Time Series Analysis of U.S. Unemployment Rate

Abstract

The unemployment rate is highly associated with people' lives and social stability. Analyzing the unemployment rate can effectively help related departments to enforce positive policies. In this report, we built a seasonal ARIMA model and periodogram analysis based on the U.S. unemployment rate data from Jan, 1948 to Nov, 2016 to predict the unemployment rate of next ten months. Our model prediction shows that the U.S. unemployment rate has a decreasing trend in the following ten months which may imply that the government is making effective policies, while the periodogram analysis wasn't useful. We would like to suggest the U.S. related department to continue the ongoing policies to further eliminate the unemployment rate. Still, there are some limitations in our model, some improvements might be helpful to make a more accurate prediction.

Introduction

This reports analyzes the monthly U.S. unemployment rate from Janauary in 1948 to November 2016. The dataset is proved by applied statistical analysis (astsa) package [1] which contains 827 observations. Based on these time period of data, we are going to use time series analysis to predict the future 10 months' unemployment rate. The data has strong seasonality and some severe peaks, possibly was affected by seasonal events such as extreme weather or inconsistent policies. The severe worldwide economic recession leads the unemployment rate to the peak in early 1980s [2], and the elimination of unemployment rate has become a major economic concern in the U.S. Statistically, the lower crime rate around 1990s was due to the decline in unemployment rate to some extent [3]. Also, as a fourth's year student, the unemployment rate is important area of interest. The method implemented is seasonal autoregressive integrated moving average model with the consideration of seasonality to make the prediction.

Statistical Method

To start with, we generated the monthly time series plot of the unemployment rate in U.S. from Jan, 1948 to Nov, 2016 to observe the general trend. As Figure 1 shows that the monthly unemployment rate fluctuates with seasonal effects. It started at a low rate and reached to the peak around 1983 because of the economic recession, then the rate dropped while experienced a dramatic increase in 2010, and steadily decreased to around 5% at the end of 2016.

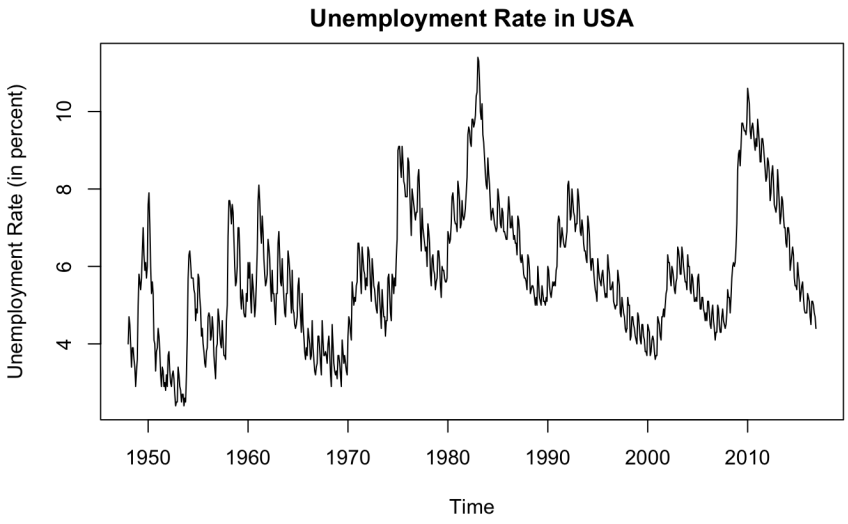

Figure 1: The Monthly U.S. Unemployment from 1946 to 2016

The data is clearly not stationary due to its irregular fluctuation, then we will apply log transformation. By differecing the logged data $\nabla \log(x_t)$, the transformation stabilizes the variance. While it still exists the seasonal trend according to the first plot of Figure 2. Since the data is recorded monthly, I tried to propose a twelfth-order difference on the differenced logged data. $\nabla_{12}\nabla \log(x_t)$ looks stationary as it has constant mean around 0, constant variance and the seasonal trend is removed. Now it's ready to fit the mode with the differencing order of non-seasonal component d=1, and the differencing order of seasonal conponent D=1, and the time span of cycle S=12.

