Time series study of Civilian Unemployment Rate from 1948 to 2016

Team member: Li Huang, Yiying Wu

1. Abstract

The unemployment rate is always a hot topic across the world, and especially concerned with us who are living and studying in the USA and probably trying to find a job after graduation. Therefore, we choose this dataset that is the summary of unemployment rate, which represents the number of unemployed as a percentage of the labor force [1].

The goal of our research is to analyze the properties of the time series object unemployment rate and to identify an adequate model to explain the data. By analyzing several possible model and select the most appropriate model to explain the time process that produced the data. After we selected the model, we will check the adequacy of the model using residual analysis and model diagnostics techniques, for example using back-testing procedures to validate the selected model on a testing.

The main findings of our research are the seasonal pattern of the unemployment rate that is appearing yearly in the past 60 some years. It seems the unemployment reached a victory low until January of 2016, however, by our forecast, the unemployment rate in 2016 is going to increase.

2. Introduction

"The U.S. unemployment rate just fell below 5% for the first time since 2008… The economy is better than it was in the Great Recession, but not even President Obama is ready to declare it's booming…"[2]. To understand how the unemployment rate is going to change before our graduation, we need to fit a time series model for the data and make a forecast to see if the economy is booming or it will still be hard to find a job.

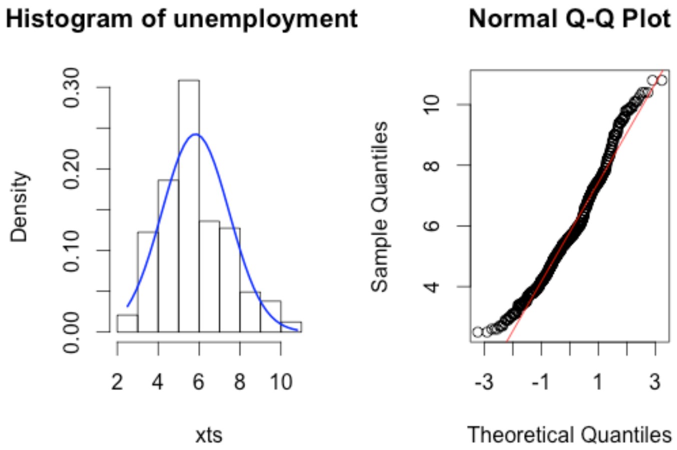
In this project, we applied time series analysis; check stationarity and seasonal pattern before fitting seasonal ARIMA models to the unemployment rate data and make forecasts.

3. Methodology

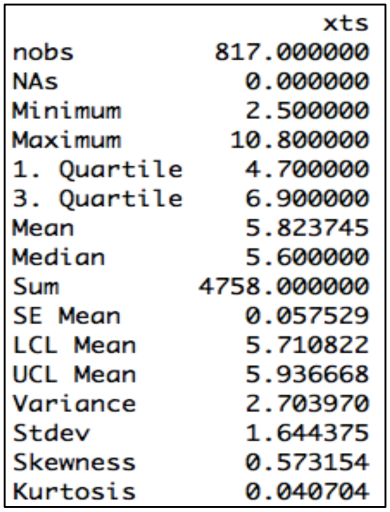
3.1. Descriptive analyze of Unemployment rate

The original dataset of the unemployment rate include two variables, date and unemployment rate. The unemployment rate is monthly distributing continuous data, with minimum value 2.5, median 5.6, and maximum value 10.8. From figure 1, we could observe the mostly frequently unemployment rate is around 5 to 6 percent. The data is not perfectly normally distributed but still acceptable.

