STAT 443 FINAL PROJECT

Forecasting Flu Data by Tom Guo, Joy Yang, Kyle Ye

1. Introduction

1.1 Statement of the Problem

Seasonal influenza is a recurring public health challenge, causing significant illness, hospitalizations, and deaths annually. Monitoring and forecasting flu-related patient visits is essential for effective public health readiness and resource allocation. Understanding trends in flu-related patient data allows healthcare systems to anticipate surges in demand and respond to outbreaks more effectively. Our project focuses on analyzing flu-related patient data at the national level to identify patterns and develop forecasting models. By doing so, we aim to provide insights that can support decision-making for public health interventions, particularly in a post-COVID era where healthcare systems face increased strain.

1.2 Motivation

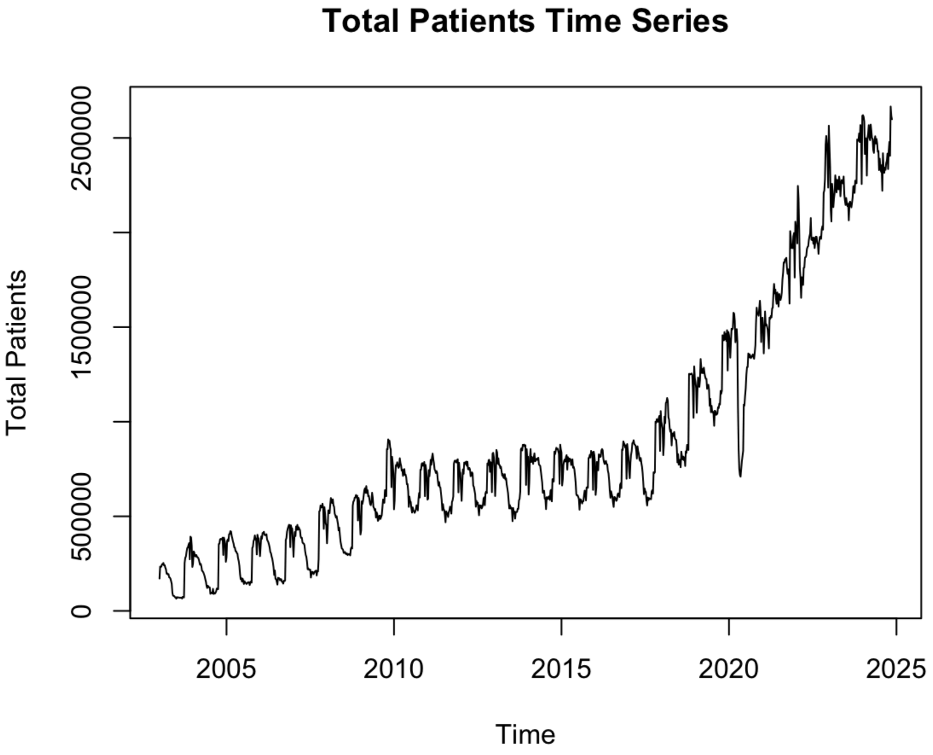
The motivation for this project stems from the profound impact of the COVID pandemic. COVID started when we were in high school, and lockdown was called off after our first year of University. As a group who lived through a pandemic at a relatively early age, we are particularly interested in examining its impacts on flu patient visits.

1.3 Objective

The primary objective of this project is to analyze the total number of flu-related patient visits weekly from January 2003 to December 2024 and forecast five years in the future. This analysis will help uncover trends, seasonal patterns, and potential shifts in healthcare demand post-pandemic.

1.4 Data

The dataset used in this project comes from the National, Regional, and State Level Outpatient Illness and Viral Surveillance. It spans from 1997 to 2024 and provides weekly observations. However, due to missing data points and the abundance of earlier data, we chose to focus on the period from January 2003 to December 2024, where no missing data points are present. The observations in the dataset are recorded weekly, but some years contain 53 weeks instead of the usual 52 weeks. To simplify the analysis and ensure consistency (and under the recommendation of our professor), the data was standardized into a time series with a frequency of 52 weeks per year. This approach allowed us to maintain a uniform structure for modelling and analysis, as we lacked a reliable way to handle irregular time series data. Below is a plot of our time series.



The plot reveals several important trends and patterns over the analyzed period from 2003 to 2024. From 2003 to 2010, there is a slight upward linear trend in the number of flu-related patient visits, indicating a gradual increase in total patients during this time. From 2010 to 2018, the trend appears relatively constant, with no significant long-term changes. However, from 2018 onwards, there is a pronounced linear upward trend, much steeper than the one observed between 2003 and 2010. This substantial increase corresponds to the timing of the COVID pandemic, which likely influenced the dynamics of patient visits.

Despite the significant upward trend after 2018, the data shows a large dip in the middle of 2020, which stands out as a probable outlier. This dip may be attributed to the implementation of COVID regulations like social distancing and lockdowns, or a widespread round of immunization shots reducing flu-related visits during that period. Seasonality is also clearly evident in the data, as expected given the cyclical nature of flu seasons. Moreover, the influence of climate and seasonal changes on flu activity is well-documented, and the consistent periodic fluctuations in the data align with this understanding.

It is important to note that forecasting this dataset poses significant challenges due to the clear change point introduced by the COVID pandemic. The event caused abrupt changes in how we viewed health care in the 20th century, making the post-pandemic data unlikely to follow the same dynamics as pre-pandemic data. Throughout our analysis, we will remain mindful of this change point and its potential impact on the accuracy of our forecasts.

