STA457 Final Project

Yasmine Hemmati

1 Abstract

In this report, an in-depth analysis of monthly unemployment is conducted of monthly unemployment in the US using historical data from 1948-1978. We explored the U.S unemployment time series data, transformed the data to convert it to a stationary process by applying the box-cox transformation and differencing, then selected a SARIMA model by performing necessary diagnostics. After selecting our final model, we will use this model to forecast the data 10 months into the future. Our results indicated that monthly unemployment will have a slight upward trend for 10 months after 1978. We conclude that we will see an overall rise in monthly unemployment just under a year after 1978.

Keywords: time series, U.S unemployment, spectral analysis, forecasting, seasonal ARIMA

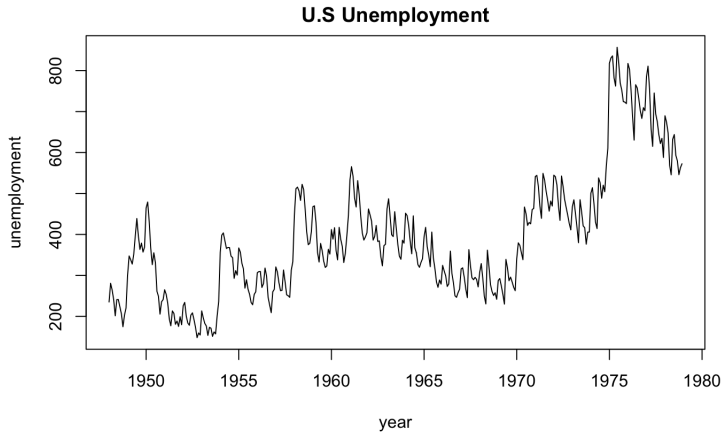
2 Introduction

A country's unemployment rate is one of the best-known labor market measures [1] and provides insights into the health of a country's economy. Forecasting unemployment is typically a difficult task for policymakers. High unemployment tends to hurt growth in labor productivity, a country's Gross Domestic Product (GDP) and is linked to higher rates of poverty, homelessness, income inequality and crime [2]. Various factors can cause an increase in unemployment, such as an economic recession, demographic trends such as an aging population, technological advancements that automate jobs, globalization with employers sending jobs overseas to reduce labor costs, etc [3]. The Covid-19 pandemic we are currently in left 10 million Americans jobless in its first two weeks [4]. The data we will use to construct our model is the "unemp" time series data from the R package astsa. This data contains monthly U.S unemployment from 1948-1978 and has 372 observations. This report aims to find a model to forecast US unemployment. In this report, we will explore the US unemployment time series data and perform various diagnostics to find a model that can forecast future monthly unemployment.

3 Statistical Methods

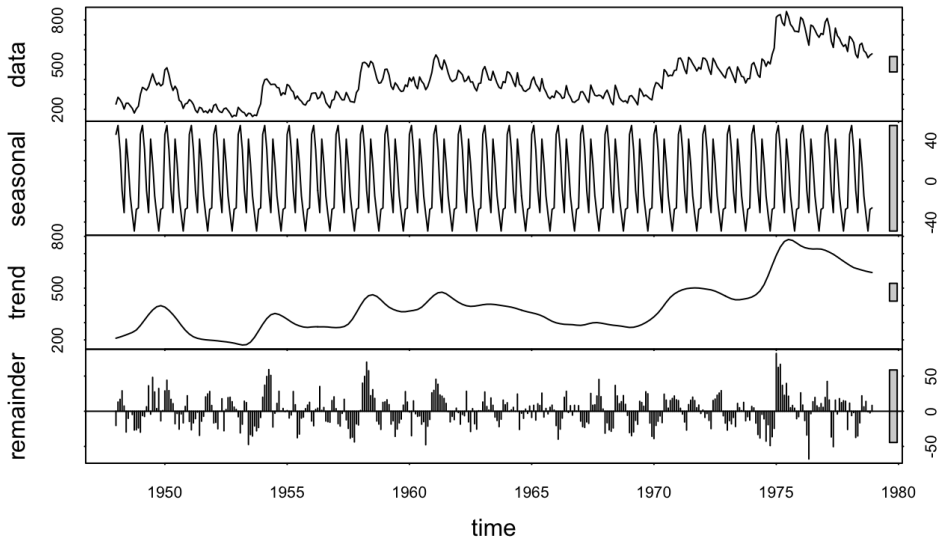
3.1 Data Visualizations

First we will visually inspect the time series plot of the U.S monthly unemployment time series data.



