Analysis & Forecasts of Consumer Price Index for All Urban Consumers in the U.S. City

Group 22
Joyce Chen: j982chen@uwaterloo.ca
Linda Li: c592li@uwaterloo.ca
Skylar Zeng: j44zeng@uwaterloo.ca
Zixuan Zhu: z298zhu@uwaterloo.ca
Yingxiang Li: y2979li@uwaterloo.ca

Table of Contents

Responsibilities

Project, Plan, Data, Conclusion, References: Skylar Zeng
Dataset collection, Analysis-Regression: Linda Li
Residual Diagnostic, Analysis-Smoothing (Differencing, Decomposition): Zixuan Zhu
Dataset collection, Analysis-Smoothing (Holt-Winters), Conclusion: Joyce Chen
Analysis-Box Jenkins: Yingxiang Li
Powerpoint, Presentation: All members

Project

This report delves into the analysis of the Consumer Price Index for All Urban Consumers, with a specific focus on Lodging Away from Home in the U.S. City Average from the Federal Reserve Bank of St. Louis website. This study particularly emphasizes lodging, as it reflects fluctuations in prices paid by urban consumers for accommodation while travelling. The accommodations include a diverse range of lodging options like hotels and motels, highlighting their significance in understanding consumer expenditure patterns and economic trends related to travel. For example, if the index shows a significant increase, it could indicate rising travel costs, which could impact consumer spending habits, the travel industry, and the overall economy. This report aims to project future trends in the CPI for this category, offering insights into anticipated economic and consumer behavioural shifts.

Plan

Based on the R output, there is seasonality and trend within the dataset, posing challenges for accurate forecasting. To address these obstacles, the study proposes the application of Regression, Smoothing methods, and Box-Jenkins models. These approaches will develop a range of candidate models. Through the comparative evaluation of the models generated by these methods, we aim to identify the most effective forecasting model, considering both accuracy and applicability to the dataset's characteristics. This selection process will enable us to overcome the identified obstacles and achieve reliable forecasts.

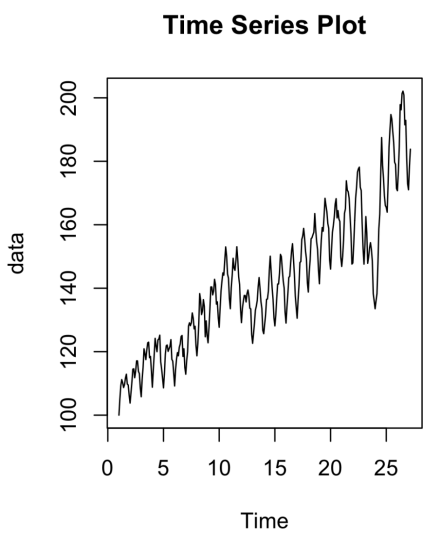
Data

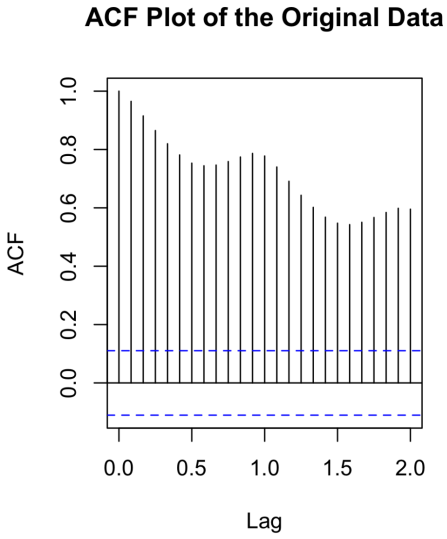
We select the consumer price index (CPI) values for the first day of every month and form a dataset from Feb 1st, 2000 to Jan 1st, 2024. We apply Box-Cox transformation (log the original dataset) to our data to stabilize the variance. We also separated data from Feb 1st, 2022 to Jan 1st, 2024 as the test set and data from Feb 1st, 2000 to Jan 1st, 2022 as the training set for the following approaches.