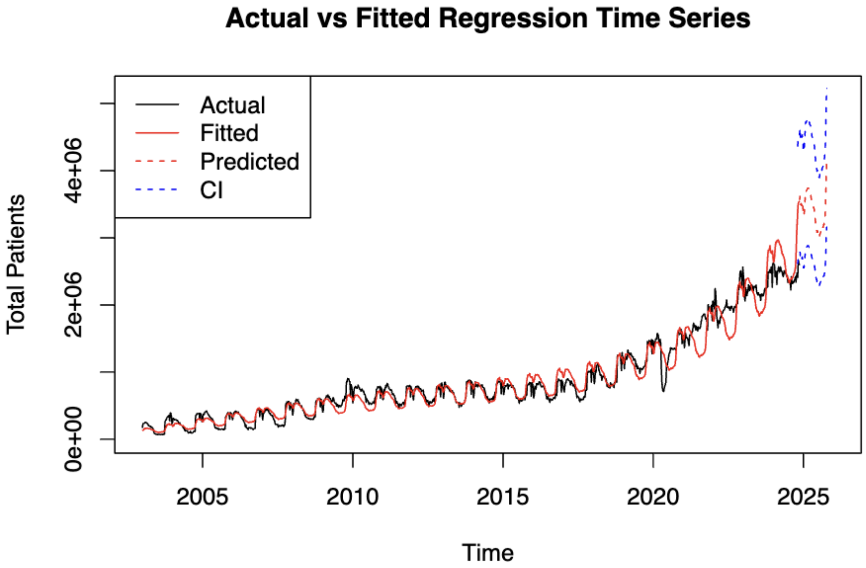
2. Modelling

2.1 Unregularized Regression Models

The first step for regression modelling was to determine the appropriate transformation for our data to stabilize the variance. We applied the Box-Cox transformation and identified an optimal lambda value of 0.222. Once the transformation was applied, we split the dataset into training and testing sets. The training set included all data points up to the start of 2023, while the testing set comprised data from January 2023 to December 2024. This split ensured that the testing set represented approximately 8.5% of the total data, providing a sufficient number of observations for evaluation without compromising the training set size.

With bias-variance trade-off in mind, we proceeded to fit polynomial regression models from degree 1 to degree 7, to capture the trends and non-linear relationships in the data. Initially, we used mean squared error (MSE) to evaluate model performance, as done in class. However, due to the large scale of our data, where observations often exceeded five-digit numbers, the resulting MSE values were disproportionately large, making them challenging to interpret. To address this issue, we opted to use absolute mean error (AME) instead, as it provided a more interpretable measure of error for our dataset. For the rest of this report, AME will be used.

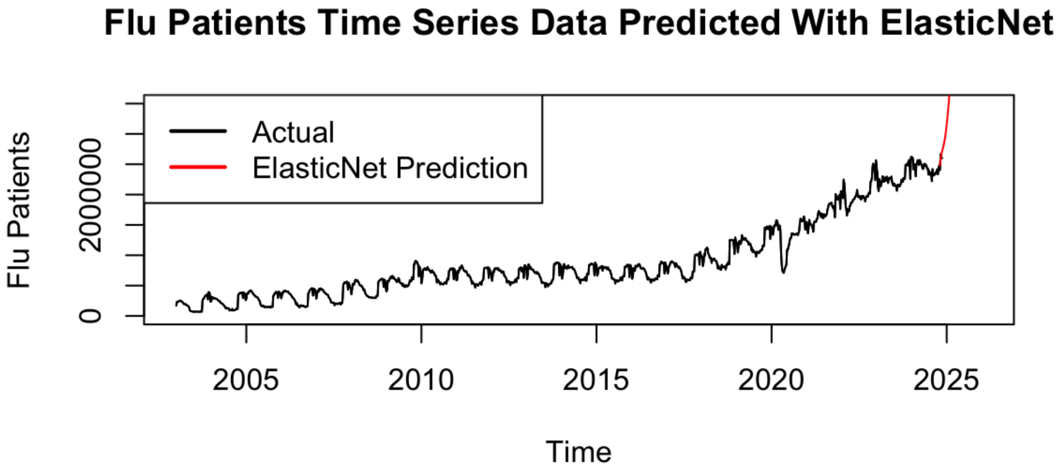
After evaluating the models, we found that the 4th-degree polynomial regression model marginally yielded the lowest AME, making it the most suitable unregularized model for our dataset. Using the 4th-degree model to forecast the next year, we got the following result.



We can see very clearly that before COVID, the fitted models were relatively accurate and followed the same shape with similar seasonality as the actual data. The 4th-degree polynomial model is the best fit indicating a non-linear relationship in the data. While the model fits well within the observed range, the 95% confidence intervals as seen are very wide which is likely due to the change point, but also could be the potential unpredictability of higher-degree polynomials. From the residuals vs fitted values plot, refer to appendix, figure 1, the linear decrease paste halfway suggests the model may be underfitted. When we run the residual, ACF, and QQ plots, refer to appendix, figure 9, we see clear evidence against homogeneity and randomness. Therefore our model is not adequate for our data. The point prediction values are valid but the confidence intervals are not valid because the assumption of residuals is not valid.