We observe that this time series is not stationary, as the mean and variance appears to change with time. Roughly constant peaks is also an indication of hints of seasonality within the data. We see an upward trend in unemployment around 1975, and this can be associated with the 1973–1975 recession that was caused by the 1973 oil crisis and a series of economic measures taken by President Nixon in response to inflation [7]. We will do further analysis to check whether or not the time series is seasonal and stationary.

To check if there is a trend and seasonality in our data, we can use the stl() function to decompose the data into its trend and seasonal components. Looking at the trend and seasonality components, we see a trend and yearly seasonality within the data. Moreover, this decomposition shows an upward trend in unemployment and suggests we must remove this upward trend.

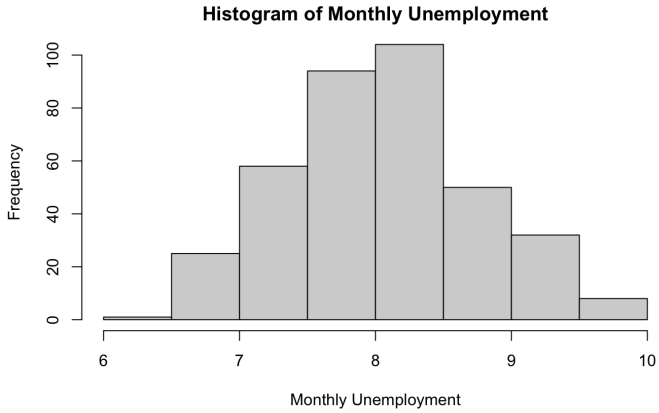


Moreover, we can also plot the histogram of monthly unemployment and see that the data is not normally distributed.

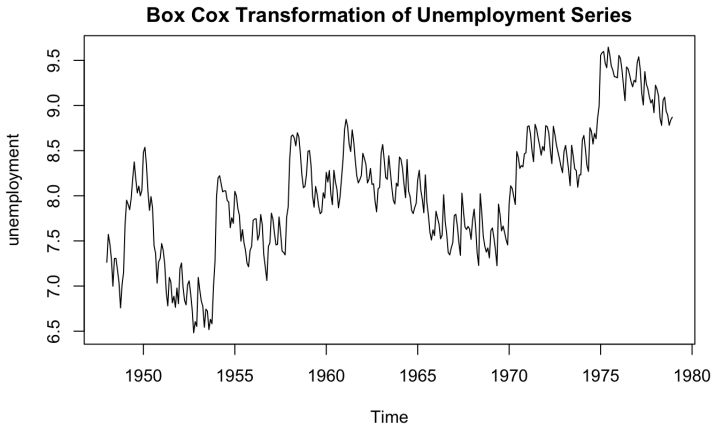
3.2 Data Transformations

3.2.1 Box-cox transformation

We will use the Box-cox transformation to stabilize the variance and transform our data to resemble a normal distribution. Using the boxcox() function, the data can then be transformed using the following equation $y_i^{(\lambda)} = \frac{y_i^{\lambda}-1}{\lambda}$. To find the optimal λ value, I plotted the log-likelihoods corresponding to λ values ranging from -2 to 1 and found the maximum occurs when λ is approximately 0.1. After applying the box-cox transformation, the histogram looks as follows. We can see the data resembles more of a normal distribution now.

