CSC 425

Final Project

Civilian Unemployment Rate

Tzu-Hao Peng

Non-Technical Report

The dataset is from quandl and it is about the unemployment rate from January 1948 to January 2018. The data is collected in the US for people 16 years of age and older. There are two variables including date and values with 841 observations. The purpose of this project is to predict the unemployment rate for the next 12 months.

From the exploratory of the data, the time series plot shows a seasonal pattern. The unemployment rate are not always high or always low. It fluctuated over years. More specifically, the peaks went over almost every 10 years. However, after 2000s, the unemployment rate fluctuated more often than before, there is a peak in every 5 years. In my opinion, I think the problem is related to the financial crisis between 2007 to 2009.

From the model fitting, I examined the correlation with some plots and test the data with two models. One with seasonal component and the other without seasonal component. The result shows that the model with the seasonal component would be the better model because of the smaller error values. The seasonal component is a sum of the periodical short-term movements having a length of less than one year. It represents intra-year fluctuations that repeat every year with respect to timing, direction and magnitude.

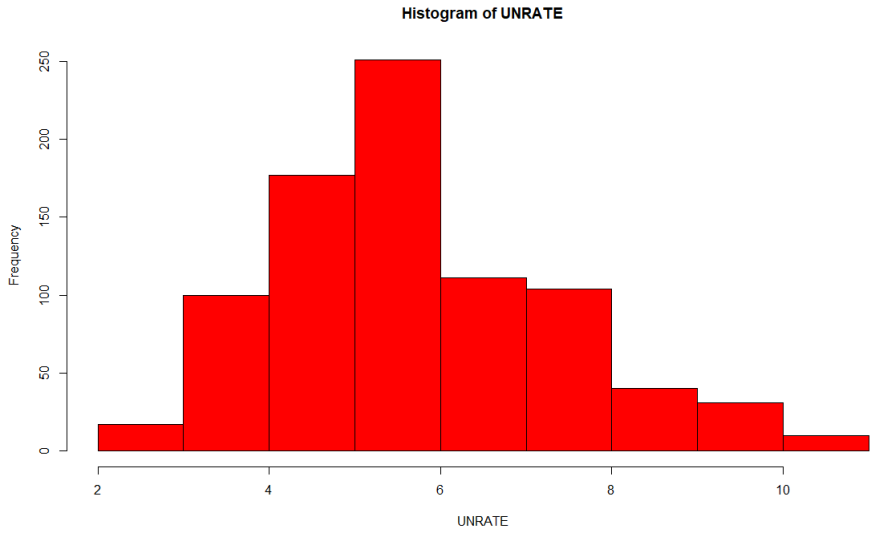
From the Residual analysis and model diagnostics, we can see from the plots that the model is slightly normally distributed with stationary but not serial correlated. Serial correlation is the relationship between a given variable and itself over various time intervals. Serial correlations are often found in repeating patterns, when the level of a variable effects its future level. After several analysis and diagnostics, I would say this is a valid model.

Last but not least, from the forecast analysis, I made the forecast of the future 12 months from the final model. The forecast shows that the unemployment rate would decline first in the next few months and rise after reaching the lowest point, compare to the time series trend in the past 60 years, I would say the prediction is appropriate.

Technical Report

a. Exploratory analysis of the data.

