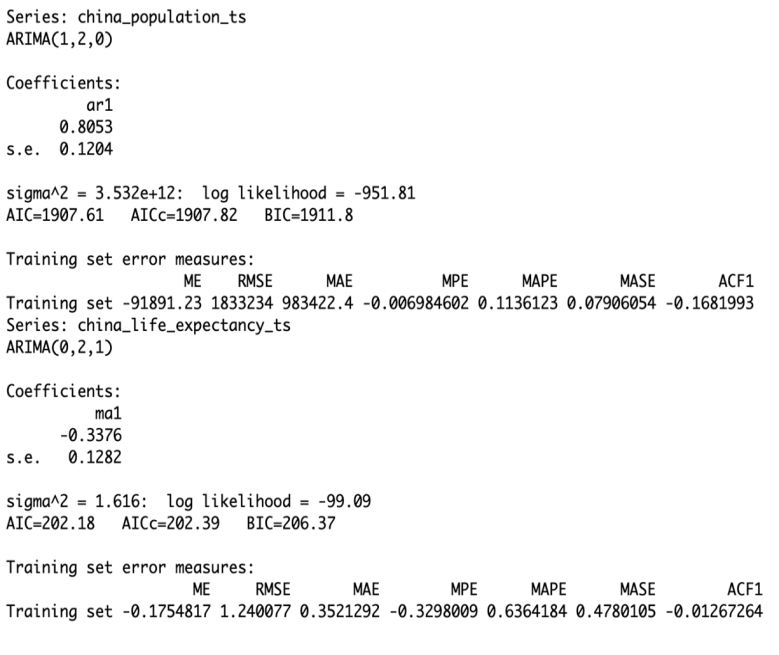
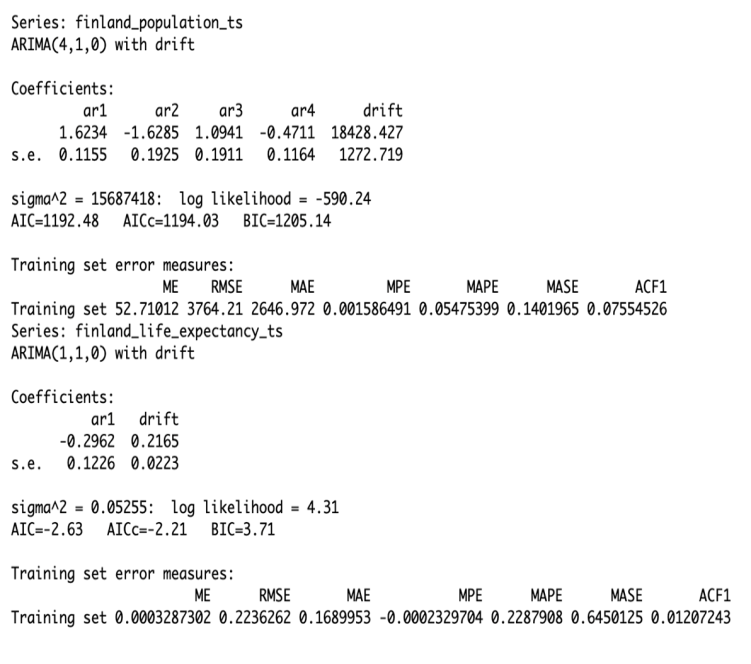


From the above ACF plots, we can observe that China and Finland ACF plots show significant values up to 16 lags, but Tuvalu shows 15 lags suggesting the presence of seasonality in the life expectancy data. Only the first lag is significant for all the three countries in the PACF plots.