Figure 1: Histogram and QQ plot of Unemployment rate



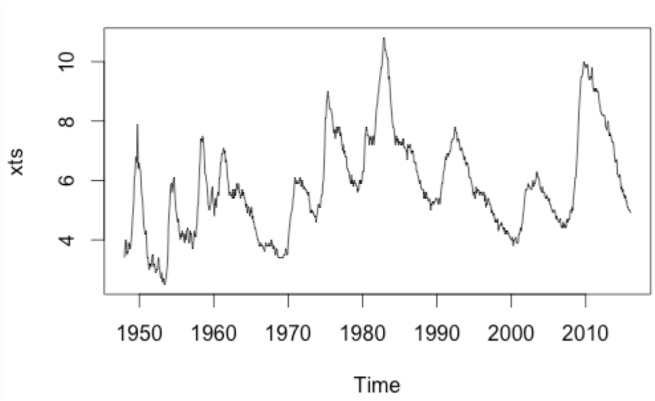
To find out how the time series object of the unemployment rate(xts)'s distribution looks like, we make a basic summary of xts and the finding is below:
Figure 2: basic summary of the time series object unemployment rate (xts)



```
                      xts
nobs           817.000000
NAs              0.000000
Minimum          2.500000
Maximum         10.800000
1. Quartile      4.700000
3. Quartile      6.900000
Mean             5.823745
Median           5.600000
Sum           4758.000000
SE Mean          0.057529
LCL Mean         5.710822
UCL Mean         5.936668
Variance         2.703970
Stdev            1.644375
Skewness         0.573154
Kurtosis         0.040704
```

Figure 2. The skewness is 0.57 showing the distribution is a little bit skewed and the kurtosis is 0.04, showing the distribution has a little bit fat tile. But the value of skewness and kurtosis is very small, we could assume the result of our analyze would be appropriate. To make sure we select the best suitable model for the data, we will also compare a model build on logarithm of xts in later part 3.3.

To observe the pattern of unemployment rate over time, we created a time plot of the series object. Figure 3 shows how the unemployment rate over time from 1948 to 2016.
Figure 3: Time plot of unemployment rate



From figure 3, we could observe the plot shows several characteristics of the series. First, as expected, the unemployment rate exhibits strong cyclical pattern signifying expansions and contractions of the US economy. For example, the unemployment increased dramatically during 2007 and 2009 because of the US recession. The pattern does not have a fixed period because economic expansions and contractions have no fixed durations. Second, the unemployment rate shows a slightly upward trend. There are several possible explanations for the trend, including the increase in labor forces and participations, and advances in technology. Third, the unemployment rate rose quickly and declined slowly. This asymmetric behavior indicates that unemployment rate does not follow a linear time series model.

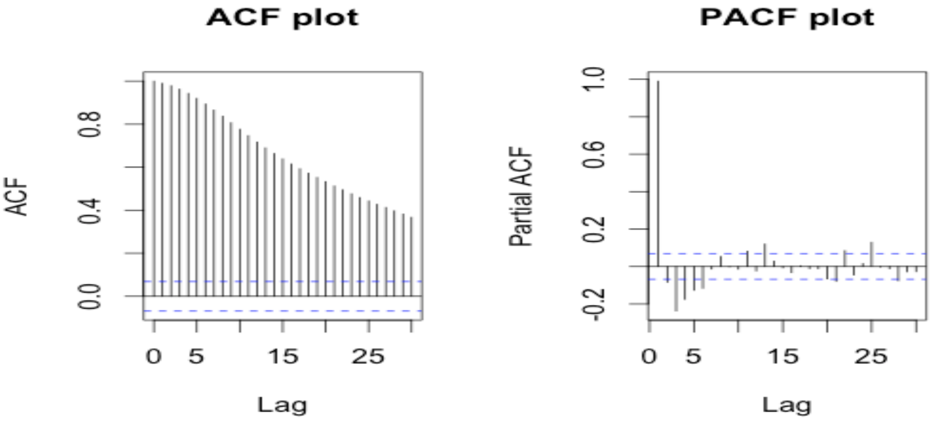
3.2. Pattern discovery of Unemployment rate

In order to build a model to fit the data, we need to firstly check the assumptions of non-stationarity and seasonal pattern in the data.