2.2 Regularized Regression Models

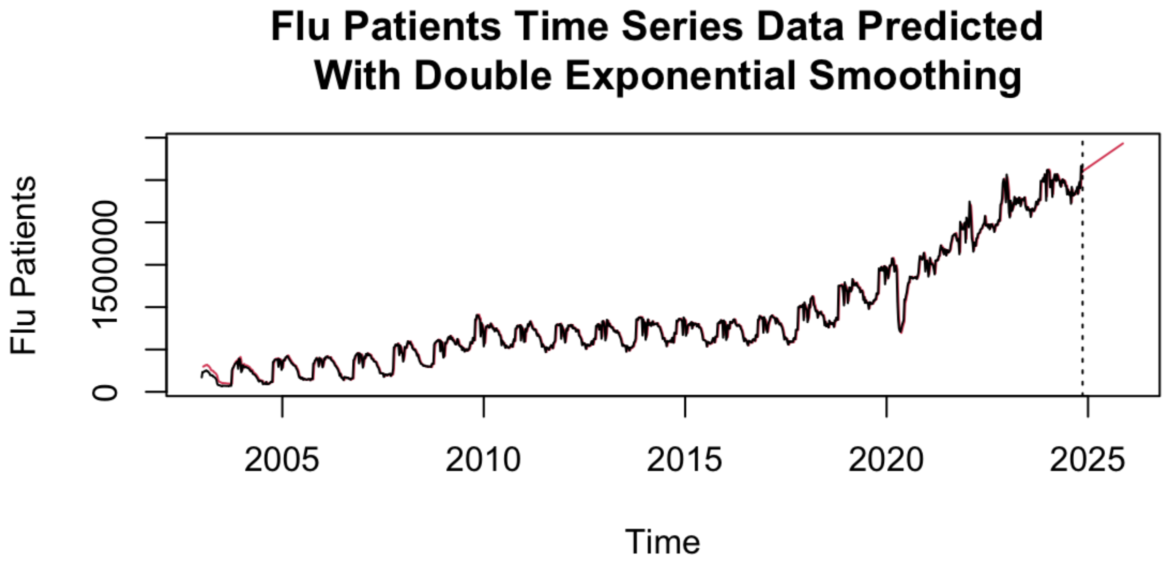
Next, we consider regularized regression models, fitting Lasso, Ridge Regression, and Elastic Net models. We fit these models following a similar procedure to Assignment 2, Question 1, using five different alpha values: 0, 0.25, 0.5, 0.75, and 1. Using the same train and test split as in the polynomial regression analysis, we evaluated models with polynomial degrees ranging from 1 to 7 again. Among these, the Elastic Net model with alpha = 0.25 gave the smallest ASE, indicating it provided the best balance between bias and variance for this dataset. The forecast of the next year using the optimal model is given below.



We can see that the prediction of 2025 is essentially just increasing exponentially, which is not an adequate forecast because seasonality is ignored. Referring to appendix, figures 2, the plot is linear and suggests the model is appropriately capturing the general relationship between the predictors and the outcome. However, from appendix, figure 3, a decreasing linear trend that disappears after lag 9, indicates autocorrelation for the first 9 lags. The residuals suggest that the model may not fully account for the time dependencies or autocorrelation structure in the data, particularly in the early lags. We will learn more about this issue when fitting the Box-Jenkins ARIMA and SAMIRA models.

2.3 Holt-Winters Models

For the Holt-Winters models, the training and testing data sets were adjusted. The training set spanned from 2003 to 2021, while the testing set included data from 2022 to the end, giving approximately 13% of the data to the test set, a better ratio. The Holt-Winters models were fitted according to the approach we used in Assignment 3, Question 1. We found that the double exponential smoothing model had the smallest AME. However, we believe this result may have been due to chance. Theoretically, we expected the additive Holt-Winters model to perform better, as the data exhibits clear seasonality with a roughly constant effect each year. The impact of the COVID-19 surge, which appears in the later part of the data, likely influenced the performance of both the additive and multiplicative Holt-Winters models. Despite this, the ACF plot of the double exponential smoothing model, refer to appendix, figure 4, shows stationary behaviour, particularly when considering the ACF at lags 0.1 and 0.13 as outliers. Below is the plot of the 2025 prediction.

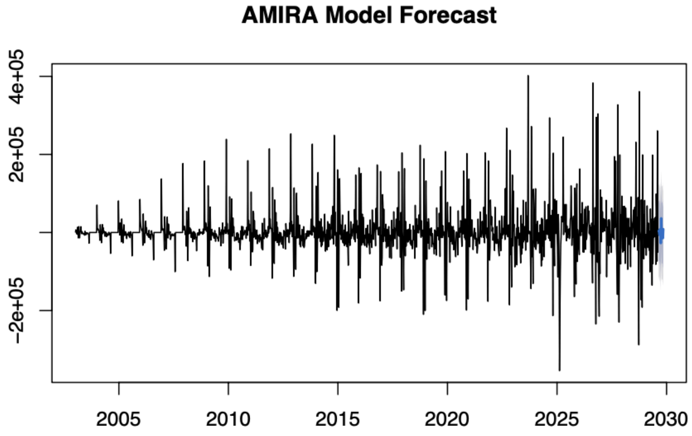


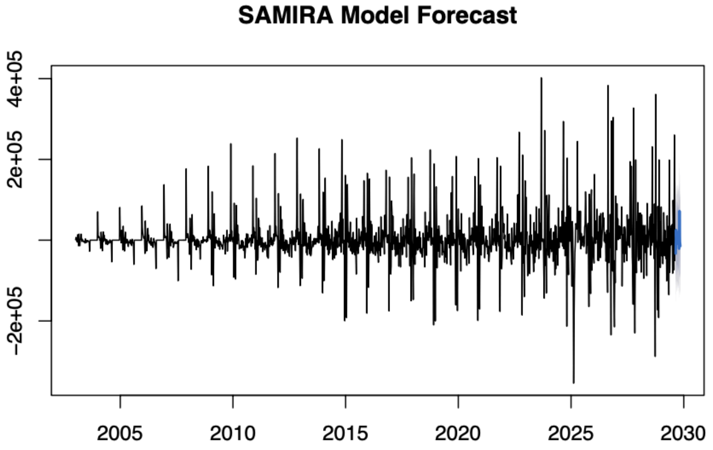
We can see that although the trend is forecasted with some accuracy, the seasonal component is visibly missing. For comparison the additive and multiplicative models were both used to make predictions as well, refer to appendix, figures 5 and 6 for the plots. The multiplicative Holt-Winters model does not seem to give an accurate forecast, however, the additive model does seem to be much more accurate. From the AME calculated earlier, if we assume double exponential smoothing had a low AME by chance, then additive Holt-Winters had the lowest AME, aligning with the hypothesis from earlier and the plots.