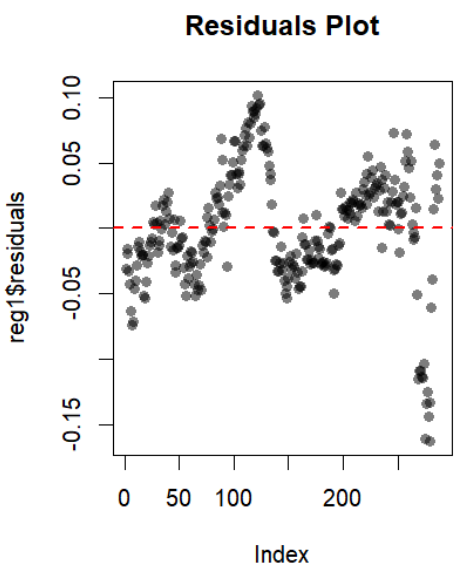
Clearly, from the time series plot, we notice apparent signs of both trend and seasonality. ACF plot showcases a sine curve decay with most of the spikes crossing beyond the confidence limits (blue lines), indicating non-stationarity and correlation among the original data.

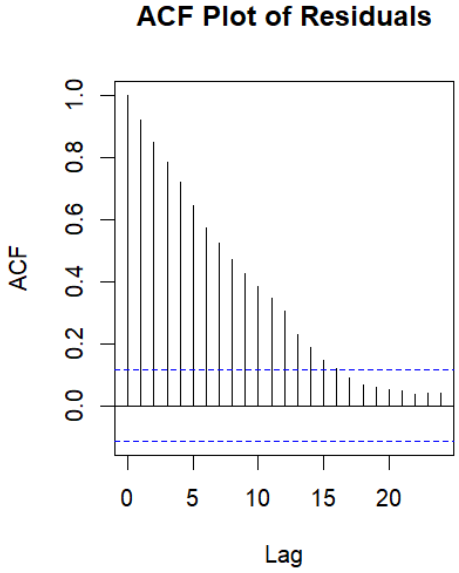
For residual diagnostic, we test the four properties: normality, constant mean, constant variance, and randomness (independence). Firstly, from the Shapiro-Wilk Test for normality, the p-value is much lower than 0.05 (use this default level for the whole report), which shows that the residuals are not normally distributed. Next, inside the residual plot, there seem to be some patterns and hence, the mean of residuals is not constant. The ACF plot of residuals also shows a slow exponential decay, meaning that residuals are highly dependent and correlated. Furthermore, the p-value result from the Fligner-Killeen Test is higher than 0.05, which suggests

constant variance for residuals. Last but not least, both difference sign and run test have p-value output lower than 0.05, which suggests non-randomness for residuals.



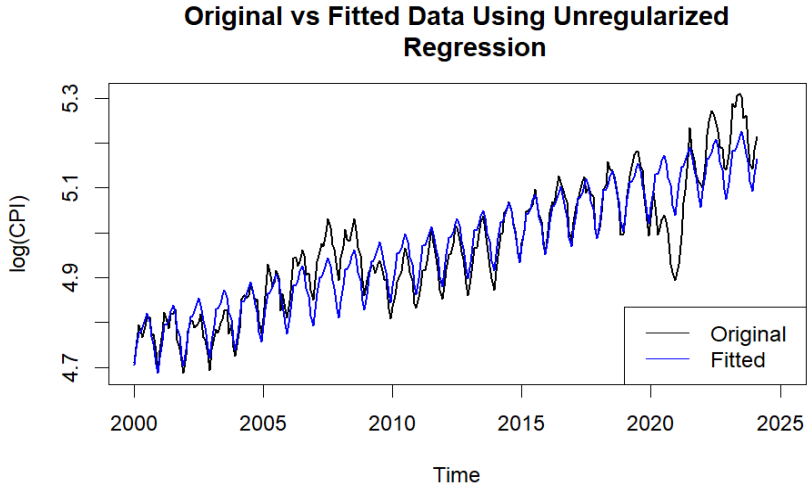




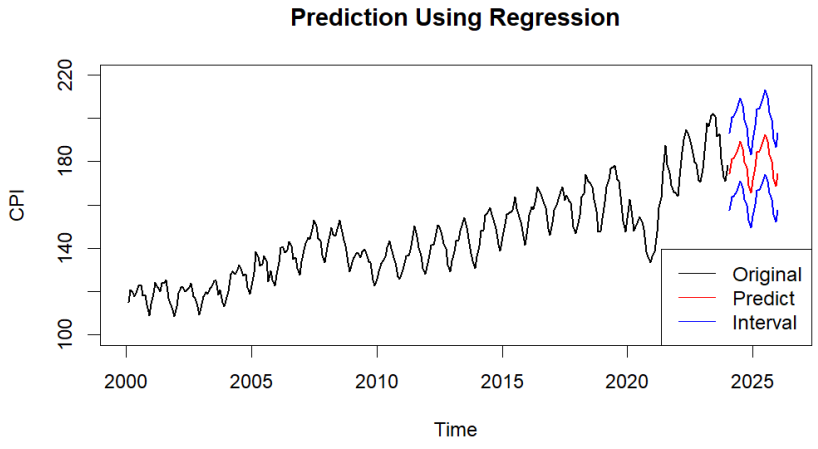


Analysis - Regression

The first strategy used to analyze the data is unregularized regression with orthogonal polynomials. Since the data has strong seasonality with a period of 12, indicator variables are also included. Through fitting polynomials of various degrees on the training set and calculating the average prediction squared error (APSE), it is observed that APSE is minimized when employing a degree 1 polynomial with seasonal components.