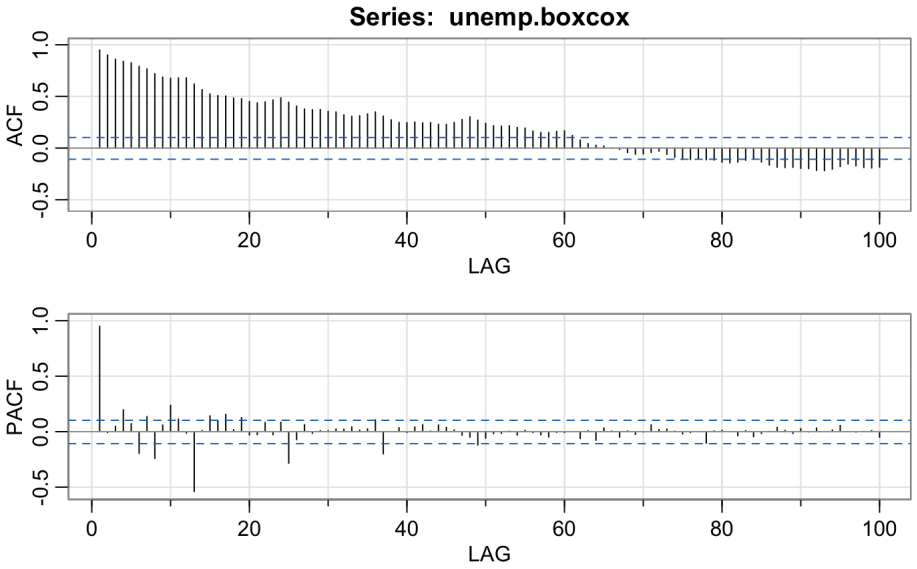


Moreover, the variance of the data prior to applying the box-cox transformation is 24040 and the variance after we applied the box-cox transformation is 0.4858! We can also visualize the data after applying the box-cox transformation.

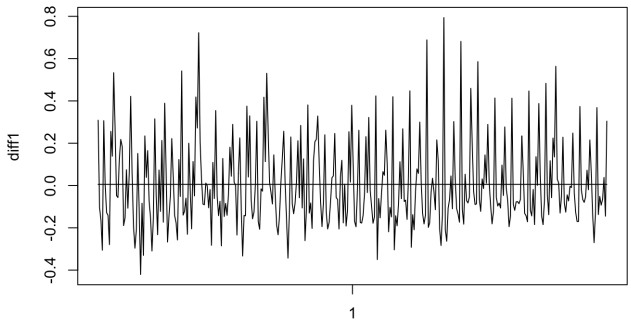


3.2.2 Differencing

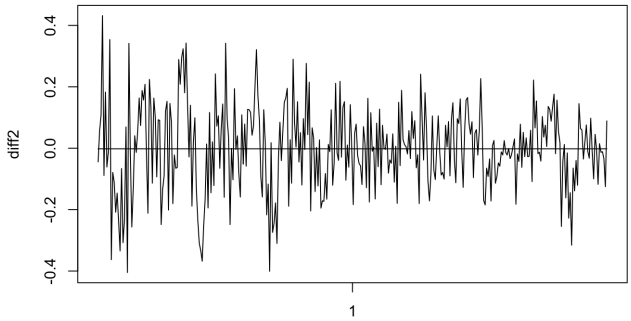
After applying the box-cox transformation, we see that the sample ACF decays slowly to zero as h increases. This slow decay in the sample AFC is an indication that differencing is needed. Moreover, we can see hints of seasonal behavior in the ACF. Notice that at the end of each year, hence lags that are multiples of 12, there is a small rise in the ACF, implying a seasonal pattern.



We begin by taking the first difference. The variance after taking the first difference is 0.04217187. Using the monthplot() function, we can plot the month plot of data after taking the first difference:



We appear to have achieved a stable process that is stationary; there do not appear to be trends or seasonality present. To affirm the stationary of the transformed data, we will apply the Augmented Dickey-Fuller test for stationary hypothesis testing to obtain a p-value of 0.01. Thus, we accept the null hypothesis that we have a stationary distribution. Recall that after decomposing the data, we notice seasonal behavior similar to what we've seen previously. Moreover, from the decomposition and form of the PACF, as shown previously, we notice seasonal behavior. So we will try taking the seasonal difference to see if we can reduce the variance further. After taking the seasonal difference as well, the variance is now 0.01944! Using the monthplot() function, we can plot the month plot of the data after taking the first difference and then the seasonal difference:

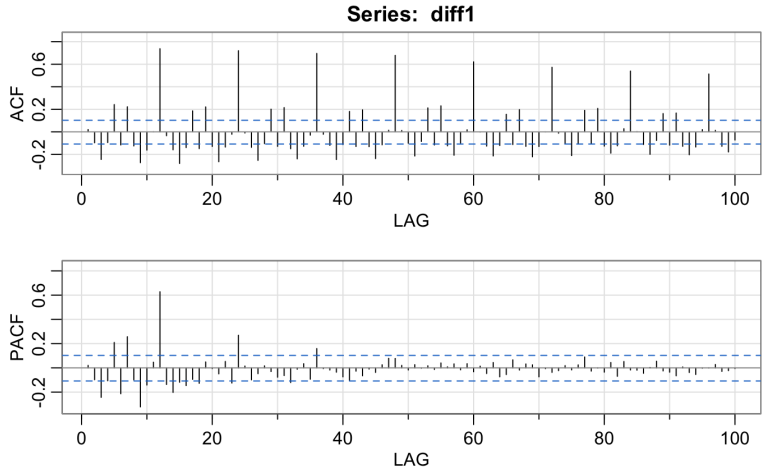


The series appears stationary from the plot. To affirm the stationary of the transformed data, we will apply the Augmented Dickey-Fuller test for stationary hypothesis testing and obtain a p-value of 0.01. Thus, we accept the null hypothesis that we have a stationary distribution. Therefore, we get a stationary process by applying the box-cox transformation and then taking the first difference. Furthermore, we also get a stationary process by taking the box-cox transformation and then taking the first and seasonal differences.

4 Results

4.1 Proposed Models

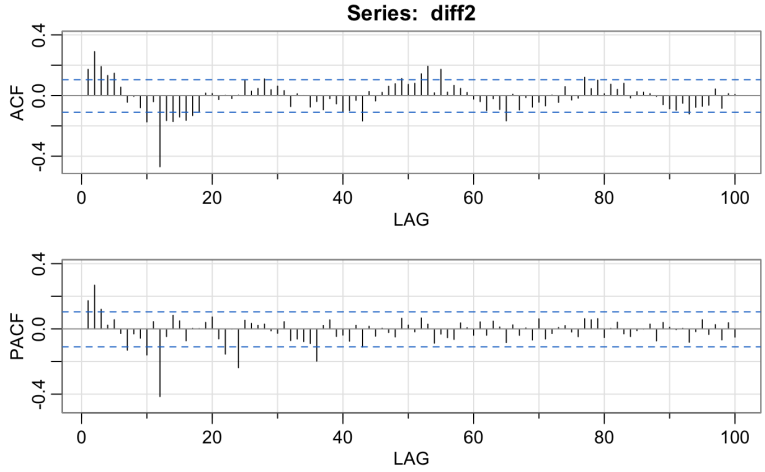
ARIMA models do not account for seasonal trend; hence SARIMA models may be a better fit for this data. Our aim now is to find an appropriate (S)ARIMA model based on the ACF and PACF of the differenced data. Taking the first difference only of the transformed data gives us a stationary distribution, and taking the first difference and then the seasonal difference gives us a stationary process. Thus, we will use both ACF and PACF plots to help us find the best model. We will begin by plotting the sample ACF and PACF of the first differenced data:



Nonseasonal: We notice that there is a significant spike at lag 2 in the ACF and PACF. This suggests an MA(2) and AR(2) component.

Seasonal: The significant spike at lag 12 in the ACF suggests a seasonal MA(1) component (Q = 1). In this case, we only took the first difference, so we have d=1 and D=0.

Therefore a candidate model we get from this analysis is the ARIMA$(2,0,2)\times(0,1,1)_{12}$ model. We will now inspect the ACF and PACF plot of the data after taking the seasonal difference, difference lag 12, and first difference, the difference lag 1.



Non seasonal: The PACF cuts off lag 3 while the ACF tails off, suggesting an AR(3) component. There is also a significant spike in the ACF and PACF at lag 2, suggesting we should also try an AR(2) component and include an MA(2) component in the model.

Seasonal: The significant spike at lag 12 in the ACF suggests a seasonal MA(1) component (Q = 1). In this case, we took both the seasonal and first difference, so we have d = D= 1.

Therefore, as part of our preliminary analysis, we will fit the following models as well: ARIMA$(3,1,0)\times(0,1,1)_{12}$, ARIMA$(3,1,2)\times(0,1,1)_{12}$, ARIMA$(2,1,2)\times(0,1,1)_{12}$.

Note that we will fit the models without the last 10 observations and use the last 10 observations to test the final model. I will now check to see which of the four models we have so far have parameter estimates that are statistically significant. Below you will find the estimated parameters for each of the proposed models.