Discovery of non-stationarity

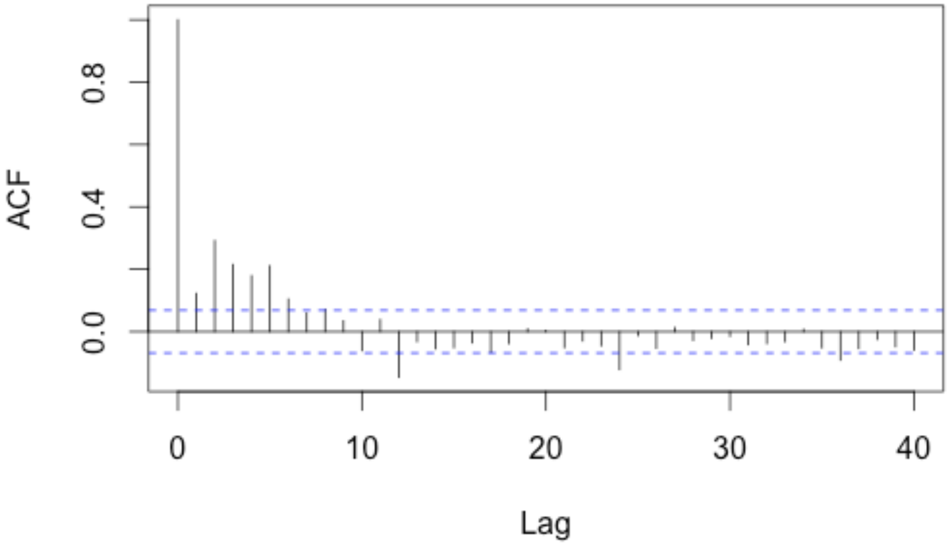
We tested white noise hypothesis of the unemployment rate by Ljung-Box test at lag 3, 6, 20, and found out we could reject the white noise, and therefore there is serial correlation in the data. However, we have observed a slightly trend in the data, using Dickey-fuller test, we cannot reject non-stationary at lag 3, 6, 20. And non-stationarity is also confirmed in Figure 4, where there is slow decay in ACF plot.

Figure 4: ACF and PACF plot of unemployment rate



Discovery of seasonal pattern
To confirm if there is seasonal pattern in the data, we plot the autocorrelation of first difference of unemployment rate in Figure 5. And it seems there is a repeated pattern at lag 12, 24, 36, which is a yearly pattern in the dat. Thus, the seasonality is confirmed in the data.

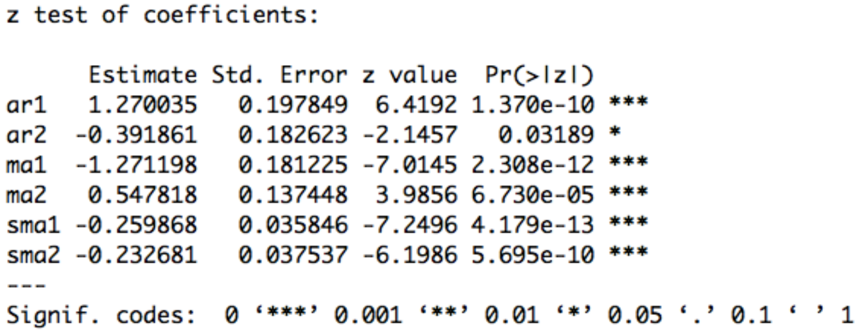
Figure 5: ACF plot of first difference of unemployment rate



3.3. Model fitting and selection

3.3.1 model fitting on time series object unemployment rate

Firstly, we fit auto.arima model on the unemployment rate with seasonality trend, and the recommended model is ARIMA(2,1,2)(0,0,2)[12], which is model m1 with BIC as -357.05.
The summary of model m1 is as below, all the coefficients seems significant to be included in the model.