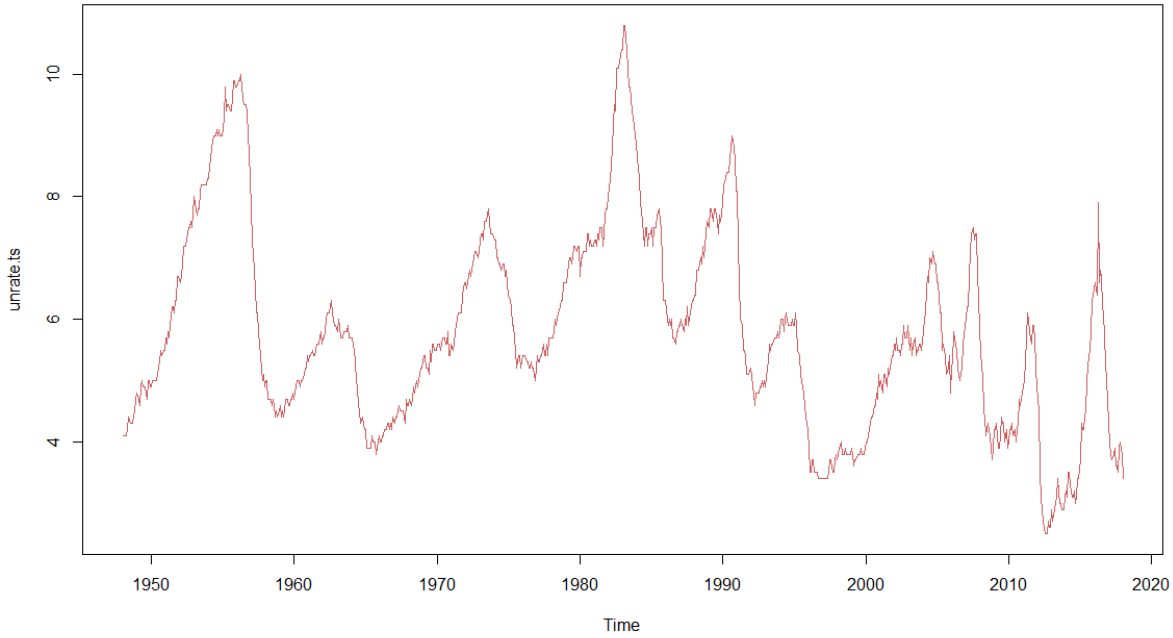
The data is from quandl and it is about the unemployment rate from January 1948 to January 2018. This data are restricted to people 16 years of age and older, who currently reside in 1 of the 50 states or the District of Columbia, who do not reside in institutions (e.g., penal and mental facilities, homes for the aged), and who are not on active duty in the Armed Forces. The data includes 2 variables and 841 observations with no missing values.



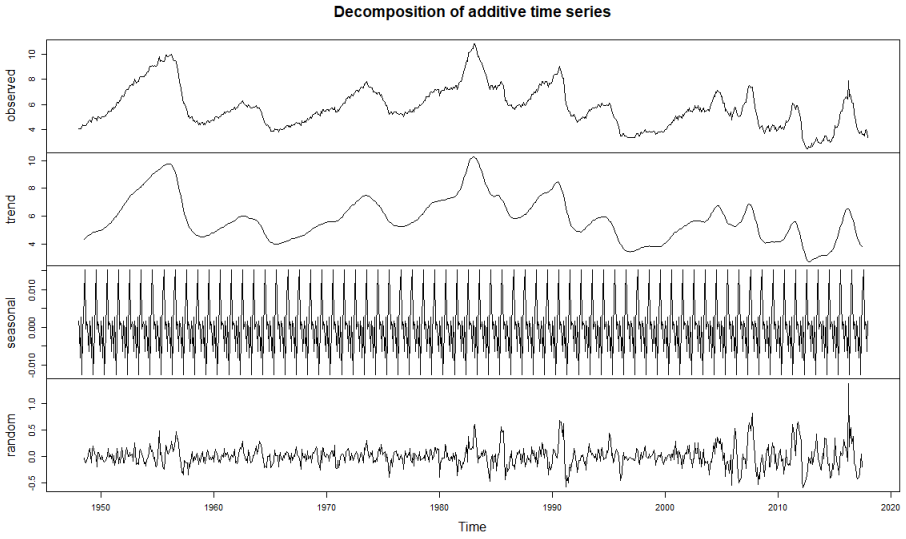
First, I am going to see the histogram of the unemployment rate. It is a little bit right skewed and most of the unemployment rate are between 4 to 6.

```
> basicStats(unrate.ts)
                unrate.ts
nobs          841.000000
NAs             0.000000
Minimum         2.500000
Maximum        10.800000
1. Quartile     4.600000
3. Quartile     6.900000
Mean            5.787753
Median          5.600000
Sum          4867.500000
SE Mean         0.056374
LCL Mean        5.677101
UCL Mean        5.898404
Variance        2.672766
Stdev           1.634860
Skewness        0.616868
Kurtosis        0.088183
```

This is the basic information about this data. The mean is slightly larger than median and the skewness = 0.617. It is a right skewed.



Next, I'm going to create the time series plot to see if there are any patterns. In my opinion, the plot shows a seasonal pattern. The unemployment rate are not always high or always low. It fluctuated over years.