For further analysis, we adjusted the datasets. The first analysis focused on pre-COVID data, from 2003 to 2019. We split the training set from 2003 to the end of 2018 and the testing set from 2019 onwards. The second analysis used post-COVID data, from 2019 onwards, with the training set covering 2019 to the end of 2022 and the testing set from 2023 onwards. Both approaches had an additive Holt-Winters model that provided the lowest AME. These results suggest, that the additive Holt-Winters model consistently performs well.

2.4 Box-Jenkins Models

For the Box-Jenkins models, we began by examining the differenced data and found that after a single time differencing, the data became stationary. Using R’s built-in functions and additional libraries, we fitted ARIMA and SAMIRA models to the differenced data. Upon checking the residuals of both ARIMA models, refer to appendix, figure 7, we observed that there was no discernible trend, but the ACF plot revealed significant periodicity. The residuals seemed to follow a normal distribution, though with a much larger mean. From appendix, figure 8, this was also the case with the SAMIRA model. The following are the forecasts of the ARIMA and SAMIRA models respectively.


AMIRA Model Forecast



Since our data is seasonal as for the flu conditions, we only did one ARIMA model and fitted 5 SAMIRA models with theoretically better performance. These five models were chosen to provide a range and balance between AR, MA, and seasonal components, ensuring both non-seasonal and seasonal patterns are well-represented. This selection ensures a robust comparison to identify the model that best fits the time series while avoiding overfitting. After fitting the five models, we observe that Model 3 (SARIMA(1,0,1)x(0,1,1)[12]) has a balance of AR, MA, and seasonal MA terms, making it versatile without being overly complex. Additionally, in most cases, it tends to provide a better AIC value of 1018.5 compared to simpler models. Lastly, Model 3 often has well-behaved residuals, indicating it effectively captures both the seasonal and non-seasonal patterns in the data, making it the best model among the 5.

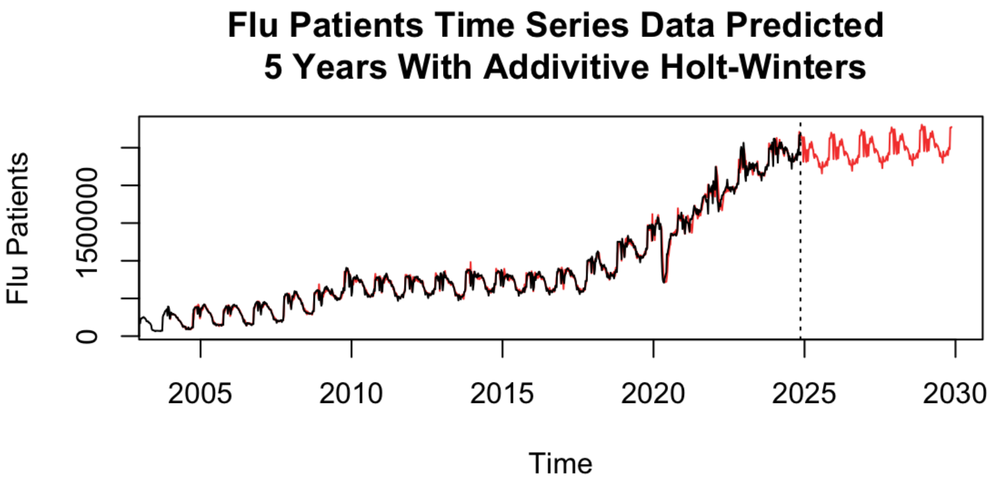
Further analysis with AR, MA, and ARMA models revealed that all models (AR(1,0,0), MA(0,0,1), and ARMA(1,0,1)) had significant p-values (< 2.2e-16), indicating that their residuals were not white noise, suggesting the models did not fully capture the data's structure. The AIC values for these models were very close, with the MA(0,0,1) model having the lowest AIC (35046.22), followed by ARMA(1,0,1) and AR(1,0,0). However, the differences in AIC were minimal, and the residual plots showed structure, indicating that the models did not adequately explain the data. Significant spikes in the ACF of the residuals further confirmed that autocorrelation remained unaccounted for. As a result, these models were deemed unsuitable for modelling our data.

3. Statistical Analysis

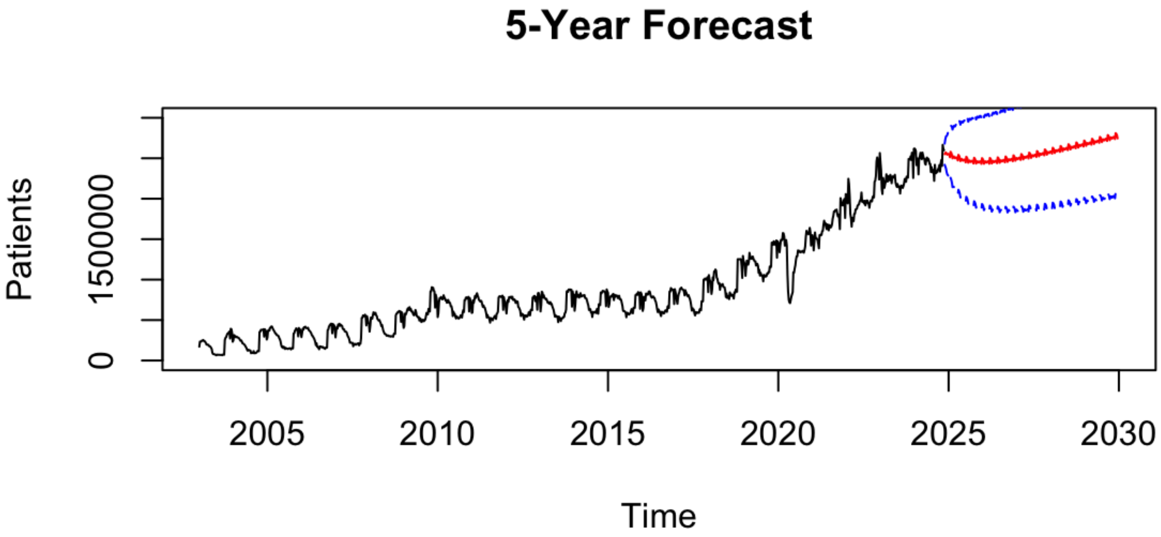
3.1 Theoretical and Result-Based Analysis

