STAT 3504 Final Project Report

U.S. Unemployment Time Series Data Analysis

Adara Morganstein

Massimo Camuso

Temple University

Abstract

The unemployment rate has long been utilized as an indicator of the state of the American labor market, as well as a gauge of the economy's health overall. Analysis of the unemployment rate over time can lead to insights into past economic trends and help to predict unemployment spikes and recovery patterns. Additionally, its measure through time serves as an interesting time series to analyze through statistical modeling.

The data set analyzed for this report includes seasonally adjusted U.S. unemployment rate data, retrieved from Kaggle and sourced from the Federal Reserve Economic Data (FRED) database. Through analysis we determined this time series is best modeled using an ARIMA (1,1,1) model. We checked model diagnostics and made sure residuals were distributed with uniform randomness, assuring confidence in our model's fit. Finally, we then used our model to forecast 6 additional points and compared them to the next 6 months' rates logged in FRED.

Introduction

FRED, an online database of economic time series data, was created in the early 1990's by the Federal Reserve Bank of St. Louis. It has been maintained by the same institution since. The analyzed data set logs unemployment rate as percentage of the labor force, qualifying members of the labor force as people aged 16 and older, currently residing in 1 of 50 states or Washington D.C., who do not reside in institutions (prison, elderly care, etc.), and who are not currently active duty in the military. This study analyzes the time series data set of this unemployment rate, logged once per month starting on January 1, 1948, and ending May 1, 2024. The data has been seasonally adjusted to account for seasonal trends. This series is hosted in FRED but the data itself is retrieved from the U.S. Bureau of Labor Statistics.

The primary objective of this report is to demonstrate the application of statistical tools and techniques in model building and evaluation for time series data. We utilized R to import the data, convert it to a time series object, and analyze it to build an appropriate model. Analysis techniques included plotting ACF and PACF, testing the necessary number of differences, and fitting an ARIMA model. The fit of the model was then evaluated utilizing its summary in R, as well as testing AIC and checking residual distribution.

Model Specifications

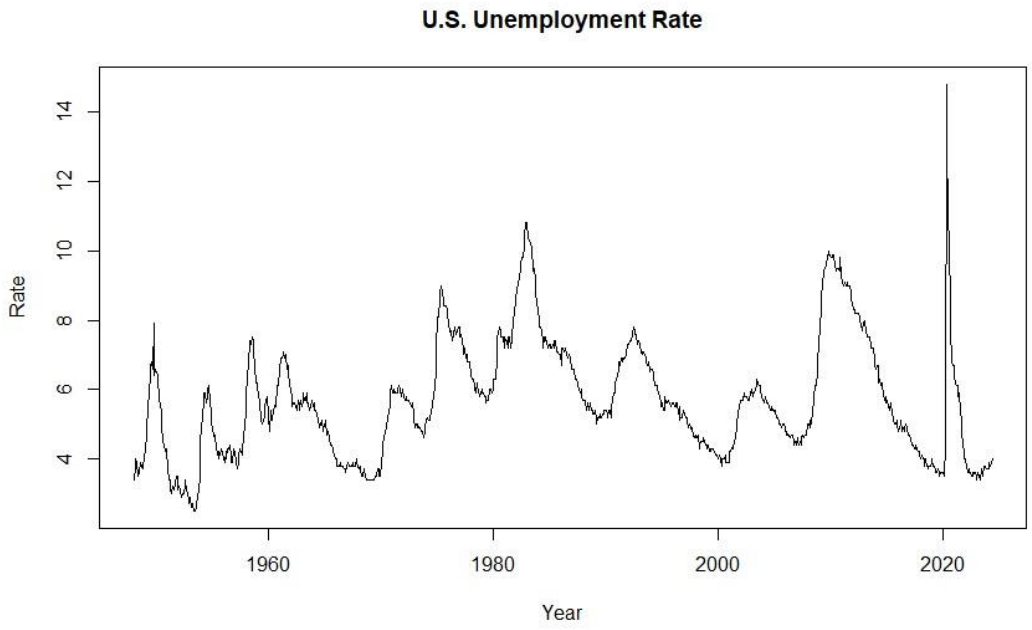

Figure 1: U.S. Unemployment Time Series

To begin specifying a model to our data, we ran several tests in R. We first plotted ACF and PACF and observed their patterns. The ACF plot (Figure 2) shows gradual decay, indicating autoregressive behavior is present. The PACF plot (Figure 3) confirms this, showing a sharp spike at lag 1 and cutting off after, indicating one autoregressive term is needed.