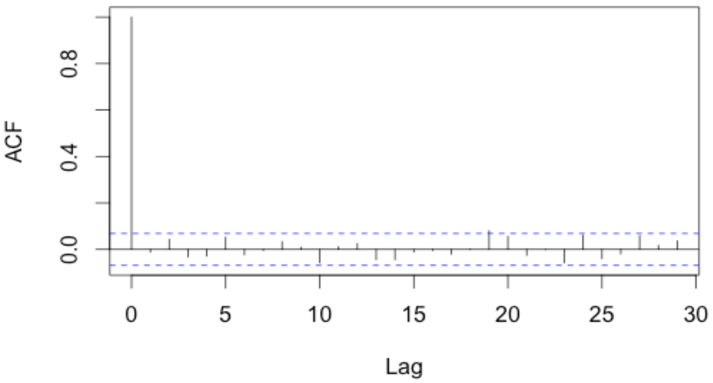
```
z test of coefficients:

      Estimate Std. Error z value  Pr(>|z|)    
ar1   1.270035   0.197849  6.4192 1.370e-10 ***
ar2  -0.391861   0.182623 -2.1457   0.03189 *  
ma1  -1.271198   0.181225 -7.0145 2.308e-12 ***
ma2   0.547818   0.137448  3.9856 6.730e-05 ***
sma1 -0.259868   0.035846 -7.2496 4.179e-13 ***
sma2 -0.232681   0.037537 -6.1986 5.695e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Diagnostics of model m1
To check if the model m1 is adequate, we used Ljung-Box test at lag 12, 15, and the white noise assumption of the residuals could not be rejected.

From the acf plot of the residuals, Figure 6, there is a relative significant correlation in the residuals at lag 19. Therefore we could say model m1 is relatively adequate, but probably not the best choice.

Figure 6: ACF plot of residuals of model m1

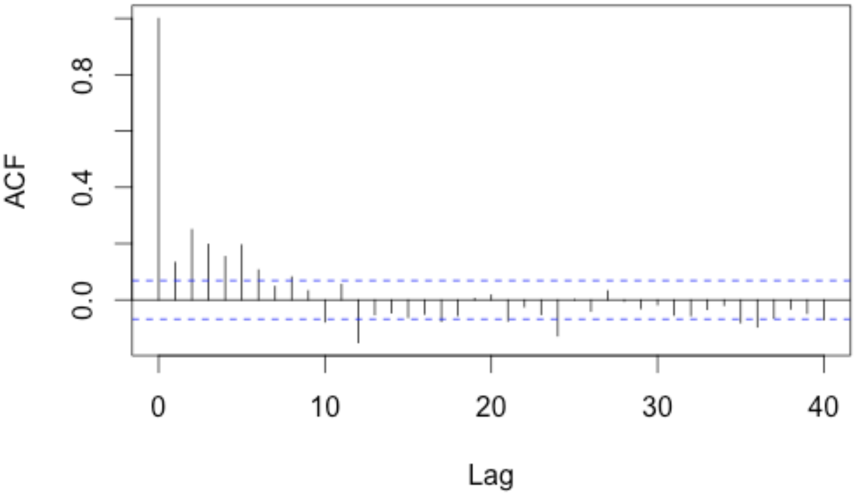


3.3.2 model fitting on logarithm of time series object unemployment rate

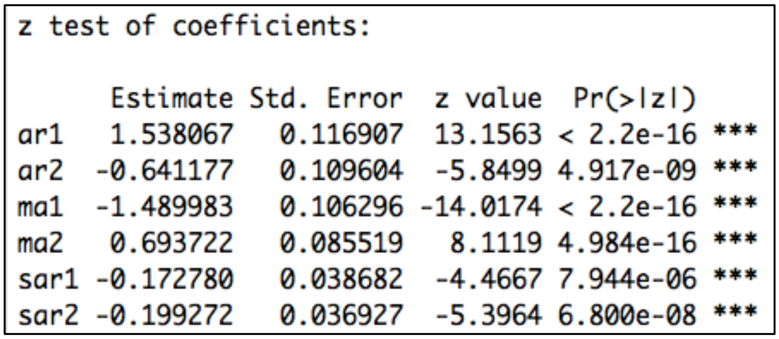
As from the time plot of unemployment plot, we observed slight trend in the data, therefore, we will try to transform the data by logarithm and fit a model to check if there is any improvement.

Applying the same analysis of unemployment rate, we analyzed logarithm of unemployment rate; for example, the ACF plot of first difference of logarithm of unemployment rate is as in figure 7. From the plot, we could observe there is significant correlation at lag 12,24,36; therefore, the logarithm data also shows seasonality.

Figure 7: ACF plot of first difference of logarithm of unemployment rate



Then, we fit auto.arima model on the logarithm of unemployment rate with seasonality trend, and the recommended model is ARIMA(2,1,2)(2,0,0)[12] , which is model m1 with BIC as -3104.7.
The summary of model m1 is as below, all the coefficients seems significant to be included in the model.

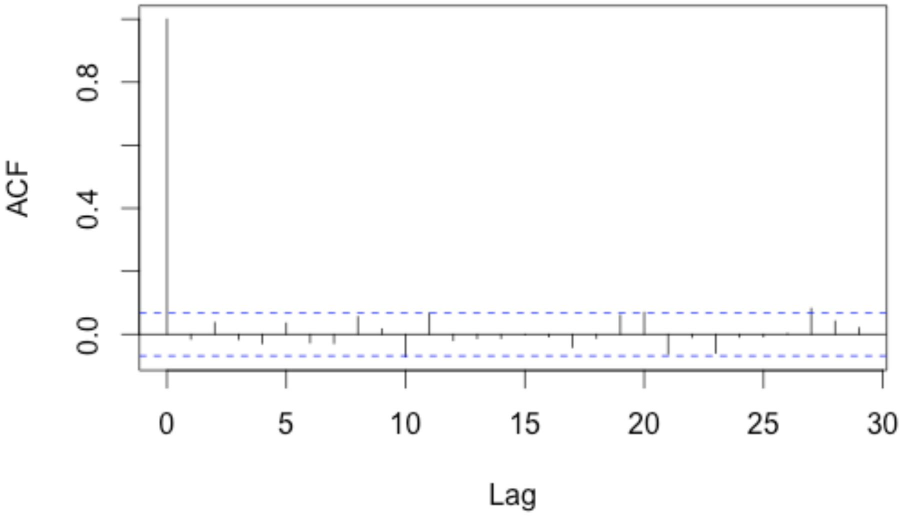