From all the models we ran, the best models, based on residual diagnostics and the lowest AME, were the additive Holt-Winters process and the SAMIRA models. This aligns with our theoretical expectations, as Holt-Winters is particularly effective for data exhibiting both trend and seasonality. The additive model performed better than the multiplicative version, likely due to the constant nature of seasonality in the flu data. Additionally, the SAMIRA model extends the ARIMA model by incorporating seasonality, which allows it to better capture the underlying patterns in the data.

3.2 Forecasting the Next 5 Years with Additive Holt-Winters



3.3 Forecasting the Next 5 Years with SAMIRA



4. Conclusion

To conclude, we make an inference based on the five-year forecast of our best two models. From the additive Holt-Winters model, we observe a slight, almost negligible linear increase in the total number of patients, with the same seasonal patterns as before. This suggests that post-COVID, the structure of the time series remains largely unchanged to pre-COVID, except for a significant upward shift in the mean. Clinically, this indicates an increased baseline of flu cases compared to pre-COVID levels, but no further sharp increases are expected. The SAMIRA model provides a similar outlook, showing a linear upward trend following a slight dip but with less significant seasonality. Together, these forecasts imply that while the flu case baseline has shifted higher, the overall dynamics of flu trends and seasonality remain stable.

5. Appendix

Figure 1: Residual vs Fitted Values of Optimal Unregularized Regression Model

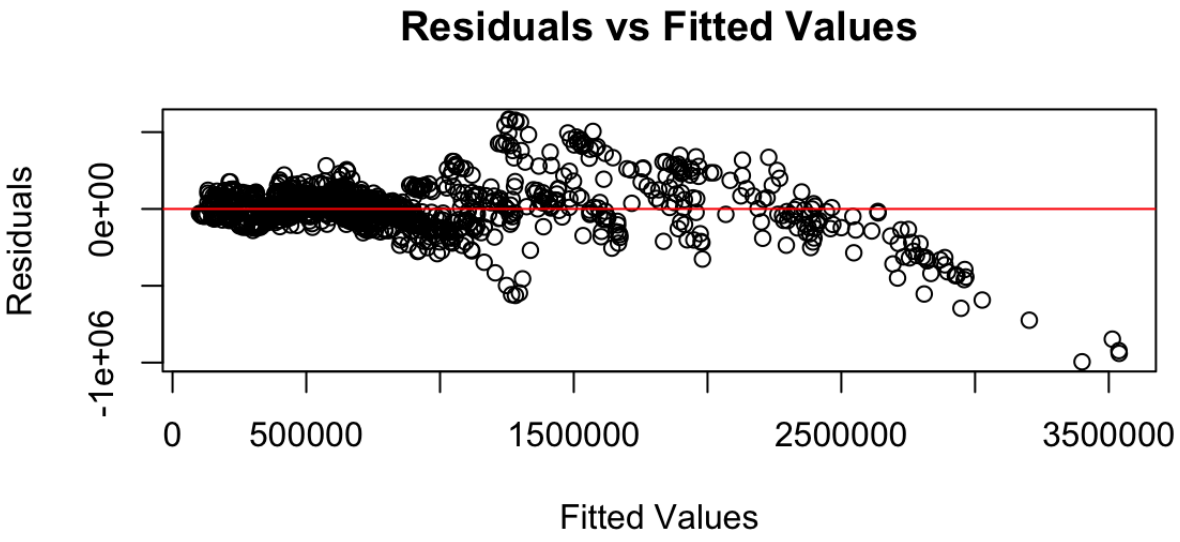


Figure 2: Observed Values vs Fitted Values of Optimal Regularized Regression Model