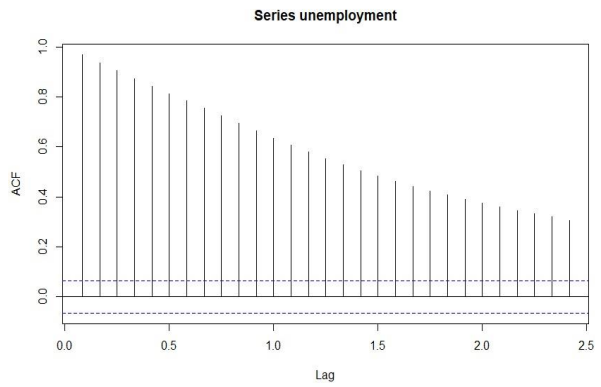

Figure 2: ACF Plot

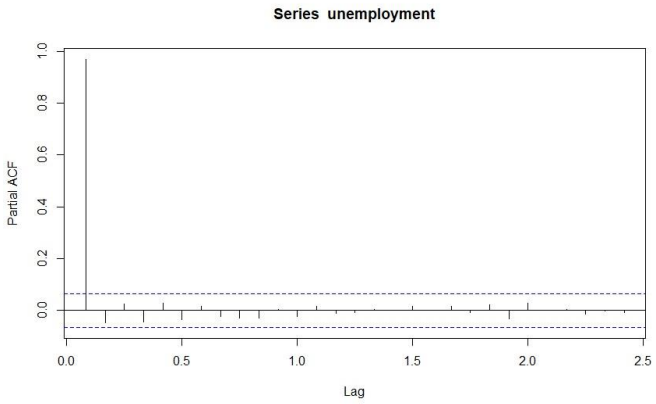

Figure 3: PACF Plot

We also ran an ndiffs() test in R to determine the number of differences required to make the series stationary, which returned a value of 1. From here, we tried fitting both an ARIMA (1,1,1) model as well as an ARIMA (1,1,0) model, and compared AIC values for both. Although the ACF plot does not have a drop indicating need for an MA term, the AIC for the ARIMA (1,1,1) was smaller by about 4, so the ARIMA (1,1,1) model fits just slightly better. This was confirmed by using the auto.arima() function in R, which also indicated that ARIMA (1,1,1) would provide the best fit to the data.