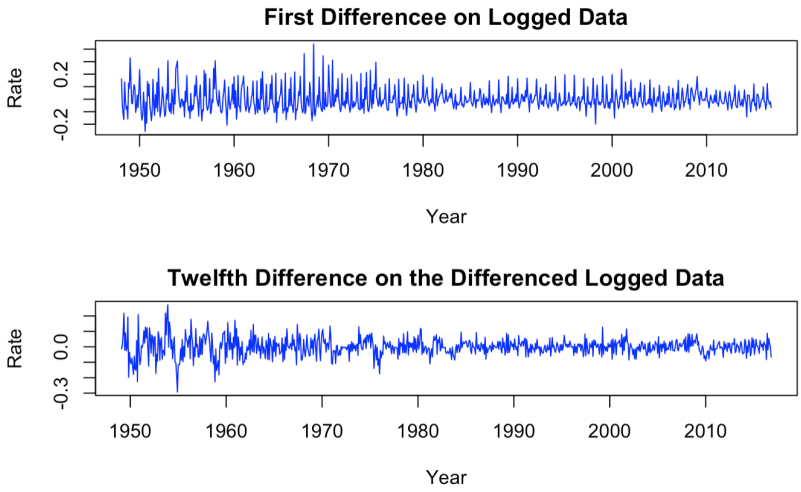

Figure 2: $\nabla \log(x_t)$ and $\nabla_{12}\nabla \log(x_t)$ Differencing Time Series

Since we will fit a SARIMA model, then we need to explicitly consider both the seasonal and non-seasonal components [4] by looking at the ACF and PACF plots.

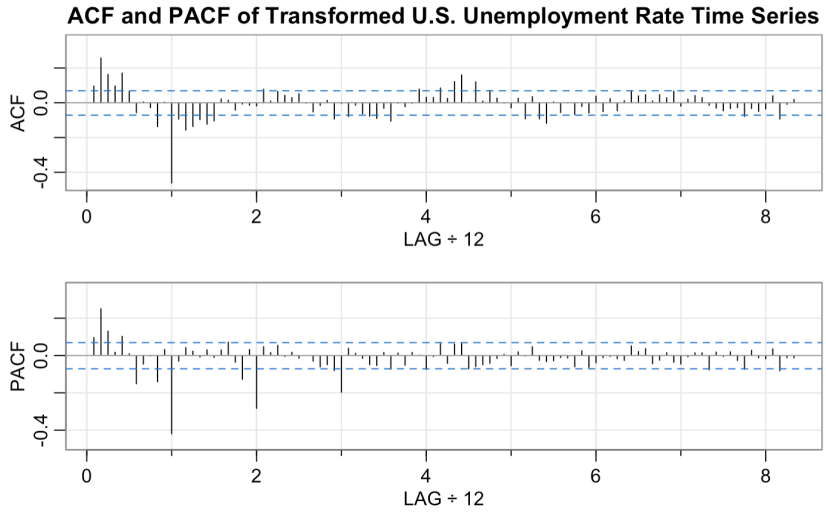

Figure 3: Sample ACF and PACF of $\nabla_{12}\nabla \log(x_t)$

It can be seemed from Figure 3 that for seasonal part, ACF is cutting off a lag 1s (s=12) whereas PACF is tailing off. This implies that P=0, Q=1. For the non-seasonal part, ACF cuts off at lag 5 when PACF is tailing off or PACF cuts off at lag 3 when ACF is tailing off. Then we could try p=3, q=0 or p=0, q=5. Then we are proposing two candidate mod-

els **ARIMA(3,1,0)×(0,1,1)$_{12}$** and **ARIMA(0,1,5)×(0,1,1)$_{12}$** for U.S. unemployment rate data, and do further model selection.

Result

Figure 4 and 5 shows the residual diagnotics for the two candidate models. The standardized residual plots for both models behave like white noise despite a peak around the year 1975. The ACF plots of both model are within the range of normal assumption. As for the residual normal QQ-plot, most of points for both models lie on the straight line except for a few tolerable outliers. The P-value of Ljung Box Statistic plots shows that some points of ARIMA(3,1,0)×(0,1,1)$_{12}$ are on the line of significance level, while all points for ARIMA(0,1,5)×(0,1,1)$_{12}$ are above the significance level. Then we fail to reject the null hypothesis of independence, while the second ARIMA(0,1,5)×(0,1,1)$_{12}$ model is comparatively better. To fully support our claim, we will check the AIC, AICc, BIC and p-values of all parameter estimates for our two models.