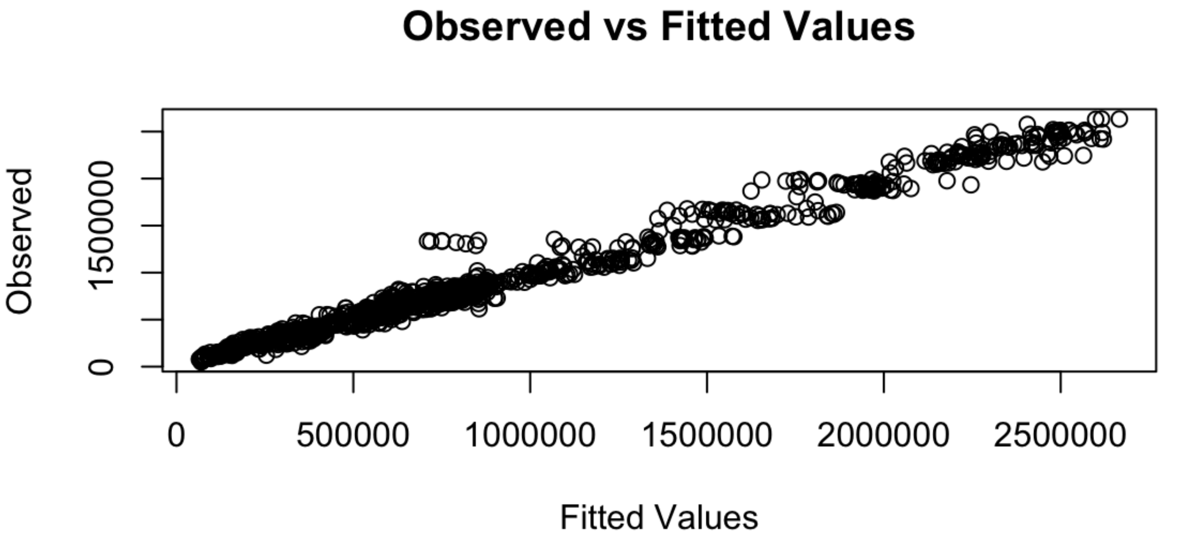


Figure 3: Residuals vs Time of Optimal Regularized Regression Model

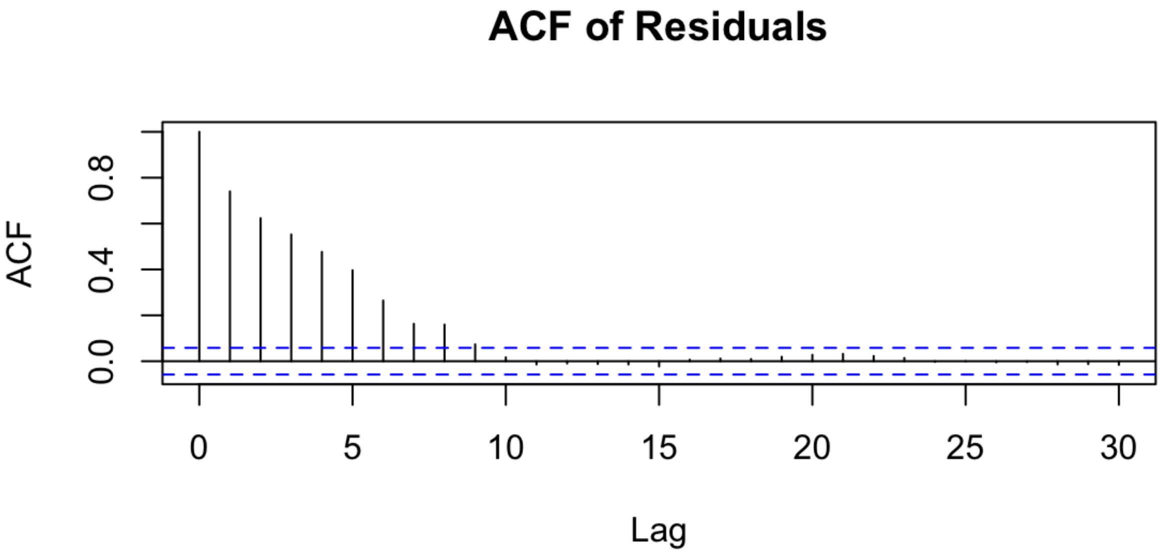


Figure 4: ACF for Double Exponential Smoothing

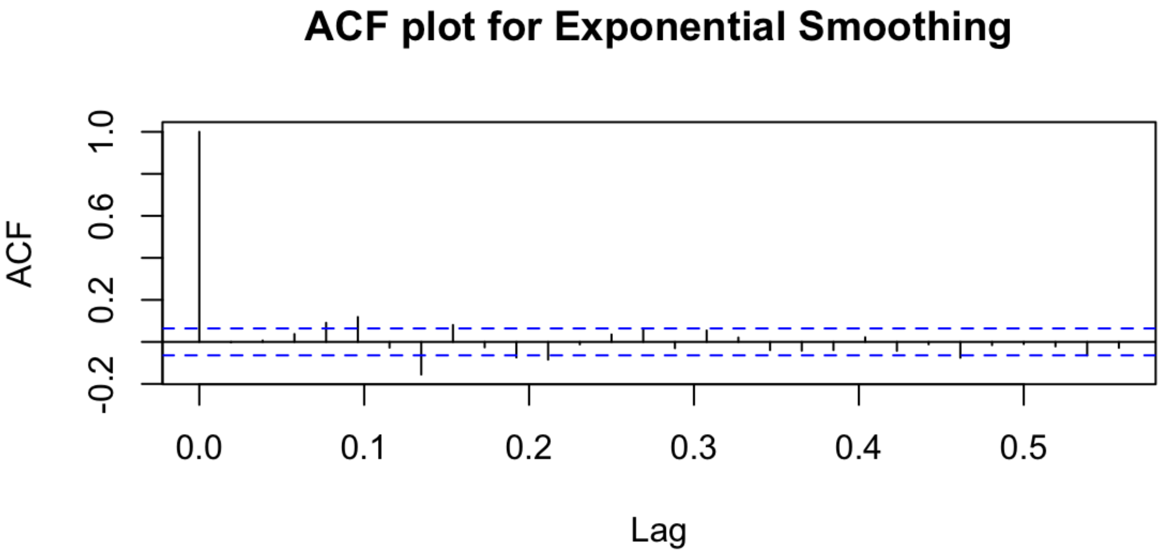


Figure 5: Additive Holt-Winters 2025 Prediction