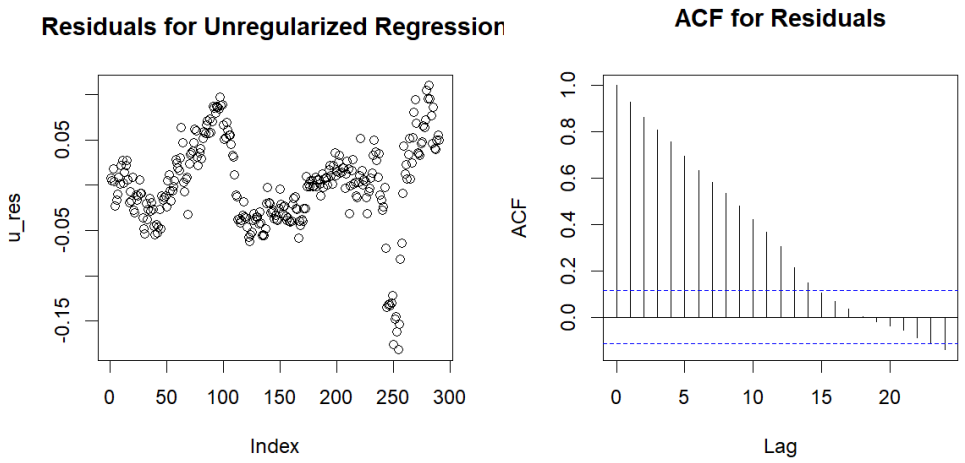


The selected model captures the increasing trend and the general seasonality in the original data. However, there are limitations for time after 2020, where a change point is presented. This is because a relatively simple model may not adequately capture such structural change.



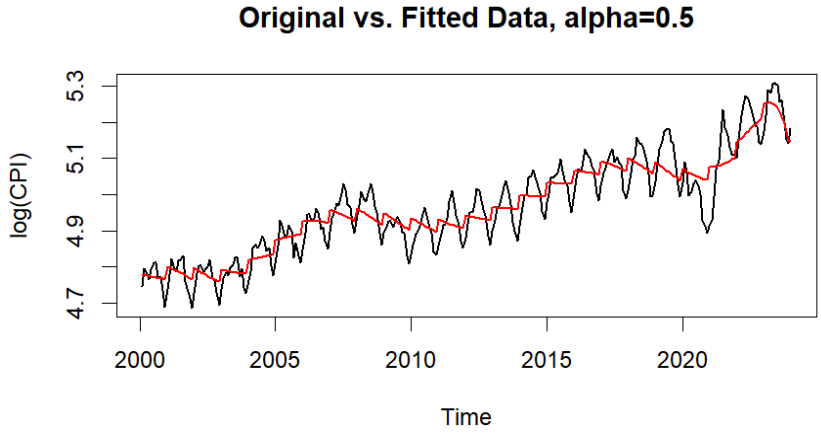
The predicted data follows the seasonal pattern of the original data, but it does not follow the increasing trend, with values less than the final values from the original data.

While the fit of the model and predicted data seem reasonable, it is also important to consider if the residuals meet all the assumptions.