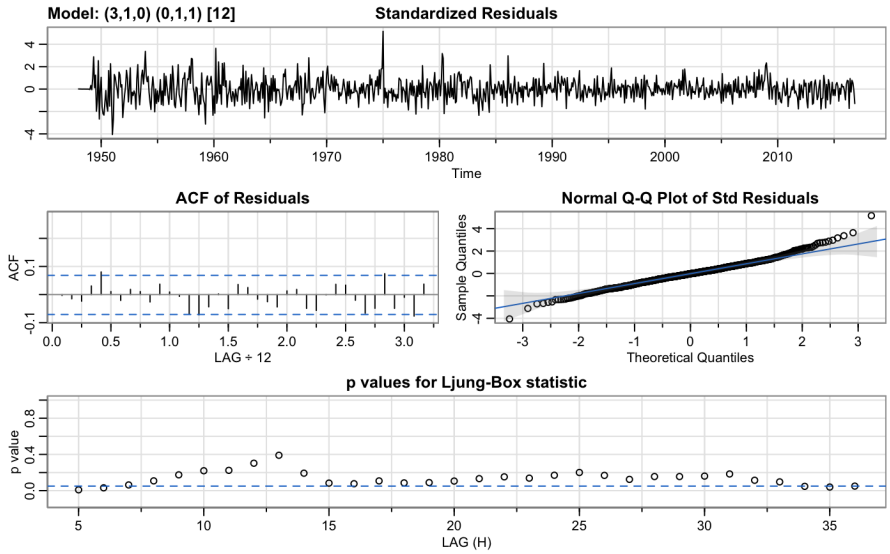

Figure 4: Residual Analysis for ARIMA(3,1,0)×(0,1,1)$_{12}$

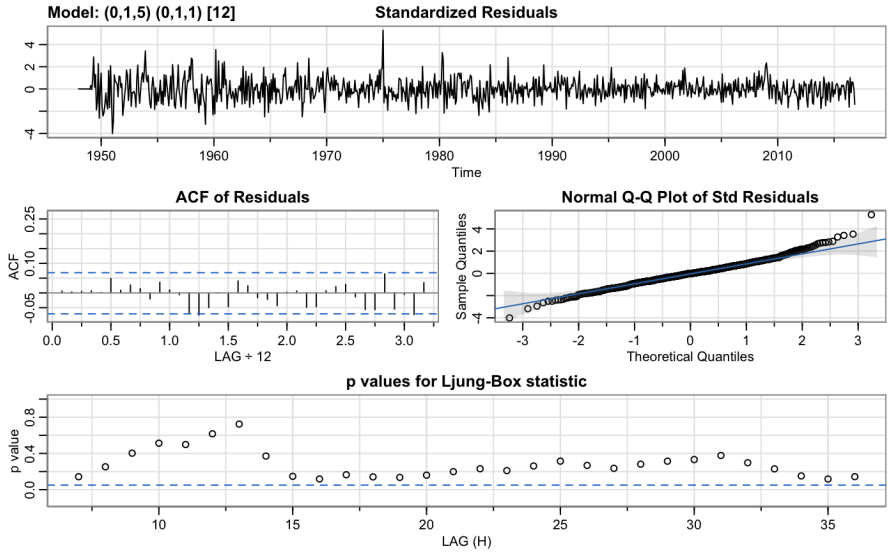

Figure 5: Residual Analysis for ARIMA(0,1,5)×(0,1,1)$_{12}$

Table 3 shows that the AIC and AICc values of ARIMA(0,1,5)×(0,1,1)$_{12}$ model are smaller than ARIMA(3,1,0)×(0,1,1)$_{12}$, despite having a slight bigger BIC value.

Table 3: AIC, AICc, BIC Values for Two Model

	AIC	AICc	BIC
ARIMA(3,1,0)×(0,1,1)$_{12}$	-0.022	-0.022	0.007
ARIMA(0,1,5)×(0,1,1)$_{12}$	-0.024	-0.024	0.016

Table 1 and 2 display the p-values for all parameter estimates of ARIMA(3,1,0)×(0,1,1)$_{12}$ and ARIMA(0,1,5)×(0,1,1)$_{12}$ models. All p-values for ARIMA(3,1,0)×(0,1,1)$_{12}$ are lower than 0.05 significance level except for the term ϕ_3. Similarly, all p-values are smaller than 0.05 significance level except for the term θ_4 in the ARIMA(0,1,5)×(0,1,1)$_{12}$ model. Then these two parameters are statistically insignificant.