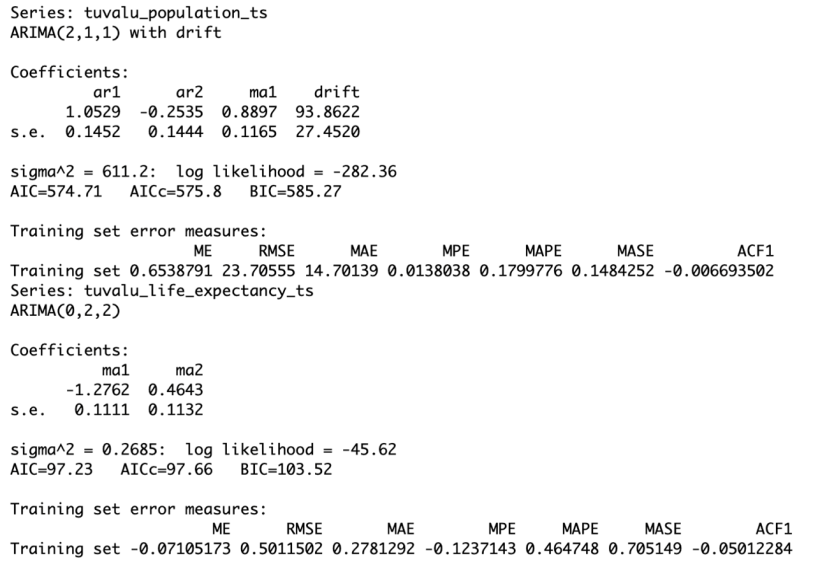
**Model Estimation:** Utilizing the identified parameters to estimate the ARIMA model that best fits the data by selecting optimal values for p, d, and q (the ARIMA parameters).

**Model Evaluation:** Assessing the model's performance through diagnostics to validate randomness in residuals, absence of autocorrelation, and consistent variance.

**Summary Statistics of the ARIMA models for all the three countries:**

**a. China b.Finland**



**c. Tuvalu**

**Forecasting:** Utilizing the ARIMA models to forecast life expectancy trends for future time periods based on historical patterns observed in the data. This involved predicting life expectancy rates for the next ten years (up to 2031) to anticipate potential changes or growth in life expectancy.

**ARIMA Forecast Plots:**

|  |  |
| --- | --- |
|  | a. The Arima model predicts that China's population will decrease over the next 10 years, but the  life expectancy is expected to go up. |
|  | b. For Finland, the model suggests that both the population and life expectancy are on the rise in the next 10 years. |

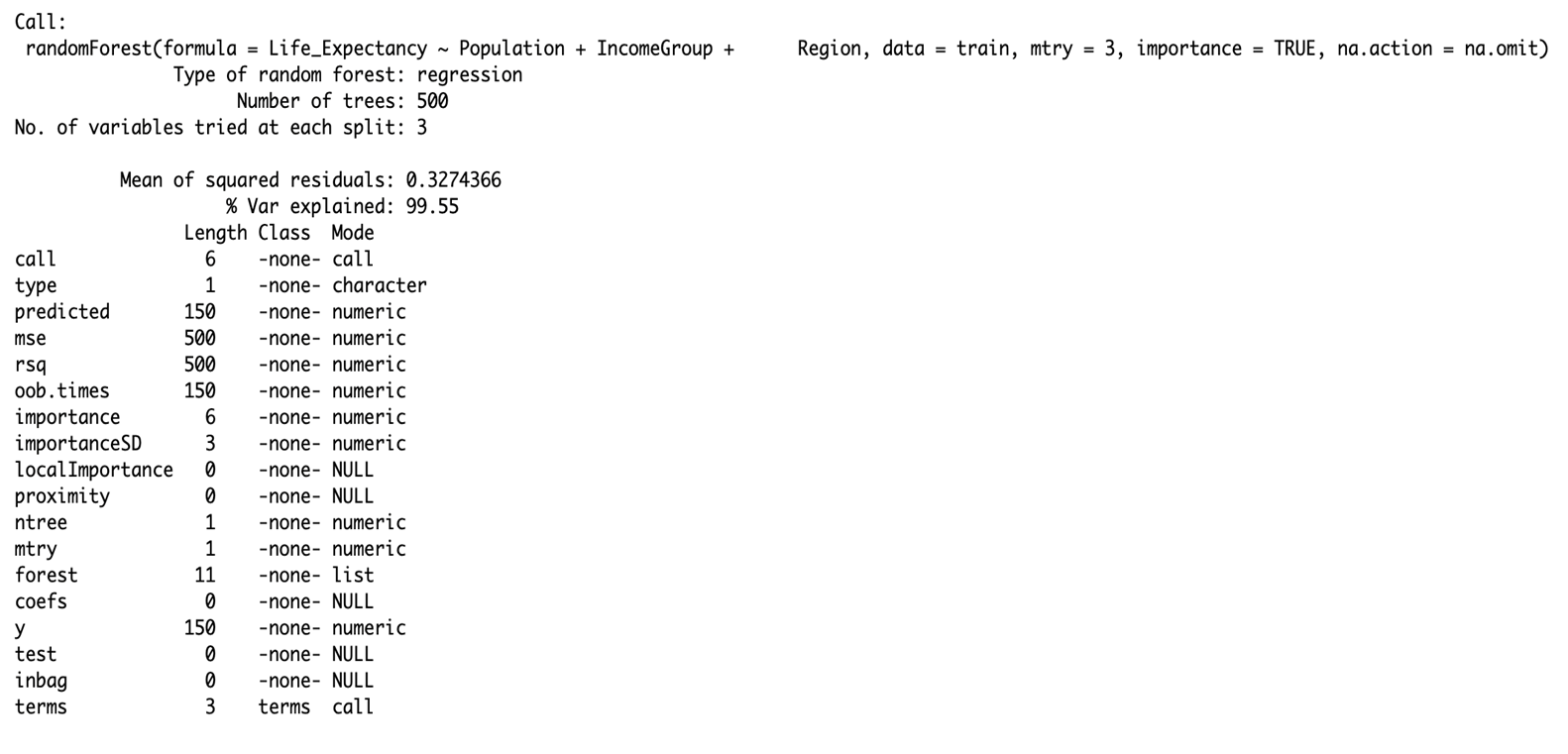
|  |  |
| --- | --- |
|  | c. In Tuvalu, both the population and life expectancy rates are expected to increase over the next 10 years. |