```
z test of coefficients:

      Estimate Std. Error  z value  Pr(>|z|)    
ar1   1.538067   0.116907  13.1563 < 2.2e-16 ***
ar2  -0.641177   0.109604  -5.8499 4.917e-09 ***
ma1  -1.489983   0.106296 -14.0174 < 2.2e-16 ***
ma2   0.693722   0.085519   8.1119 4.984e-16 ***
sar1 -0.172780   0.038682  -4.4667 7.944e-06 ***
sar2 -0.199272   0.036927  -5.3964 6.800e-08 ***
```

Diagnostics of model m2
To check if the model m1 is adequate, we used Ljung-Box test at lag 12, 15, and the white noise assumption of the residuals could be rejected at lag 12, but not at lag 15. From the acf plot of the residuals, Figure 8, there is a relative significant correlation in the residuals at lag 27. Therefore we could say model m2 is relatively adequate, but not a perfect match.

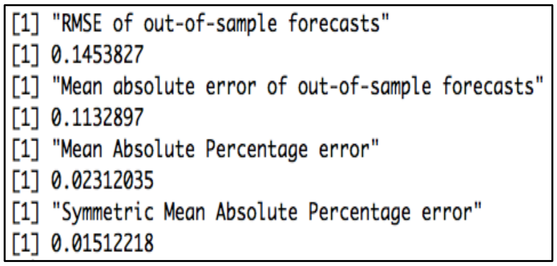
Figure 8: ACF plot of residuals of model m2



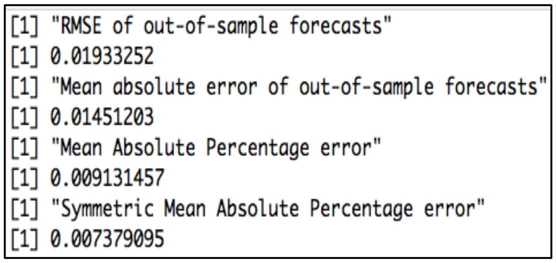
3.3.3 Model selection by BIC and Back testing method

By BIC criteria, the BIC value of model m2 is -3104.7, which is much smaller than the BIC value of model m1 at -357.05. It seems model m2 is preferred.

By Back testing method, we get the metrics for model m1 and m2 as below:
Figure 9: Back testing result of model m1(left) and m2(right)