Table 1: Parameter Estimations and P-value for ARIMA(3,1,0)×(0,1,1)$_{12}$

	Estimate	P-value
ϕ_1	0.1148	0.0011
ϕ_2	0.2023	0.0000
ϕ_3	0.0900	0.0103
Θ_1	-0.7674	0.0000

Table 1: Parameter Estimations and P-value for ARIMA(0,1,5)×(0,1,1)$_{12}$

	Estimate	P-value
θ_1	0.1029	0.0033
θ_2	0.2009	0.0000
θ_3	0.1068	0.0019
θ_4	0.0859	0.0128
θ_5	0.1331	0.0002
Θ_1	-0.7706	0.0000

Overall, these two models both show good performance, while combine the model residual diagnostics, p-values and AIC, AICc, BIC values we choose **ARIMA(0,1,5)×(0,1,1)**$_{12}$ for the future forecasting. From parameter estimation of Table 2, we can write our final fitted model:

$$\nabla_{12}\nabla \log(\hat{x}_t) = (1+0.1029B)(1+0.2009B)(1+0.1068B)(1+0.0859B)(1+0.1331B)(1-0.7706B)\hat{w}_t$$

In details, p=0 and P=0 indicates the numbers of autoregressive term, d=0 and D=0 is the orders of differencing, d=6 and D=1 are the numbers of of moving average terms, s=12 is the length of a seasonal cycle. Now we are able to forecast the future next 10 months' values of U.S. unemployment rate.

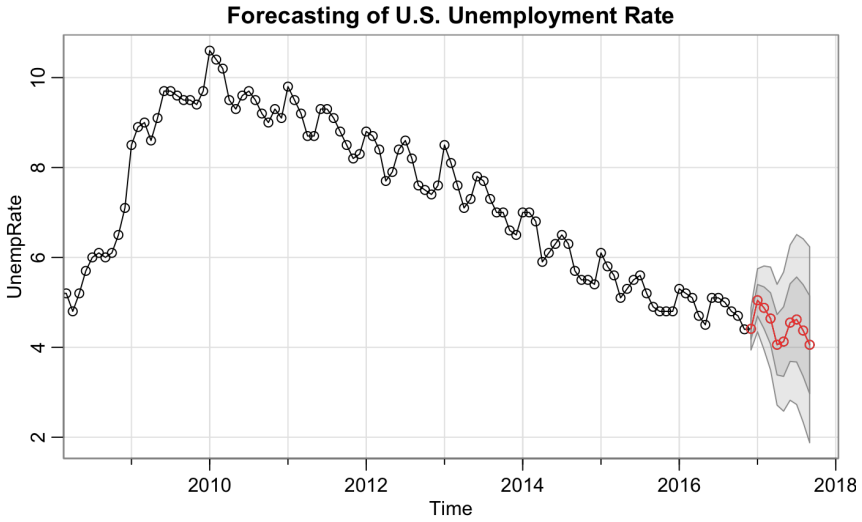

Figure 6: Ten-Months Forecast of ARIMA(0,1,5)×(0,1,1)$_{12}$ model on $logx_t$

Figure 6 forecasts the logged value of U.S. unemployment rate of next ten months in red dots, where the grey areas are the prediction interval. We see that there is a general decreasing trend with the seasonal patterns after Nov, 2016, and the details of the forecasts and 95% prediction intervals are in Table 4 on the natural scale after taking the exponential. If we continue our forecasting to 12 months, we can see a full cycle of seasonal trends.

Table 4: The Summary of First Three Predominant Periods

	Prediction	Lower Bound of 95% PI	Upper Bound of 95%
1	82.329	79.849	84.809
2	155.532	152.749	158.314
3	131.312	128.189	134.435
4	103.588	100.115	107.061
5	57.796	53.958	61.663
6	61.963	57.714	66.211
7	94.599	89.951	99.248
8	101.300	96.258	106.342
9	79.305	73.870	84.739
10	57.760	51.933	63.587

It's clear that our U.S. unemployment rate time series has periodic behavior. By doing spectral analysis, we can decompose the time series and identify the first three predominant periods.