ARIMA[2,0,2][0,1,1][12]

	Estimate	SE	t.value	p.value
ar1	1.8248	0.0640	28.4960	0.0000
ar2	-0.8419	0.0617	-13.6370	0.0000
ma1	-0.8059	0.0821	-9.8179	0.0000
ma2	0.1265	0.0575	2.1988	0.0286
sma1	-0.6905	0.0479	-14.4017	0.0000
constant	0.0053	0.0029	1.8352	0.0673

ARIMA[2,1,2][0,1,1][12]

	Estimate	SE	t.value	p.value
ar1	1.8398	0.0771	23.8613	0
ar2	-0.8948	0.0705	-12.6920	0
ma1	-1.7356	0.0952	-18.2247	0
ma2	0.8096	0.0837	9.6700	0
sma1	-0.7015	0.0459	-15.2712	0

ARIMA[3,1,2][0,1,1][12]

	Estimate	SE	t.value	p.value
ar1	1.6569	0.1166	14.2045	0.0000
ar2	-0.5303	0.1690	-3.1374	0.0019
ar3	-0.1451	0.0641	-2.2624	0.0243
ma1	-1.6347	0.1112	-14.7022	0.0000
ma2	0.6347	0.1107	5.7317	0.0000
sma1	-0.6869	0.0481	-14.2824	0.0000

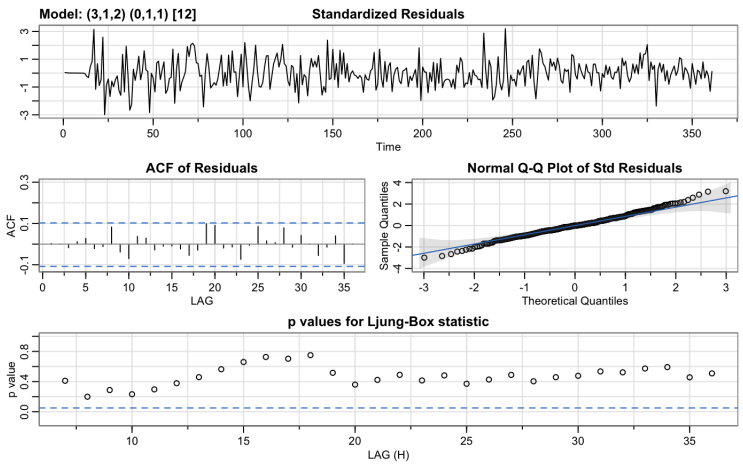
ARIMA[3,1,0][0,1,1][12]

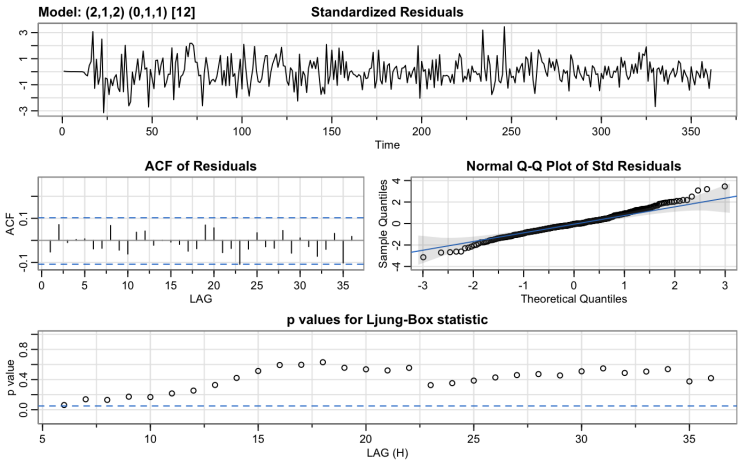
	Estimate	SE	t.value	p.value
ar1	0.0757	0.0538	1.4070	0.1603
ar2	0.1959	0.0529	3.7011	0.0002
ar3	0.0934	0.0538	1.7339	0.0838
sma1	-0.7143	0.0454	-15.7302	0.0000