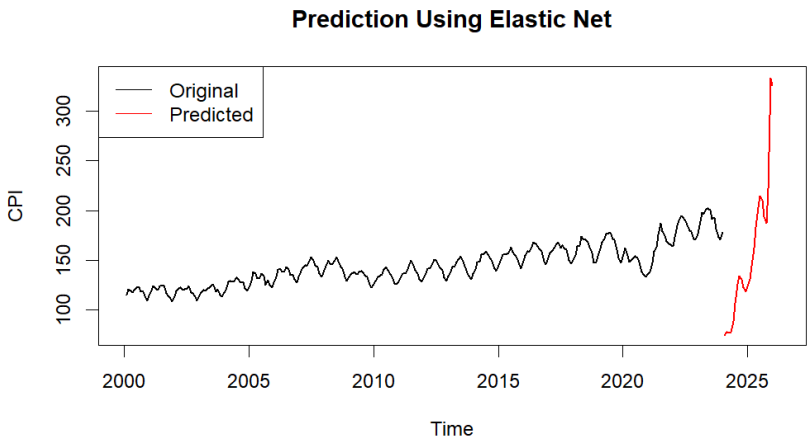


From the residuals plot, data points are not randomly distributed, indicating a violation of the constant mean assumption. This is further supported by the Autocorrelation Function (ACF) plot, which shows a slow decay. Consequently, this model cannot effectively remove the trend, therefore it might not be a suitable candidate for forecasting purposes.

The Regularized polynomial regression technique is then utilized, with cross-validation on the training set and APSE calculation on the test set. The optimal model among Ridge, Lasso and Elastic net is determined to be Elastic net regression with degree 12 polynomials.



The selected model effectively captures the increasing trend but is less effective in fitting the seasonal pattern in the original data. It captures the change in values for data after 2020 to some degree, possibly due to the adoption of a relatively complex model.



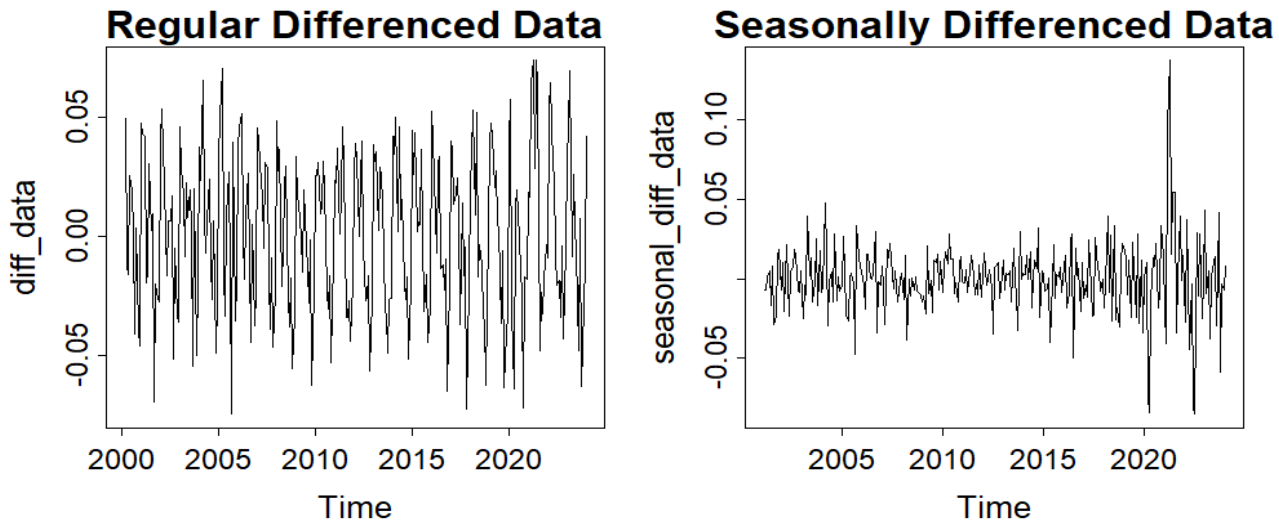
However, the complexity of this model leads to a bias-variance tradeoff that could affect the prediction accuracy. The predicted values from Elastic net regression do not appear to be reasonable, showing a significant difference in structure compared to the original data. The residual analysis is also performed, both the residuals plot and ACF plot suggest there is a strong seasonal component that has not been removed by regression. Consequently, this model may not be a suitable candidate for forecasting.