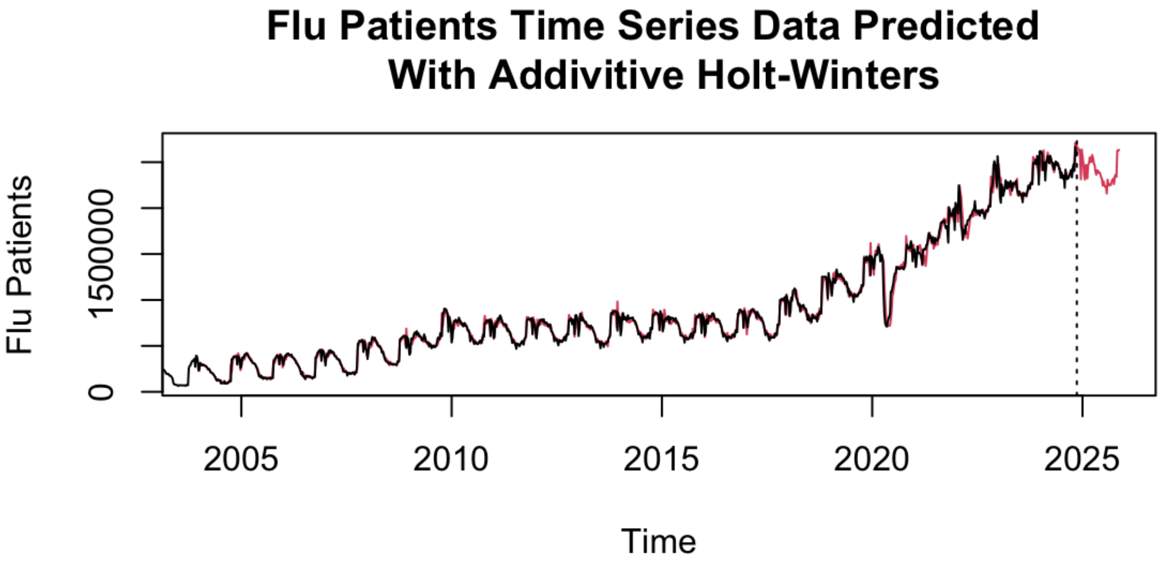


Figure 6: Multiplicative Holt-Winters 2025 Prediction

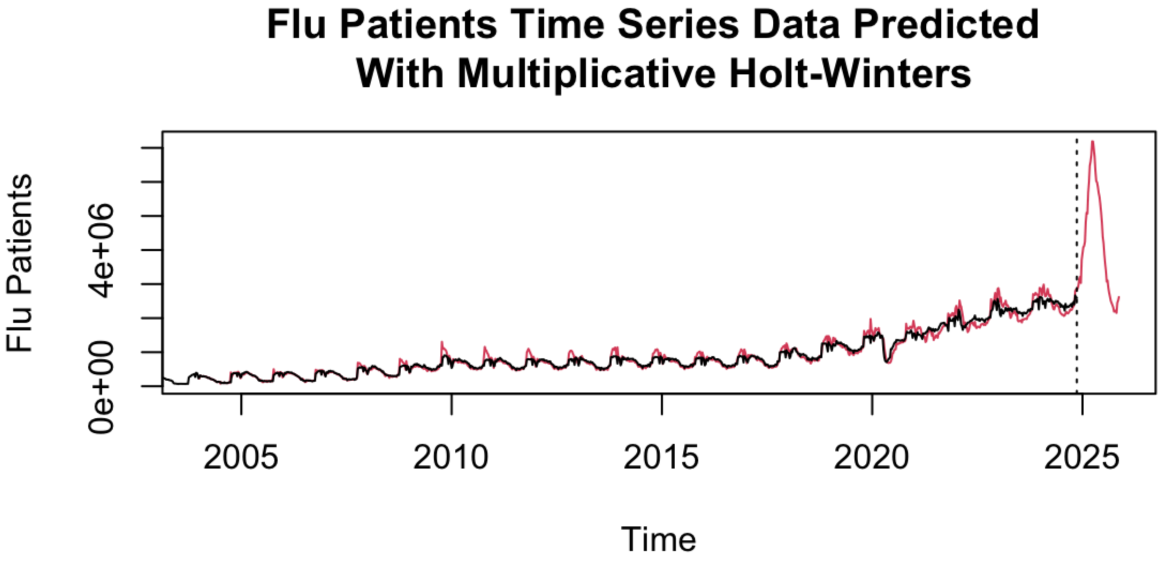


Figure 7: Residuals of ARIMA Model

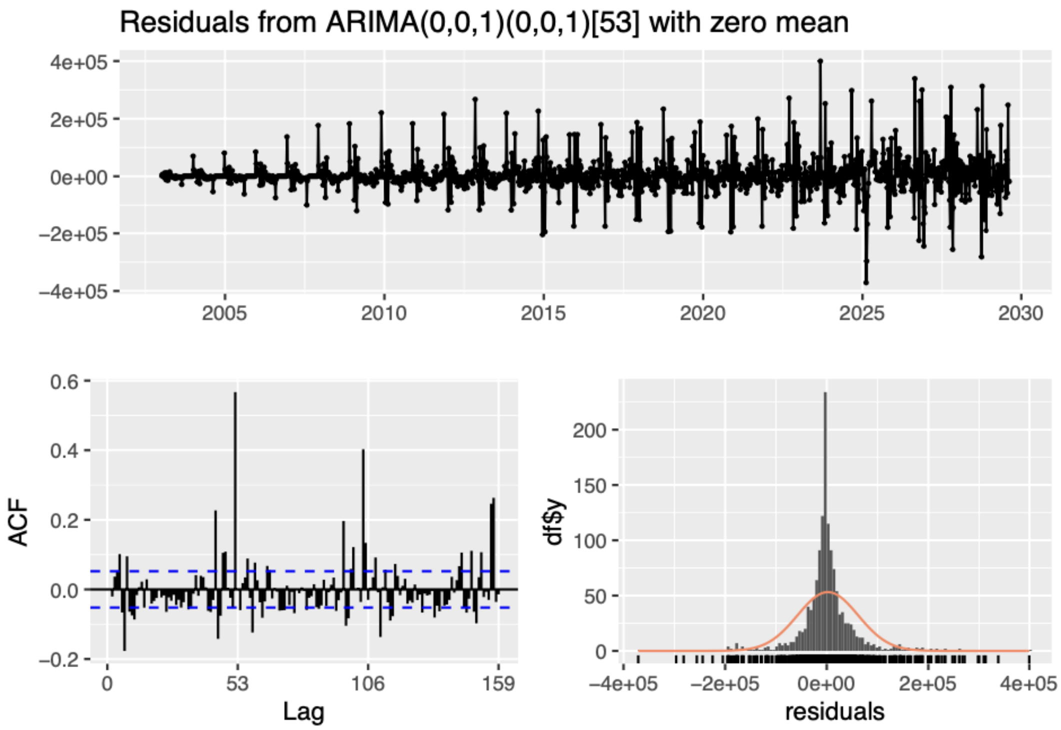


Figure 8: Residuals of SARIMA Model