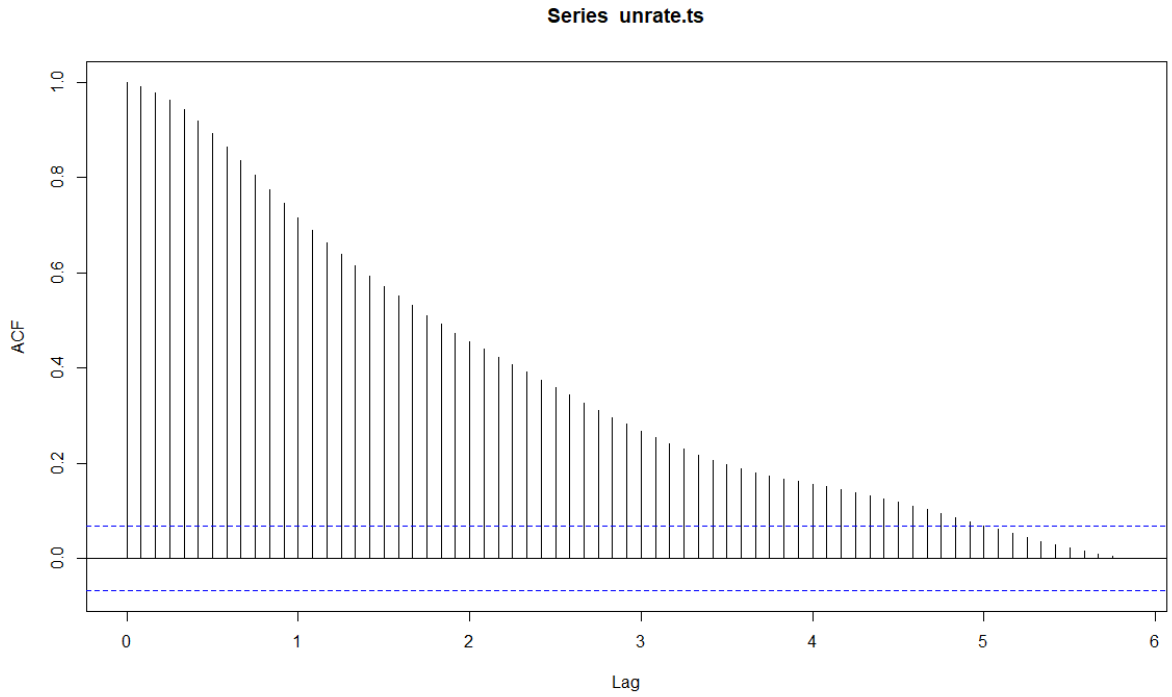


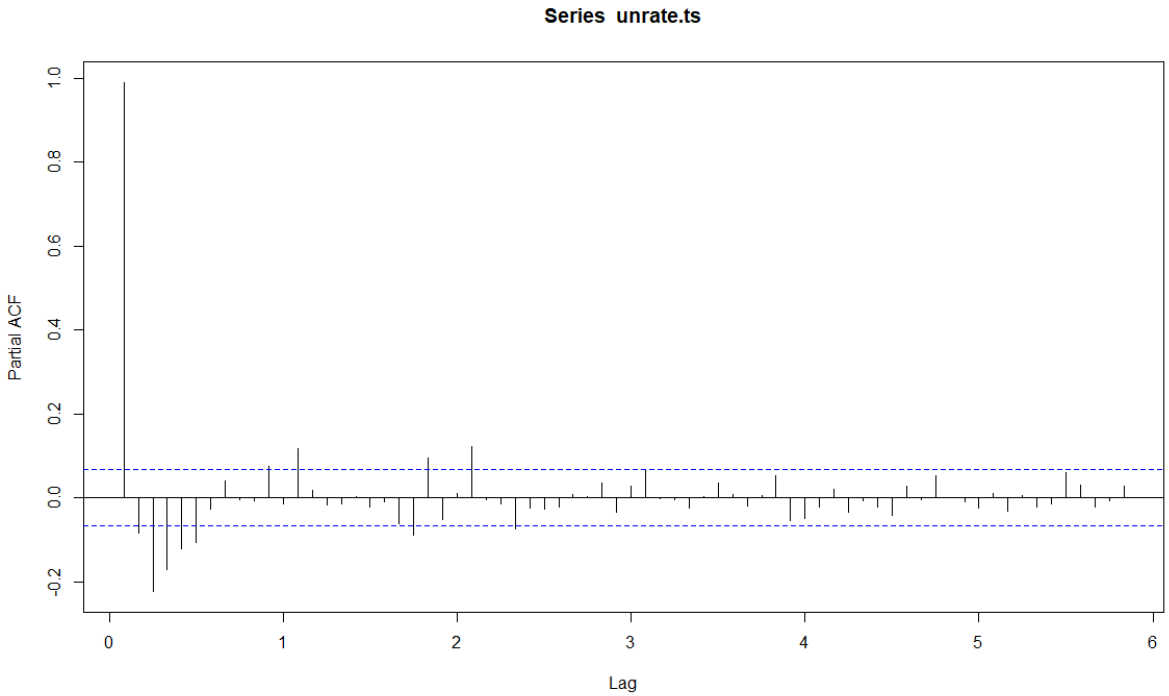
Then, I did the decomposition of the original time series. The first row is the original series. The second row shows the trend. The third row shows the seasonal pattern and the last row shows the random noise.

```
> adf.test(unrate.ts)

        Augmented Dickey-Fuller Test

data:  unrate.ts
Dickey-Fuller = -4.1, Lag order = 9, p-value = 0.01
alternative hypothesis: stationary
```