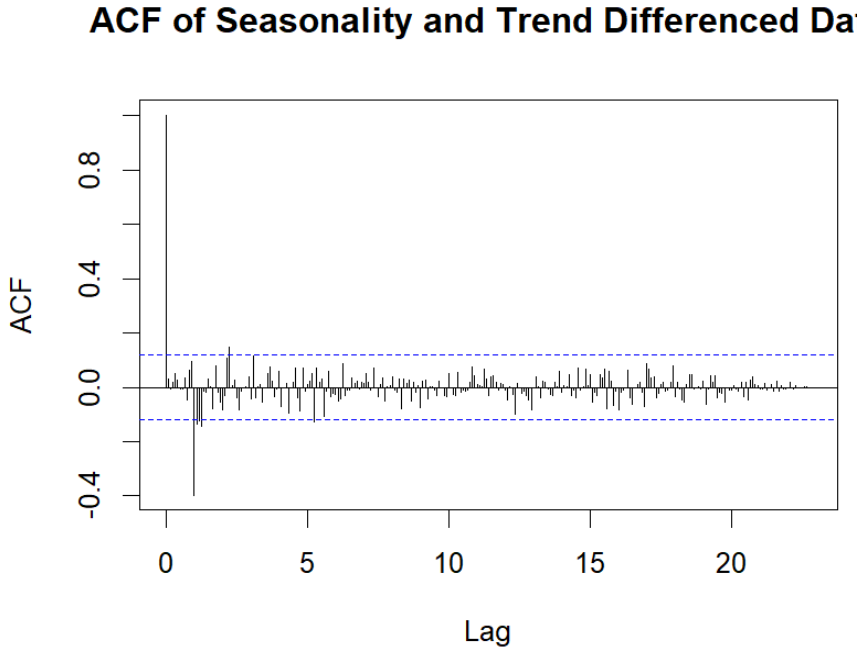
Since the APSE of the Lasso regression model is very close to that of Elastic Net, it could also be considered. However, it performs poorly in forecasting data and also exhibits seasonality in residuals. To effectively remove seasonal patterns from the original data, smoothing is needed, and it will be applied to the entire data in the next sections.

Analysis - Smoothing

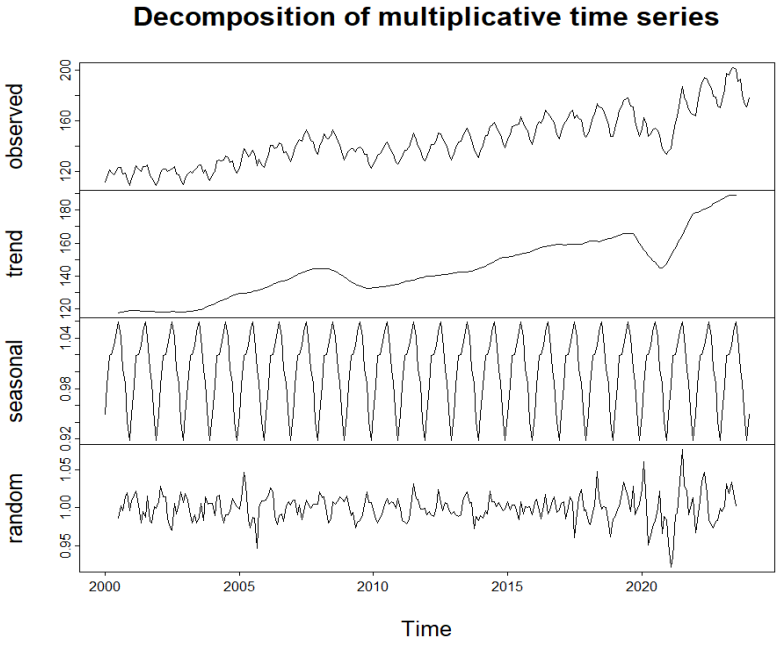
Once we would like the residuals (detrendized and deseasonalized data) to be stationary, we combine regression with the differencing method to obtain stationarity. We decided to use first-order differencing that avoids over-differencing (because we don't want variance to increase exponentially). Besides, we set the lag value as 12, which is equal to the period/frequency value.

Although there is no trend when we look at the regular differencing plots, seasonality still exists. In this scenario, we change to both seasonal and regular differencing methods and it turns out that the corresponding graph has no trend and seasonality. The ACF of seasonal and regular differencing data represents stationarity and uncorrelatedness with a fast decay graph and most of the random data lying among confidence limits (blue lines).