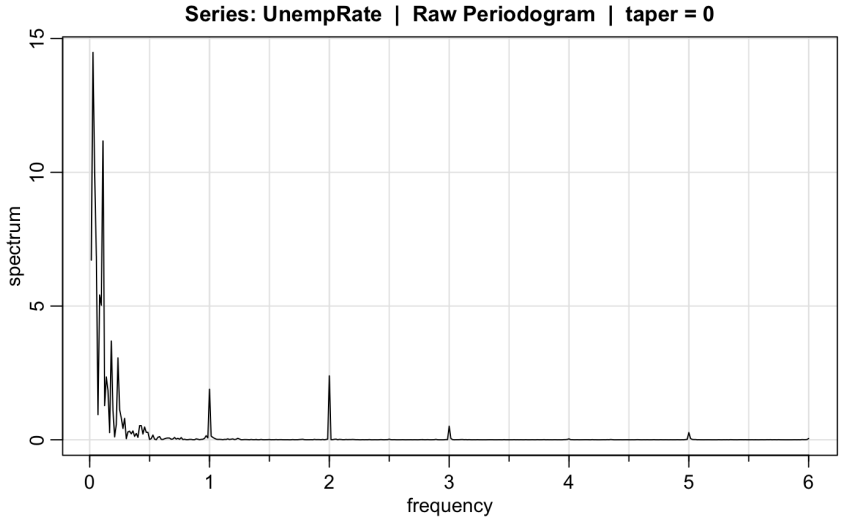

Figure 7: Periodogram of U.S. Unemployment Rate

Figure 7 shows that the first peak appears around 0.1, which is similar to the result of our seasonal effects analysis that S=12 where 1/12=0.08. More details of the first three predominant periods and their confidence interval can be found in Table 5.

Table 5: The Summary of First Three Predominant Periods

	Frequency	Period	Spectrum	0.025 Quantile	0.0975 Quantile
1	0.0278	36.0000	14.4829	3.9261	572.0440
2	0.1111	9.0000	11.1720	3.0286	441.2704
3	0.0417	24.0000	9.9625	2.7007	393.4977

Table 4 shows that all 95% confidence interval for three predominant periods are very large. In addition, all three spectrum values lie in the confidence intervals of other two spectrum values. Therefore, we are unable to identify the significance of this three dominant peaks.

Discussion

Based on the prediction of our model ARIMA(0,1,5)$\times$(0,1,1)$_{12}$, the U.S. unemployment rate is likely to go down for the next 10 months including the peak and the valley points. It means that the U.S. unemployment rate will be lower compared to the same month last year, which implies there will be more job opportunities and a better economy condition of the state. The current polices from the U.S. public sectors are effective, and they may continue to use. We also used a periodogram to identify the cyclical behavior of the series and the first three predominant periods. While the 95% confidence intervals are large to determine the significance. Although our model provides a reasonable forecast, there are still some limitations. Firstly, the 95% prediction interval is a bit wide which might challenge the reliability of the model. As we see from the standardized residual plots and the normal QQ-plots, there are some outliers that we didn't deal with before fitting the model. Also, I used log transformation to reduce the skewness of the data, while the data did not behavior perfectly after the log transformation [5]. We might consider other transformation to fit a better model. Thirdly, the AIC and BIC

values of the two candidate models contradicts with each other. Lastly, the data is from 1950 to 2016, some major historical events might heavily influence our data. If we could analyze on a more recent dataset, the prediction might be more accurate.

Reference

[1] Stoffer, D. (2021, September 5). UnempRate: U.S. unemployment rate in ASTSA: Applied statistical time series analysis. UnempRate: U.S. Unemployment Rate in astsa: Applied Statistical Time Series Analysis. Retrieved December 17, 2021,
from https://rdrr.io/cran/astsa/man/UnempRate.html

[2] Bednarzi, R. W., amp; URQUHART, M. A. (n.d.). The employment situation in 1981: New recession takes its toll. Retrieved December 17, 2021,
from https://www.bls.gov/opub/mlr/1982/03/art1full.pdf

[3] Raphael, S., amp; Winter-Ebmer, R. (2001). Identifying the effect of unemployment on crime. The Journal of Law and Economics, 44(1), 259–283. https://doi.org/10.1086/320275

[4] Graves, A. (2020, July 21). Time Series Forecasting with a Sarima model. Medium. Retrieved December 17, 2021, from https://towardsdatascience.com/time-series-forecasting-with-a-sarima-model-db051b7ae459

[5] Feng, C., Wang, H., Lu, N., Chen, T., He, H., Lu, Y., amp; Tu, X. M. (2014, April). Log-transformation and its implications for data analysis. Shanghai archives of psychiatry. Retrieved December 17, 2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4120293/