For the ARIMA$(2,0,2) \times (0,1,1)_{12}$ we see that the constant parameter estimate has a p-value of 0.0673, which is greater than $\alpha = 0.05$, hence we cannot be confident that this parameter is statistically significant. Therefore we will scrap this model. For the ARIMA$(3,1,0) \times (0,1,1)_{12}$ model, we see that the ar3 parameter has a p-value of 0.0836 and the ar1 parameter has a p-value of 0.1603. Therefore these parameter values are greater than $\alpha = 0.05$. Since the p-value is above our significance threshold, we cannot claim that this parameter is non-zero and statistically significant. Therefore, we will scrap this model as well. As we can see, the p-values of all parameter estimates are less than $\alpha = 0.05$, hence are statistically significant for the ARIMA$(2,1,2) \times (0,1,1)_{12}$ and ARIMA$(3,1,2) \times (0,1,1)_{12}$ models. We will do further analysis of these two models so we can select the best one.

4.2 Model Diagnostics and Final Model

We will perform further model diagnostics on the candidate models to select the best model.

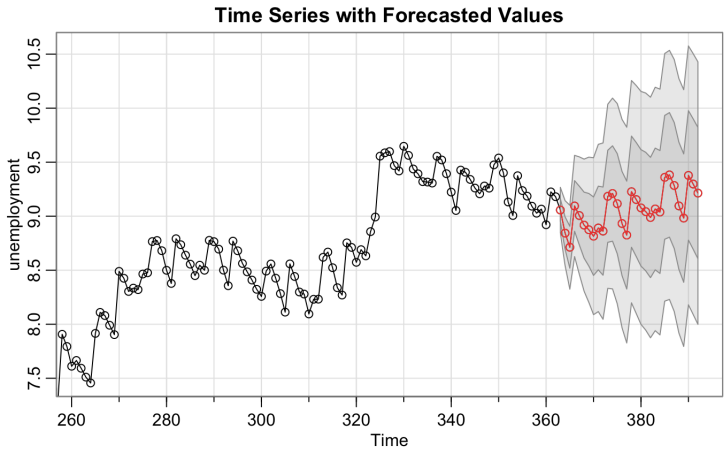




The standard residuals of both models show no pattern, indicating that the residuals are independent. Thus, normal Q-Q plot of the residuals shows that the assumption of normality is reasonable, with the exception of the possible outliers. There are a few outliers exceeding three standard deviations of the mean. We also see that there are significant spikes in the ACF residual plots for the ARIMA$(3,1,2)\times(0,1,1)_{12}$ and ARIMA$(2,0,2)\times(0,1,1)_{12}$ models but it's not quite enough to be significant at the 5% level for both models. Thus for both models, the ACF of the standardized residuals shows no apparent departure from the model assumptions. The p-values for the Ljung-Box statistics are all above a reasonable significance level for the majority of lags for all three models, which means we do not reject the null hypothesis that the residuals are independent. Hence, both models appear to fit well and satisfy model assumptions. We can also compare the AICc of both the ARIMA$(3,1,2)\times(0,1,1)_{12}$ and ARIMA$(2,0,2)\times(0,1,1)_{12}$ model to help us select our final model. We get the AICc value of the ARIMA$(3,1,2)\times(0,1,1)_{12}$ is -1.5944 and the AICc value of the ARIMA$(2,1,2)\times(0,1,1)_{12}$ is -1.574. The difference between the AICc values are very small. Given that we see that there are more p-values for the Ljung-Box statistics that are below a reasonable significance level for the ARIMA$(2,1,2)\times(0,1,1)_{12}$ and the AICc of this model is larger than the AICc value of the ARIMA$(3,1,2)\times(0,1,1)_{12}$ model, I will scrap this model. I will select the ARIMA$(3,1,2)\times(0,1,1)_{12}$ model as the final model.

4.3 Forecasting

Next, we will forecast the box-cox transformed data 10 months into the future, recalling that we omitted the last 10 observations of the data when we fit the model for testing purposes.



Our results are summarized in the table below, where the 95% confidence interval upper and lower bound are shown with the predicted or forecasted values alongside the actual values that were omitted for comparison purposes.