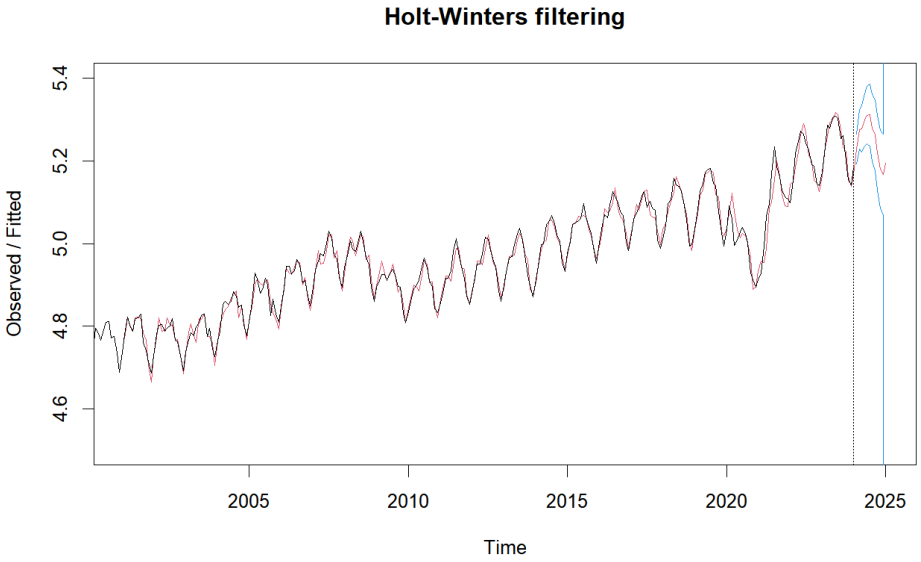


Comparing additive time series with multiplicative form, each of the four components is almost the same as the corresponding one in both decomposition plots. Notice that the range of the seasonal component is substantial relative to the error term range, which implies that the seasonal component is a reasonable component that needs to be included.



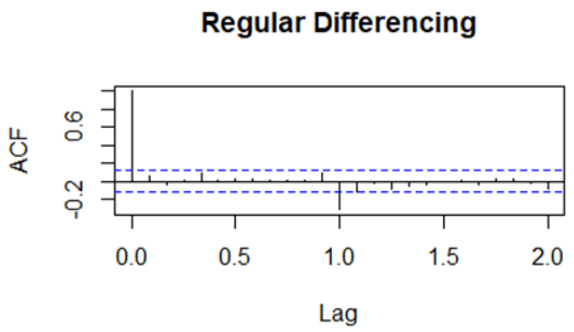
We use the Holt-Winters method simply because we wish to extend exponential smoothing to accommodate projecting the trend for the dataset into the future and to accommodate seasonality as well. For our dataset, we cannot use simple and double exponential smoothing methods because there are seasonal and trend updates.

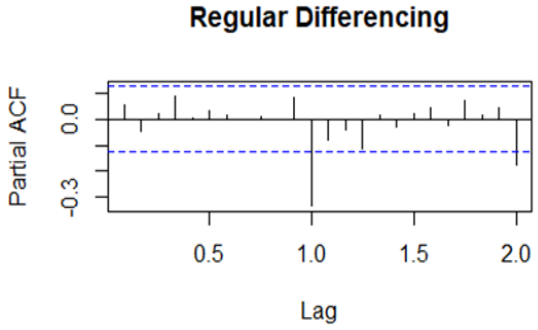
To choose the better one from the additive and multiplicative Holt-Winters model, we use the previously chosen test and train set and aim to predict the next 2 years' CPI values until Jan 1st, 2025. Then, we calculate and choose the model with a lower MSE value. The R output shows that the value for the multiplicative Holt-Winters model (0.002847429) is lower than that for the additive model (0.003091109). In the generated plot, the red fitted value line matches close to the black original data line. Thus, our final decision is to use the multiplicative Holt-Winters model to estimate and predict the dataset. The following plot shows the fit and prediction for the original dataset (without log).