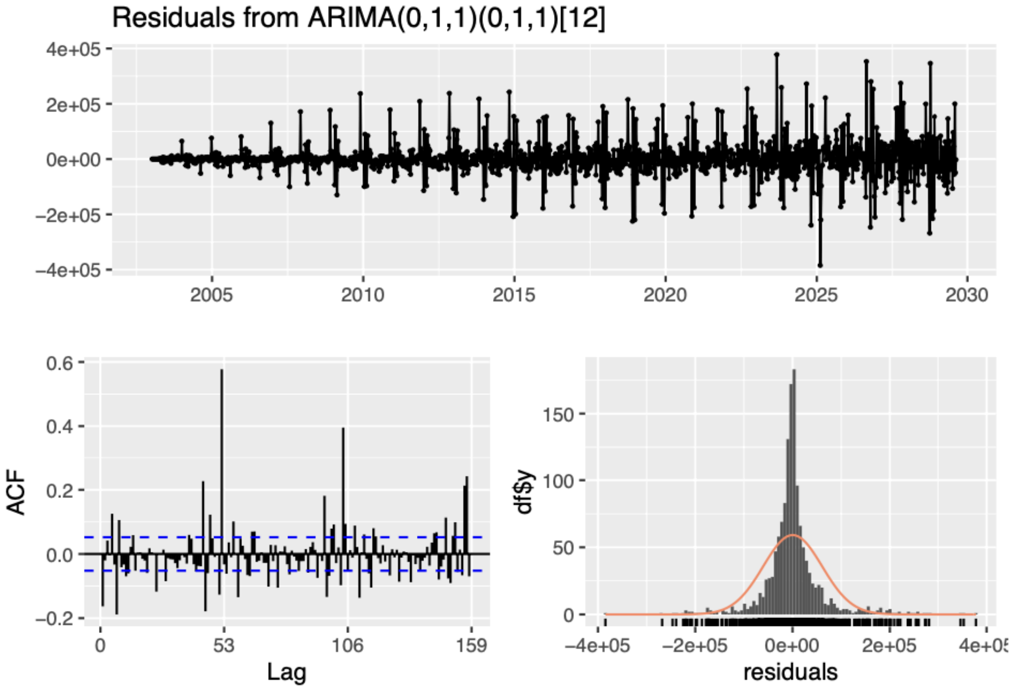
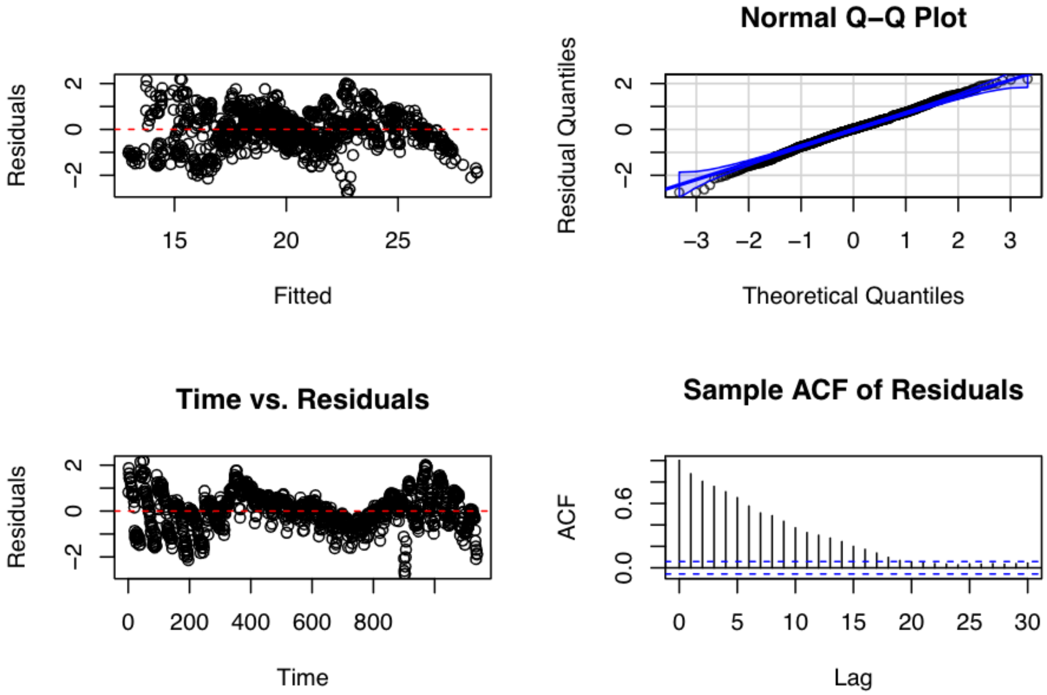


Figure 9: Residuals for Regularized Regression Model


Residuals
Fitted
Normal Q−Q Plot
Residual Quantiles
Theoretical Quantiles
Time vs. Residuals
Residuals
Time
Sample ACF of Residuals
ACF
Lag