Last, I did the Augmented Dickey-Fuller Test to see if my dataset is stationary. And the p-value = 0.01 < 0.05 shows that the dataset is stationary. I don't have to do any transformation of the data.

b. Model fitting.





First, I did ACF and PACF plots to see the auto-correlation. From the ACF plot, there is serial correlated with a gently linear decay speed. From the PACF plot, there is a high peak at the beginning and the series seems to be cut off at lag 4.

```
> auto.arima(unrate.ts, ic = "aic")
Series: unrate.ts
ARIMA(2,1,2)(0,0,2)[12]

Coefficients:
         ar1      ar2      ma1     ma2     sma1     sma2
      1.3135  -0.4299  -1.3120  0.5759  -0.2575  -0.2293
s.e.  0.1959   0.1817   0.1783  0.1357   0.0361   0.0372

sigma^2 estimated as 0.03514:  log likelihood=215.78
AIC=-417.55   AICc=-417.42   BIC=-384.42
```

In order to select the model more accurately, I did the auto arima test to select the best model with the lowest AIC. As a result, ARIMA(2, 1, 2)(0, 0, 2)[12] with AIC = -417.55 would be the best model.

```
> aa = arima(unrate.ts, order = c(2, 1, 2), seasonal = list(order = c(0, 0, 2), perio
d = 12))
> confint(aa)
             2.5 %      97.5 %
ar1    0.9296551  1.69739139
ar2   -0.7860422 -0.07379416
ma1   -1.6614045 -0.96252999
ma2    0.3099339  0.84187363
sma1  -0.3281949 -0.18672427
sma2  -0.3022829 -0.15637517
```

Next, I computed the confidence intervals for this fitted model to see if 0 is included in the 95% CI. As a result, none of the outputs are included 0. The model is significant. ARIMA(2, 1, 2)(0, 0, 2)[12] with AIC = -417.55 would be the best model.

```
> backtest(aa, unrate.ts, orig = 600, h = 1)
[1] "RMSE of out-of-sample forecasts"
[1] 0.2506962
[1] "Mean absolute error of out-of-sample forecasts"
[1] 0.1825133
[1] "Mean Absolute Percentage error"
[1] 0.03937801
[1] "Symmetric Mean Absolute Percentage error"
[1] 0.03922849
```