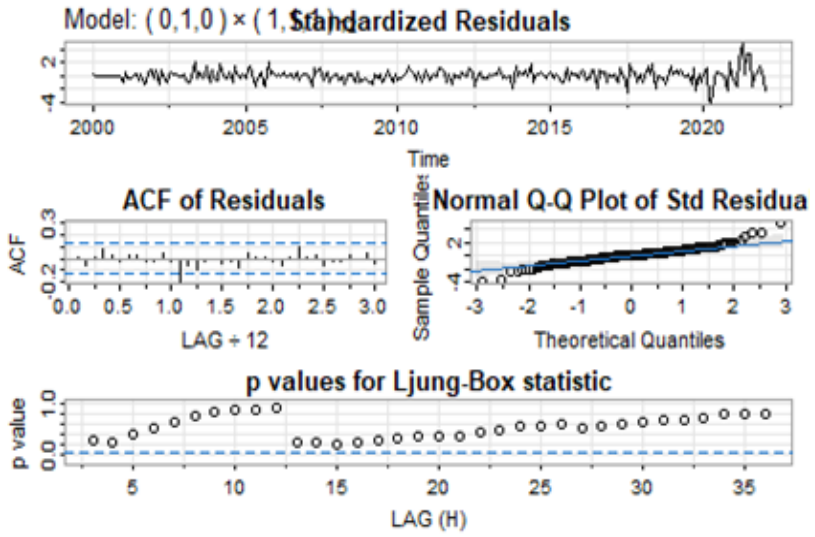
Analysis - Box Jenkins

In this pivotal stage of analysis, it becomes imperative to ascertain the most adept predictor, a process necessitating the elimination of both trend and seasonality through differential techniques. The utilization of differencing serves as an indispensable tool in this endeavour, with the smoothing method indicating a seasonal cycle of 12 months, thereby rendering first-order differencing efficacious in mitigating the prevailing trend. In this case, the number of regular differencing and seasonal differencing can be determined. Still, other parameters need to be decided by using Auto-Correlation Function (ACF) and Partial Auto-Correlation Function (PACF) plots.





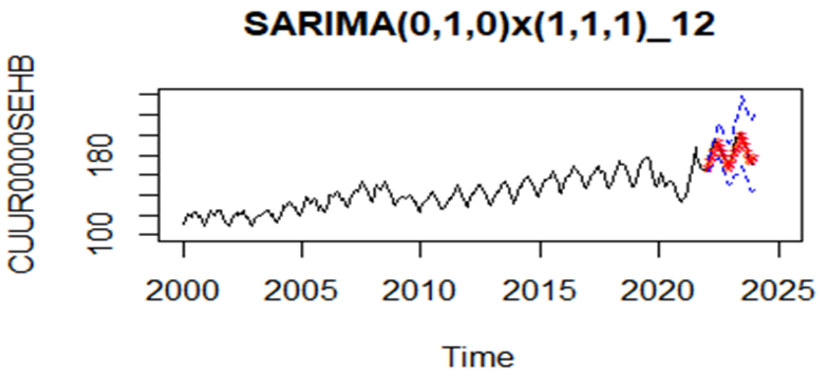
According to the ACF and PACF plots, ignoring the spikes every 12 lags, this suggests a non-seasonal model. There is no correlation and partial correlation for the non-seasonal model. Therefore, both p and q are chosen to be 0 and no other potential values. Moreover, P and Q can be identified by ignoring all spikes except the seasonal lags. Firstly, ACF cuts off after lag 1 or indicates exponential decay. And PACF cuts off at lag 1 or 2 or potential exponential decay. Therefore, 3 candidate models are SARIMA(0,1,0) × (1,1,1)_12, SARIMA(0,1,0) × (0,1,1)_12 and SARIMA(0,1,0) × (2,1,0)_12.



From the standardized residuals plot, all three models illustrate no clear pattern and almost all values are around 0, which means these models have constant mean 0 and constant variance. Moreover, the normal Q-Q plot shows most points lie on the blue line. Hence, these models meet the normal assumption. In addition, the residuals are also uncorrelated since no spikes are above the blue line.

Now, in order to decide the final model, AICc or BIC are good if interest is in the fit. Unfortunately, AICc and BIC are very close and cannot be used to determine the final model according to R output. Hence, we need to shift the model to the testing set and see how well every model fits according to the APSE value, then the lowest APSE indicates the best model.

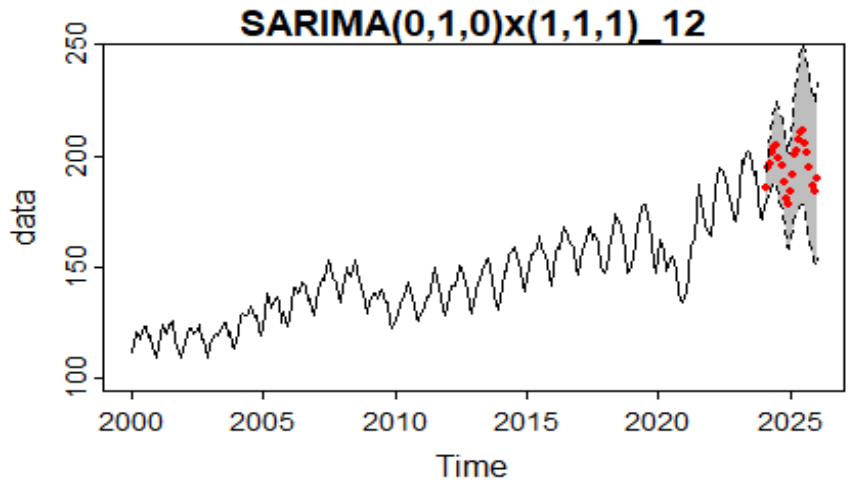
Therefore, upon meticulous examination of the R output, SARIMA(0,1,0) × (1,1,1)_12 provided the smallest APSE. It is unequivocally evident that the first one is the foremost model, emerging as the optimal choice, indisputably boasting superiority.