**b. Random Forest Regression Model**

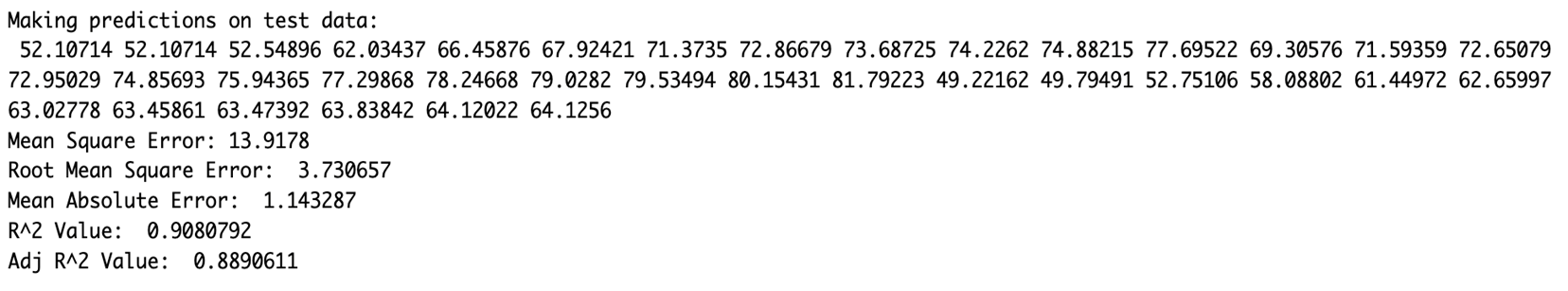
The randomForest() function is used to build a Random Forest Regression model (RF\_model) with the target variable. Life\_Expectancy predicted by predictors such as Population, Income Group, and Region.

Parameters include mtry = 3 (number of variables randomly sampled as candidates at each split), importance = TRUE (calculating variable importance), and na.action = na.omit (handling missing values).

**Summary Statistics and Evaluation of Random Forest Regression Model:**



Performance metrics like Mean Squared Error (RF\_mse), Root Mean Squared Error (RF\_rmse), Mean Absolute Error (RF\_mae), R-squared (RF\_r2), and Adjusted R-squared (RF\_adjR2) are calculated to evaluate the model's predictive accuracy.



**Summary of Random Forest Regression Model:**

From the above values, we can observe that the error values are small, which implies the model’s predictive power is good. The variation explained is 99.55% for the model, which is extremely high suggesting a good fit.