Model Fitting and Diagnostics

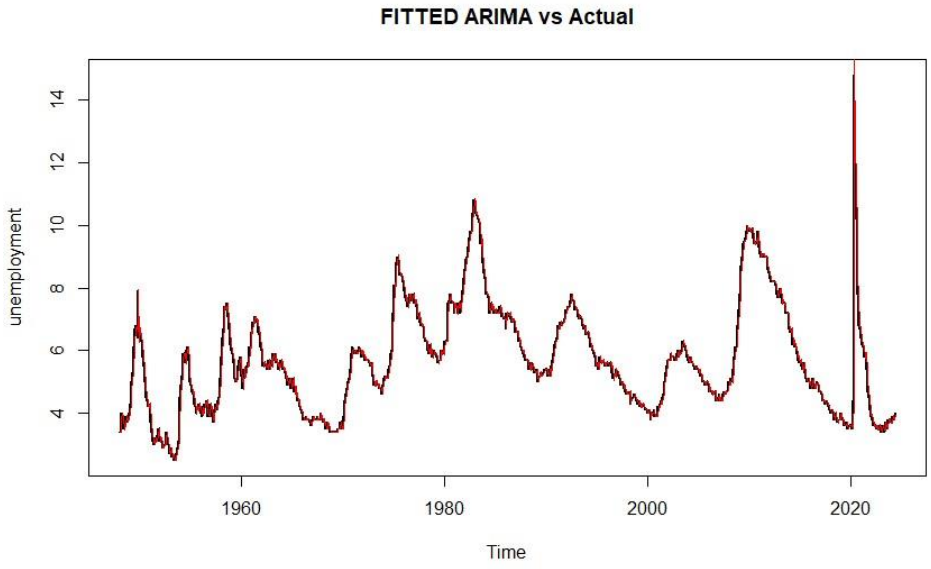

Figure 4: ARIMA (1,1,1) vs. Actual Data (red)

We plotted our ARIMA (1,1,1) model against the real data values to visually evaluate the fit. The AIC for this model is 1002.35. Coefficients and standard error for the AR and MA terms are pictured in the appendix. For model diagnostics, we performed a Ljung-Box test and evaluated residual distribution. P-value for the Ljung-Box test was 0.999, so we accepted the null hypothesis and concluded that there was no statistically significant autocorrelation of the residuals.

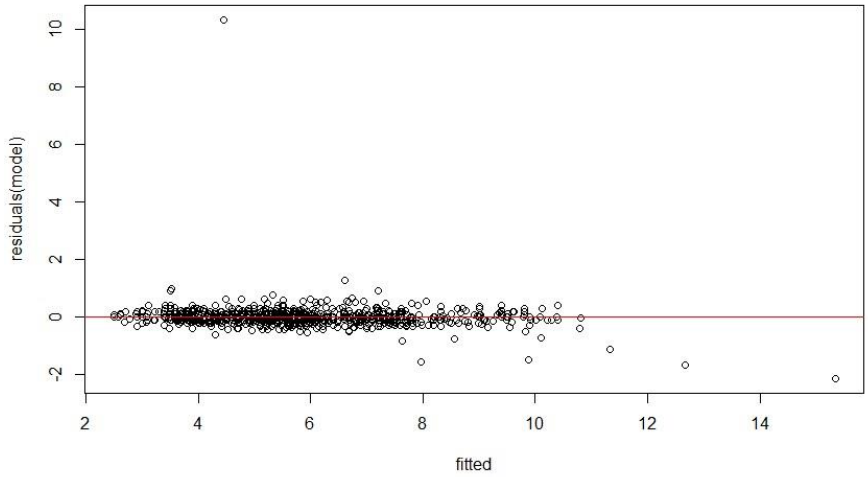

Figure 5: Residuals Plot

To further test the fit of the model, we evaluated the distribution of the residuals. For the most part, residuals are randomly distributed around 0 (Figure 5) besides one significant outlier with a residual of 10.35. This residual is from the difference between prediction and actual value for April 2020. It is significant to note that this was the first measurement taken after the start of the Covid-19 pandemic, which may have had an effect. The next largest absolute value of a residual is 2.14, which is considerably less.

Forecasting

Since the dataset we chose from Kaggle stopped recording the FRED unemployment data in May, we used our model to forecast the unemployment rates from June to November and compared the forecasted rates to the actual recorded unemployment rates. In our “Forecasted Unemployment Rates June – November" plot (Figure 6), we overlaid our models forecasted unemployment rates over the actual recorded unemployment rates for the months of JuneNovember. The biggest discrepancy between our model's forecasted unemployment rate and the actual unemployment rate occurred in July, when our model predicted an unemployment rate of 3.9 and the actual unemployment rate was 4.3. For every other month, our model was never off by more than 0.2, and we were able to successfully capture the true unemployment rate for each month within the 80% confidence intervals provided in our R output (see Figure 9, appendix).