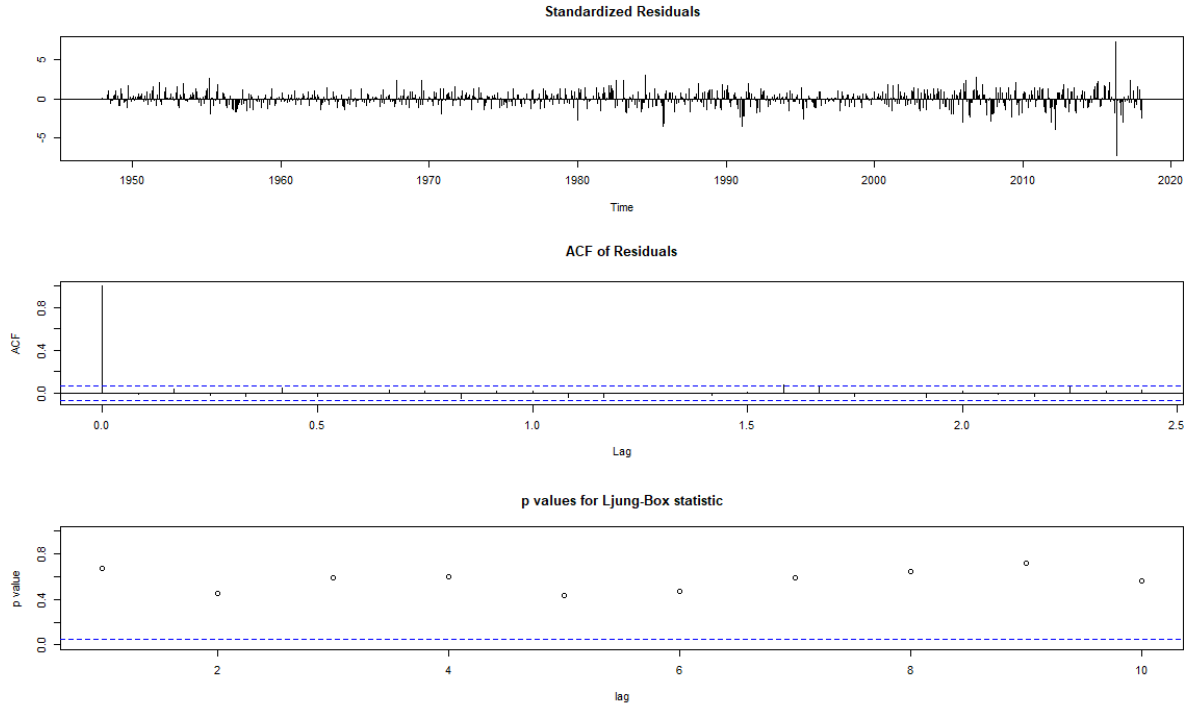
Last, I did backtesting for the final model, and the result is shown above: RMSE of out-of-sample forecasts = 0.251, MAE of out-of-sample forecasts = 0.183, MAPE = 0.039, Symmetric MAPE = 0.039.

```
> aa2 = arima(unrate.ts, order = c(2, 1, 2))
> backtest(aa2, unrate.ts, orig = 600, h = 1)
[1] "RMSE of out-of-sample forecasts"
[1] 0.2627739
[1] "Mean absolute error of out-of-sample forecasts"
[1] 0.191221
[1] "Mean Absolute Percentage error"
[1] 0.04082669
[1] "Symmetric Mean Absolute Percentage error"
[1] 0.04068736
```

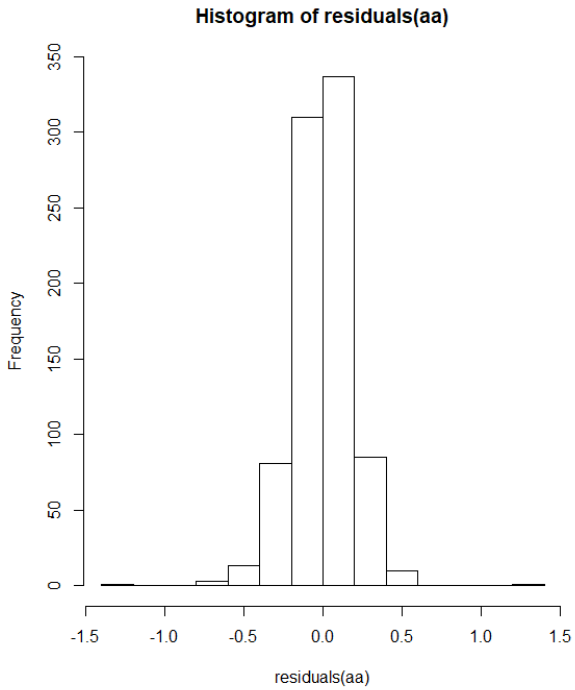
However, I did one more backtesting in order to remove the seasonal component. The result is shown above, RMSE = 0.263, MAE out-of-sample forecasts = 0.191, MAPE = 0.041, Symmetric MAPE = 0.041.