**c. Linear Regression Model**

**a. For base model:**

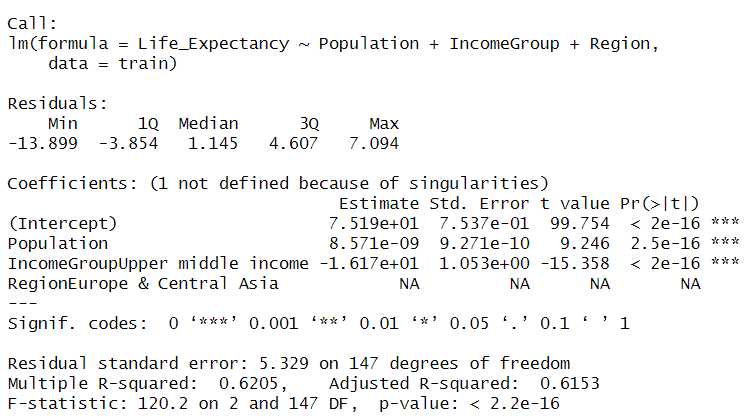
1. Using the lm() function, a Linear Regression model (LR\_model) is created with Life\_Expectancy as the dependent variable and Population, IncomeGroup, and Region as independent predictors using the training dataset (train).

2. The summary() function provides an overview of the Linear Regression model (LR\_model), presenting details such as coefficients, standard errors, t-values, and p-values for each predictor.

3. Predictions (LR\_preds) are made on the test dataset using the created Linear Regression model.

4. Performance metrics such as Mean Squared Error (LR\_mse), Root Mean Squared Error (LR\_rmse), Mean Absolute Error (LR\_mae), R-squared (LR\_r2), and Adjusted R-squared (LR\_adjR2) are computed to evaluate the model's predictive accuracy.

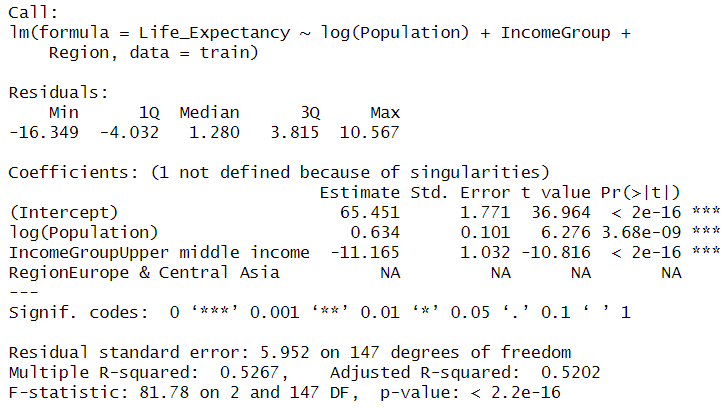
**Summary Statistics of the model:**



**b. For log transformed model:**

In this Linear Regression model, a log transformation is added to one of our predictor variables which is Population to analyze which model performs better.

**Summary Statistics of the model:**



**Interpretation for the "Upper middle income" category in the "IncomeGroup" variable:**

The coefficient value of -11.165 suggests that, compared to the reference category (likely the lowest income group), being classified as an "Upper middle income" country is associated with an average decrease of approximately 11.165 years in life expectancy. However, it's essential to note that this interpretation is in relation to the reference category within the "IncomeGroup" variable. If the reference category changes or if more categories are added, the interpretation would adjust accordingly.

**Comparision of the above two Linear Regression models:**

From the summary statistics of the above two models, the original linear regression model without the log transformation of the population field performs better among them. This is evidenced by lower error metrics (RMSE, MAE) and a higher adjusted R-squared, indicating improved predictive accuracy compared to the model utilizing the log-transformed population variable.

**Model Comparisions:**

For the chosen datasets and the problem statement, we decided to compare and derive insights from the two models **Random Forest Regression model** and **Linear Regression model.** Following are the various error values and R Squared, Adj R Squared values of both the models:

|  |  |
| --- | --- |
|  | The Random Forest model exhibits superior performance with notably lower MSE, RMSE, and MAE, displaying enhanced precision in predictions. Additionally, its higher R-squared value indicates a superior fit to the data compared to the Linear Regression model. |