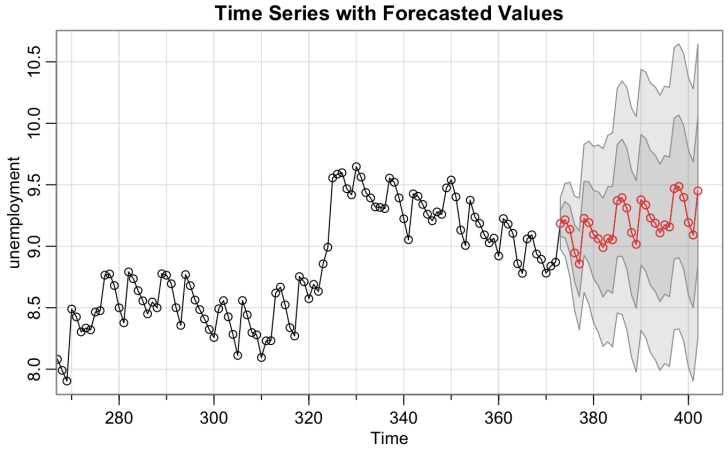
Prediction <dbl>	95% Lower Bound <dbl>	95% Upper Bound <dbl>	Actual <dbl>
9.057181	8.884602	9.229760	9.105165
8.843434	8.596285	9.090583	8.858014
8.712658	8.393433	9.031883	8.779992
9.094582	8.709100	9.480065	9.059562
9.006063	8.559502	9.452623	9.093041
8.917467	8.415240	9.419695	8.937070
8.874649	8.322223	9.427075	8.893844
8.814788	8.217596	9.411981	8.781024
8.890815	8.254140	9.527490	8.838361
8.861858	8.190747	9.532970	8.870245

As shown in the table above, our model performed very well. All of the predicted values are close to the actual values, and all of the actual values fall within the 95% confidence interval of the predicted values.

Furthermore, since we forecasted the model using the box-cox transformed data, we can then apply the inverse of the box-cox transformation in order to get the true monthly unemployment prediction values for interpretation purposes. Below is the table of the results of our forecast with the 95% confidence interval upper and lower bounds with the predicted or forecasted values alongside the actual values.

95% Upper Bound <dbl>	95% Lower Bound <dbl>	Prediction <dbl>	Actual <dbl>
565.5886	704.9298	631.8101	647.9
480.5976	661.1721	564.4174	568.8
426.8971	646.4239	526.4472	545.7
502.5339	821.1457	644.3200	632.6
460.3015	815.1821	615.0657	643.8
422.7902	806.5052	586.9890	593.1
399.4081	814.0609	573.8373	579.7
375.0269	811.8032	555.8958	546.0
380.8767	864.8285	578.7712	562.9
366.2453	870.5250	569.9605	572.5

We can also refit the model using all of the data and then forecast 10 months into the future to make predictions for data that we do not currently have. Note in this case it was checked that the model assumptions still hold, and all results and diagnostics were almost the same as when the model was fit without the last 10 observations.



From the forecast, we see that after 1978 we will see an upward trend in monthly unemployment for 10 months. The results are summarized in the table below, where the forecasted values for the next 10 months are shown

alongside their respective 95% confidence interval upper and lower bounds. Similarly, the Forecast results table shows the model forecast using the box-cox transformed data, and the Transformed Forecast table is the inverse of the box-cox transformation applied to the predicted values to get the true prediction for interpretation purposes.

Forecast Results

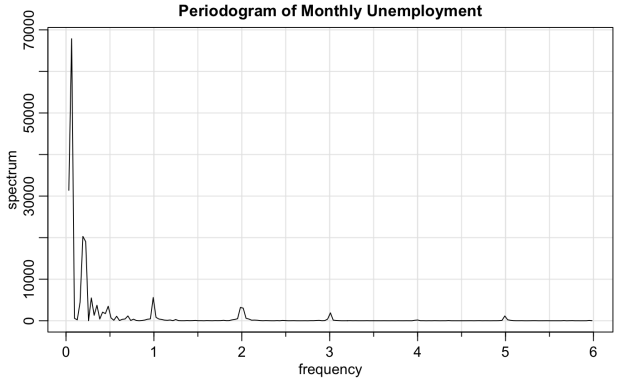
Prediction <dbl>	95% Lower Bound <dbl>	95% Upper Bound <dbl>
9.185029	9.013969	9.356090
9.215158	8.971171	9.459145
9.137771	8.822704	9.452838
8.945246	8.564893	9.325599
8.856316	8.415761	9.296871
9.226673	8.731267	9.722078
9.192583	8.647739	9.737427
9.095084	8.506179	9.683990
9.060672	8.432936	9.688408
8.989621	8.328047	9.651195