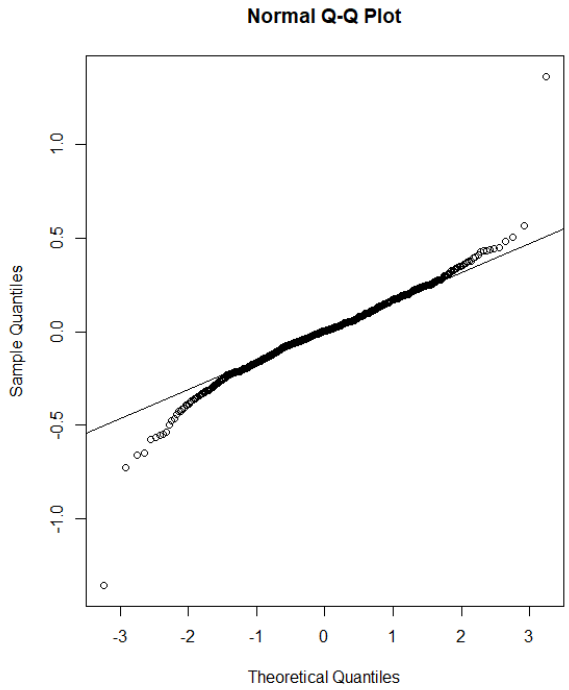
In conclusion, I would say the first model without removing seasonal component would be the better model because of the slightly lower error values.

C. Residual analysis and model diagnostics.



First, I did the diagnostics plots to analyze the residuals and diagnostics the models. From the standardized Residuals plot above, I would say the model is stationary. From the ACF of Residuals plot above, I would say the model is not serial correlated because the plot is cut off at lag 0. From the p values for Ljung-Box statistic plot above, all of the p-values are above 0.05 in any lag order.





Next, I did the Histogram of residual plot and a normal Q-Q plot to see the distribution of the model. In my opinion, I would say the model is normally distributed.

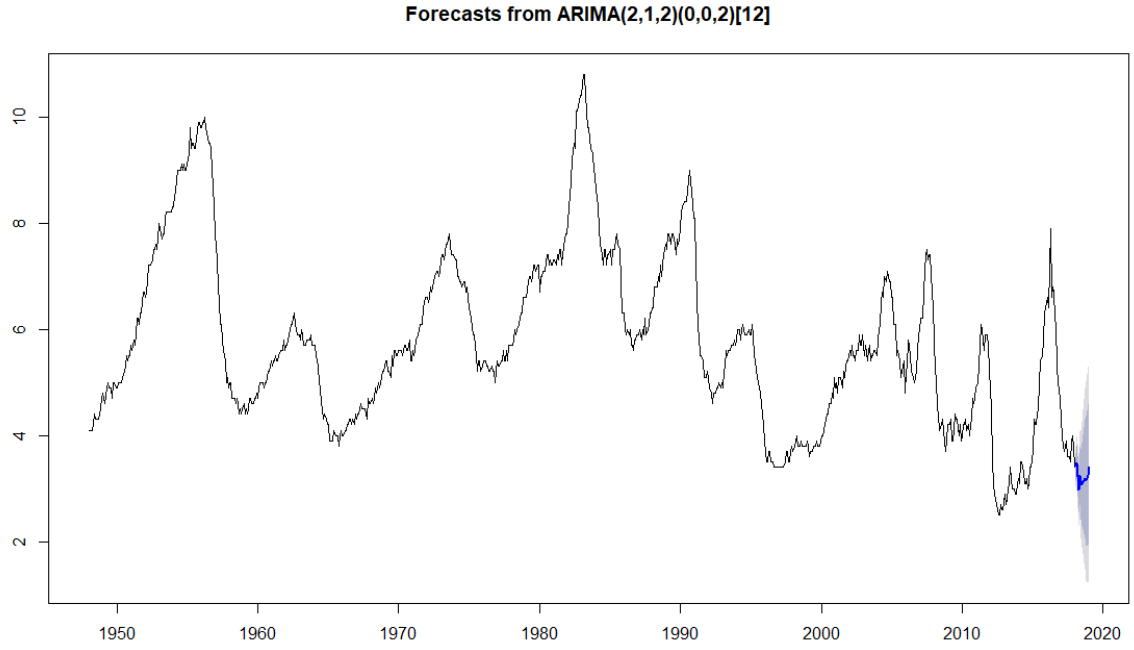
```
> jarque.bera.test(residuals(aa))

        Jarque Bera Test

data:  residuals(aa)
X-squared = 1620.2, df = 2, p-value < 2.2e-16
```