The aforementioned illustration depicts the efficacy of the model's fitment on the test set. Herein, the black line delineates the observed data points, juxtaposed with the red line denoting the fitted values generated by the model. Additionally, the blue dashed lines delineate the prediction interval.

Therefore, upon meticulous examination of the R output, it is unequivocally evident that the foremost model denoted as SARIMA(0,1,0) × (1,1,1)_12, emerges as the optimal choice, indisputably boasting superiority. The aforementioned illustration depicts the efficacy of the model's fitment on the test set. Herein, the black line delineates the observed data points, juxtaposed with the red line denoting the fitted values generated by the model. Additionally, the blue dashed lines delineate the prediction interval.

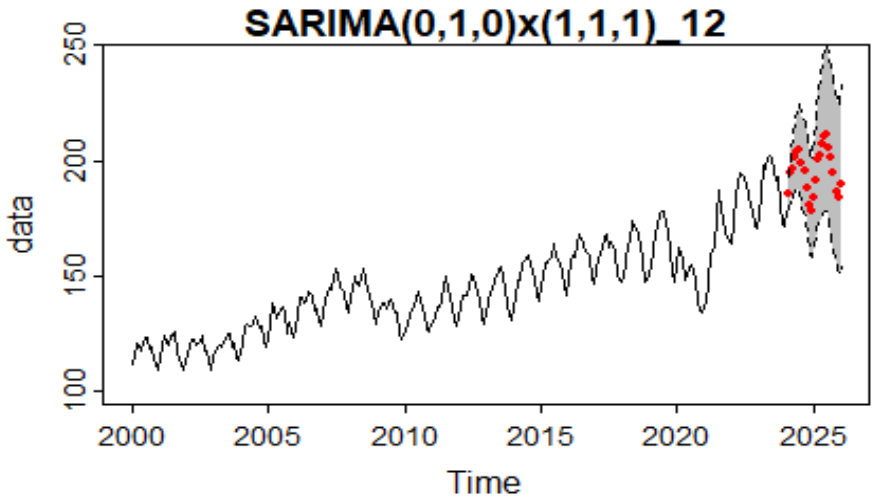
Ultimately, the culminating step in this analytical progression involves the recalibration of the model, amalgamating both the training and test datasets to ensure comprehensive coverage, thereby facilitating the generation of forecasts pertaining to future observations.



Conclusion

In conclusion, our analysis evaluated various forecasting models to determine the most effective method for predicting future CPI values. Initially, despite the Lasso regression's close APSE to the Elastic Net, its limitations in forecasting and removing seasonality shifted our focus toward models that could address these issues. The multiplicative Holt-Winters model emerged as a preferable choice over the additive version due to its lower mean squared error (MSE) and better fit to the observed data.

Ultimately, the SARIMA(0,1,0) × (1,1,1)_12 model was selected as the optimal forecasting tool. This model not only provided the best fit to the test set but also showcased accuracy in its predictions, as evidenced by the close alignment of the fitted values with the observed data. The decision to re-adjust this model using both the training and test datasets ensures that it is finely tuned for generating accurate forecasts for future observations.



The above prediction plot indicates a forecast extending two years into the future of the data with our selected optimal model, with the observed data denoted by a solid black line and the predicted values indicated by red dots, surrounded by a prediction interval represented by dashed lines.

The general upward trend in the CPI data suggests that the cost of lodging away from home is expected to increase. This could be a result of several factors including inflation, increased demand for travel, or a rise in operational costs for lodging facilities.

The predicted rise in lodging costs may also indicate consumer confidence in spending, as people are willing to pay more for travel experiences. This could suggest a shift towards valuing

experiences over goods, or a robust economic outlook where consumers are less hesitant to allocate funds for leisure. Also, notice that peaks in lodging CPI often coincide with holiday seasons and vacation periods.

It is important to note that while the model forecasts an increase in behaviour, the actual future CPI will depend on various dynamic factors including economic and political factors, consumer confidence, and global events. This forecast can serve as a reference opinion. It needs to combine with other economic indicators for more comprehensive and accurate future trends.

Reference

U.S. Bureau of Labor Statistics, Consumer Price Index for All Urban Consumers: Lodging Away from Home in U.S. City Average [CUUR0000SEHB], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/CUUR0000SEHB , March 7, 2024.