Transformed Forecast Results

95% Upper Bound <dbl>	95% Lower Bound <dbl>	Prediction <dbl>
605.7362	752.4113	675.4989
587.3066	799.7954	686.1824
538.3886	803.5088	659.0427
465.2331	758.1334	595.6655
425.9966	752.0115	568.2880
502.1115	939.7614	690.3055
477.2848	952.1678	678.1634
439.7694	931.3119	644.4895
420.6236	937.3478	632.9685
395.5322	923.3025	609.7657

4.4 Spectral Analysis

We now perform spectral analysis to identify the first three predominant periods and then find the corresponding confidence intervals.



The table below shows the first three predominant peaks of the spectral analysis alongside their respective frequencies, periods and the upper and lower bound of the 95% confidence interval.

Peak Number <dbl>	Frequency <dbl>	Period <dbl>	Spectrum <dbl>	95% CI Lower Bound <dbl>	95% CI Upper Bound <dbl>
1	0.064	15.625000	67861.59	29471.915	644089.4
2	0.032	31.250000	31392.40	13633.548	297952.3
3	0.192	5.208333	20295.94	8814.413	192633.2

We cannot establish the significance of the first peak since the periodogram ordinate is 67861.59, which lies in the confidence intervals of the second and third peak. Next we cannot establish the significance of the second peak since the periodogram ordinate is 31392.40, which lies in the confidence interval of the first peak. Furthermore, we cannot establish the significance of the third peak since the periodogram ordinate is 20295.94, which lies in the confidence interval of the second peak. Given also how wide the range all of these confidence intervals are it makes it difficult to establish a significance of any of the peaks.

5 Discussion

In this report we transformed our data to resemble that of a normal distribution and stabilized the variance using a box-cox transformation. Then we then detrended the data by taking the first difference and deseasoned the data by taking the seasonal difference. We selected the best model by performing various model diagnostics and first fitted the model without the last 10 months of observations for testing purposes. We found that when we forcasted 10 months into the future, the actual values all lied in the region we predicted with 95% confidence contains the true forecasts. We also fitted the model using all observations to make predictions on data we don't have then forecasted 10 months into the future. We found that after 1978, there will be an upward trend in U.S unemployment for 10 months. A limitation is how the parameter values were chosen for our model. The SARIMA model parameters is chosen by visually inspecting the ACF and PACF. For this reason we do not try fitting every model to find the best fit but only select a handful and then perform diagnostics to choose the best one from there. Another limitation to our model is that it does not incorporate other extraneous factors that economists agree are related to unemployment such as Gross Domestic Product(GDP) and inflation [5]. A possible area of research is developing models that can incorporate information on labor force flows, which describes how individuals move in and out of employment and unemployment or are not part of the labour force, that is governed by economic theory to form unemployment forecasts [6].

References

[1] "Indicator Description: Unemployment Rate." ILOSTAT. Accessed April 13, 2022. https://ilostat.ilo.org/resources/concepts-and-definitions/description-unemployment-rate.

[2] "Unemployment Rate." The Conference Board of Canada. Accessed April 13, 2022. https://www.conferenceboard.ca/hcp/Details/Economy/unemployement-rate.aspx.

[3] "Causes of Unemployment in the United States." Wikipedia. Wikimedia Foundation, February 19, 2016. https://en.wikipedia.org/wiki/Causes_of_Unemployment_in_the_United_States.

[4] Simpson, Stephen D. "The Cost of Unemployment to the Economy." Investopedia. Investopedia, April 11, 2022. https://www.investopedia.com/financial-edge/0811/the-cost-of-unemployment-to-the-economy.aspx.

[5] "Introduction to U.S. Economy: Unemployment." Accessed April 13, 2022. https://sgp.fas.org/crs/misc/IF10443.pdf.

[6] "The Ins and Outs of Forecasting Unemployment: Using Labor Force Flows to Forecast the Labor Market." Accessed April 13, 2022. https://www.brookings.edu/wp-content/uploads/2012/09/2012b_Barnichon.pdf.

[7] "1973–1975 Recession." Wikipedia. Wikimedia Foundation, January 11, 2022. https://en.wikipedia.org/wiki/1973%E2%80%931975_recession.