However, the Jarque Bera test reject the null hypothesis, the model is not normally distributed. In conclusion, although the model is not normally distributed, the model is not serial correlated and stationary. I would say this is a valid model.

d. Forecast analysis.

```
> final
         Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
Feb 2018       3.430384 3.190992 3.669777 3.064265 3.796503
Mar 2018       3.485582 3.146767 3.824398 2.967409 4.003756
Apr 2018       2.984600 2.548100 3.421099 2.317031 3.652168
May 2018       3.247644 2.705471 3.789817 2.418461 4.076826
Jun 2018       3.084256 2.429389 3.739122 2.082724 4.085787
Jul 2018       3.079149 2.307919 3.850379 1.899654 4.258644
Aug 2018       3.130096 2.241999 4.018194 1.771868 4.488324
Sep 2018       3.145326 2.142171 4.148482 1.611132 4.679521
Oct 2018       3.193835 2.078893 4.308778 1.488677 4.898993
Nov 2018       3.146686 1.924028 4.369344 1.276792 5.016580
Dec 2018       3.258146 1.932187 4.584105 1.230267 5.286025
Jan 2019       3.408618 1.983818 4.833418 1.229575 5.587661
```

Finally, I made the forecast of the future 12 months from the final model. The result is above.



From the plot, we could see it clearly that the unemployment rate would decline in the next few months and rise after reaching the lowest point. From the past trend, it showed there is a seasonal pattern. I would say this prediction is appropriate.

e. Analysis of the results and discussion.

In conclusion, the ACF and PACF plots showed that if I needed a seasonal model. After finding the best model with the lowest AIC and do the confidence interval test to see if the

estimated coefficients different from 0. The final model = ARIMA(2, 1, 2)(0, 0, 2)[12]. Then I did backtesting test with the seasonal component and without seasonal component, the result shows that the model ARIMA(2, 1, 2)(0, 0, 2)[12] is still better. From the residuals analysis and model diagnostics plot, we can see that residuals of the model are not correlated, the model is valid. At last, I made the forecast of the model for the next 12 months. The forecast shows that the unemployment rate would decline first in the next few months and rise after reaching the lowest point, compare to the time series trend in the past 60 years, I would say the prediction is appropriate.

Reference:

- Quandl. (n.d.). Retrieved March 06, 2018, from https://www.quandl.com/data/FRED/UNRATE-Civilian-Unemployment-Rate
- Irwin, N. (2018, February 28). How Low Can Unemployment Really Go? Economists Have No Idea. Retrieved March 06, 2018, from https://www.nytimes.com/2018/02/28/upshot/how-low-can-unemployment-really-go-economists-have-no-idea.html
- United States Unemployment Rate - Forecast. (n.d.). Retrieved March 06, 2018, from https://tradingeconomics.com/united-states/unemployment-rate/forecast

Appendix:

R code:

```
library(tseries)

library(fBasics)

library(forecast)

library(backtest)


unrate = read.csv("D:/depaul/CSC425/FREDUNRATE.csv")

unrate.ts = ts(unrate[, 2], start = c(1948, 1, 1), end = c(2018, 1, 1), frequency = 12)


basicStats(unrate.ts)


hist(unrate$Value, col = 'Red', main = "Histogram of UNRATE", xlab = "UNRATE")


plot(unrate.ts, col = "indianred", lwd = 1)


dec = decompose(unrate.ts)

plot(dec)


adf.test(unrate.ts)


acf(unrate.ts, lag.max = 70)
```

```
pacf(unrate.ts, lag.max = 70)

auto.arima(unrate.ts, ic = "aic")

aa = arima(unrate.ts, order = c(2, 1, 2), seasonal = list(order = c(0, 0, 2), period = 12))
confint(aa)

source("D:/depaul/CSC425/backtest.R")
backtest(aa, unrate.ts, orig = 600, h = 1)
aa2 = arima(unrate.ts, order = c(2, 1, 2))
backtest(aa2, unrate.ts, orig = 600, h = 1)

tsdiag(aa)

hist(residuals(aa))
qqnorm(residuals(aa))
qqline(residuals(aa))

jarque.bera.test(residuals(aa))

final = forecast(aa, 12)
final
plot(final)
```