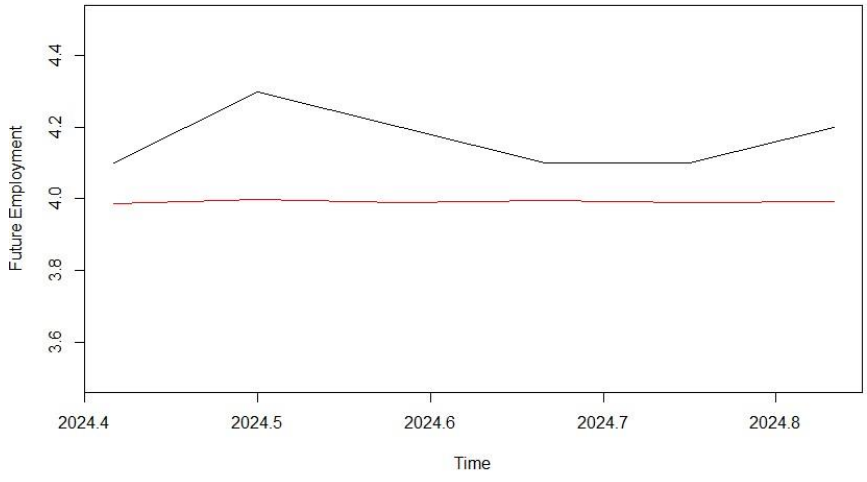

Figure 6: Forecasted Unemployment Rates, June – November (red)

Conclusion and Discussion

In this report, we analyzed and explored several model building and evaluation techniques to fit a model to a time series data set. The time series data was U.S. unemployment rates logged monthly between 1948 and 2024, acquired via Kaggle from the Federal Reserve Economic Data database. Through exploratory data analysis, we determined that the model with the best fit had one autoregressive term, one moving average term, and one level of differencing. We then evaluated the fitted ARIMA (1,1,1) model using its AIC, a Ljung-Box test, and residual plotting to confirm its goodness-of-fit.

Our model's residuals did include an outlier for the data point taken April 1, 2020. Although an educated guess predicts correlation between this large residual and Covid-19 pandemic, further investigation is recommended to confirm this. Additionally, we recommend evaluating more outside variables to further improve model fit. Currently, the model relies solely on historical unemployment data, but outside data about inflation rates or GDP growth may also serve to inform the unemployment rate.

References

SD, G. (2024, June 10). *U.S. unemployment rates*. Kaggle.

https://www.kaggle.com/datasets/guillemservera/us-unemployment-rates

Unemployment rate. FRED. (2024, December 6). https://fred.stlouisfed.org/series/UNRATE

Appendix

```
Series: unemployment
ARIMA(1,1,1)

Coefficients:
          ar1     ma1
      -0.7762  0.8322
s.e.   0.0972  0.0846

sigma^2 = 0.1741:  log likelihood = -498.17
AIC=1002.35   AICc=1002.37   BIC=1016.81
```

Figure 7: ARIMA (1,1,1) auto.arima() output

```
        Ljung-Box test

data:  Residuals from ARIMA(1,1,1)
Q* = 6.9725, df = 22, p-value = 0.999

Model df: 2.   Total lags used: 24
```

Figure 8: ARIMA (1,1,1) Model Residuals Ljung-Box Test

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Jun 2024	3.985798	3.451035	4.520562	3.167948	4.803649
Jul 2024	3.996822	3.219124	4.774521	2.807435	5.186210
Aug 2024	3.988265	3.040644	4.935887	2.539003	5.437527
Sep 2024	3.994908	2.894457	5.095358	2.311913	5.677902
Oct 2024	3.989751	2.761523	5.217980	2.111338	5.868165
Nov 2024	3.993754	2.645400	5.342108	1.931624	6.055884

Figure 9: Model Forecast Output, 6 Values

Project Contributions

Massimo – Model testing/selection, model diagnostics, and “Appendix” and “Conclusion” body paragraphs

Adara – data selection, forecasting, graphs, “Abstract” and “Introduction” and body paragraphs