**Model Comparision using Plots:**

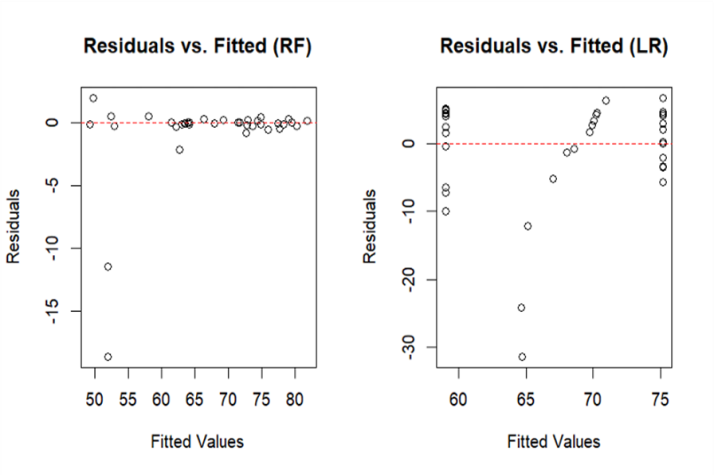
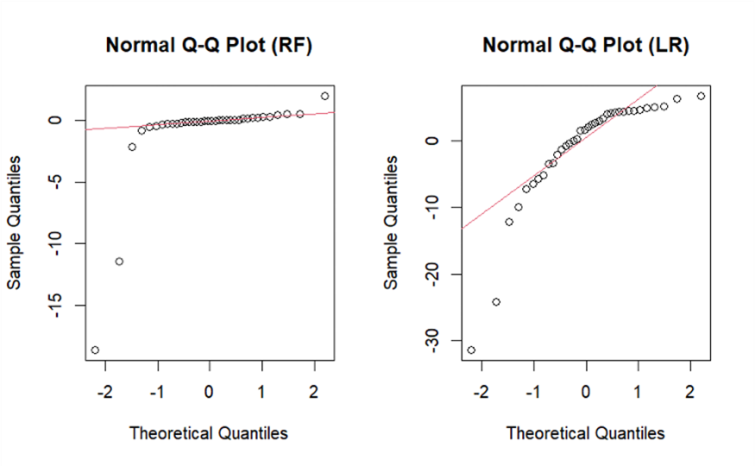
**a. Predicted vs Actual Plots:**

|  |  |
| --- | --- |
|  | Based on the visual analysis of the plots, the Random Forest Regression plot demonstrates a linear alignment between predicted and actual values, indicating a stronger model performance. Conversely, the Linear Regression plot displays deviations from linearity, suggesting less precision in predictions. Consequently, **the observation supports the superiority of the Random Forest model over Linear Regression** in this context. |

**b. Forecast Plots:**

|  |  |
| --- | --- |
|  | From the Forecast plots, it is evident that the Random Forest Regression forecast aligns closely with the original line, signifying a strong fit. In contrast, the Linear Regression forecast noticeably deviates from the original line, implying a lesser fit. Consequently, this analysis supports the assertion that the **Random Forest Regression model outperforms the Linear Regression model** in this comparison. |

**Model Diagnostics:**



**Interpretations:**

In the Q-Q plots, the Random Forest Regression model demonstrates a closer adherence of residuals to the expected normal distribution line, suggesting a better fit compared to the Linear Regression model. Similarly, in the Residuals vs Fitted plots, the Random Forest Regression model exhibits a more even distribution of residuals around the zero line, indicating better model performance, while the Linear Regression model shows more pronounced deviations. These observations further support the conclusion that the Random Forest Regression model is superior in terms of fitting the data and minimizing residuals compared to the Linear Regression model.

**Important Feature:**

|  |  |  |
| --- | --- | --- |
|  |  |  |

The feature importance analysis generated from the Random Forest Regression model indicates that "Population" emerges as the most influential predictor for estimating life expectancy. The plot illustrates that population size contributes significantly to the predictive power of the model, accounting for approximately 60% of the overall importance. Conversely, "Income Group" exhibits the least impact, contributing around 11% to the model's predictive capacity. This analysis highlights the crucial role of population size in determining life expectancy, indicating its substantial influence among the predictors considered.

**References:**

[1] Link to the population dataset: [**https://data.worldbank.org/indicator/SP.POP.TOTL**](https://data.worldbank.org/indicator/SP.POP.TOTL)

[2] Link to the life expectancy dataset: [**https://data.worldbank.org/indicator/SP.DYN.LE00.IN**](https://data.worldbank.org/indicator/SP.DYN.LE00.IN)

[3] <https://github.com/kiranshahi/Life-expectancy-prediction>

[4]https://www.emro.who.int/emhj-volume-26-2020/volume-26-issue-2/analysis-of-life-expectancy-across-countries-using-a-decision-tree.html