```
[1] "RMSE of out-of-sample forecasts"
[1] 0.1453827
[1] "Mean absolute error of out-of-sample forecasts"
[1] 0.1132897
[1] "Mean Absolute Percentage error"
[1] 0.02312035
[1] "Symmetric Mean Absolute Percentage error"
[1] 0.01512218
```



```
[1] "RMSE of out-of-sample forecasts"
[1] 0.01933252
[1] "Mean absolute error of out-of-sample forecasts"
[1] 0.01451203
[1] "Mean Absolute Percentage error"
[1] 0.009131457
[1] "Symmetric Mean Absolute Percentage error"
[1] 0.007379095
```

The mean absolute percentage error for the model m1 is 2.3%, for model m2 is 0.9%, which is smaller than the MAPE of m1.

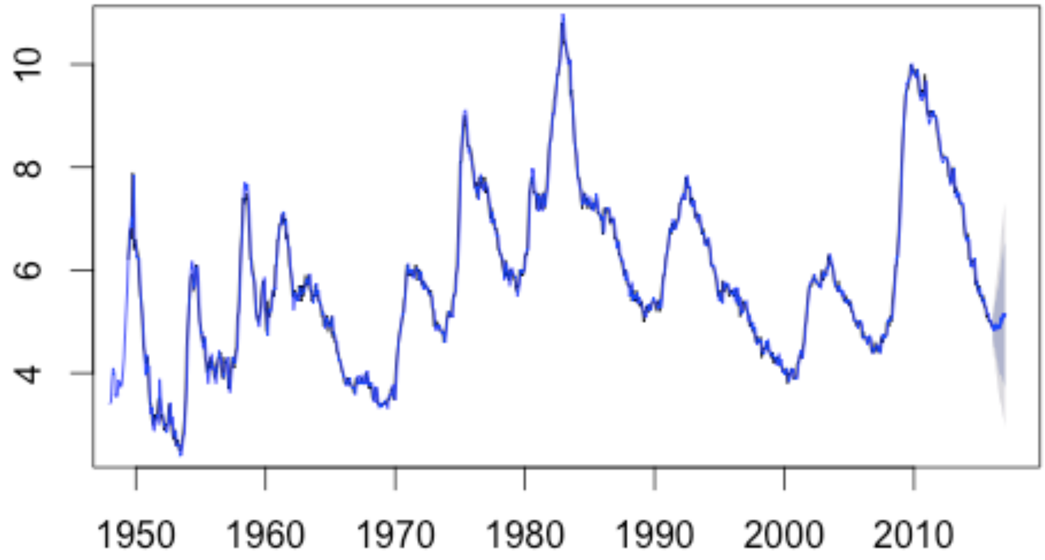
So from both the BIC criteria and the result of Back-testing, we prefer model m2. The model could be write in back shift operator as:

$$\mathbf{(1- 1.54B +0.64B^2)\ (1-B)\ (1+0.17B^{12}+0.2B^{24}) * \log (X_t)}$$
$$\mathbf{=(1-1.49B+0.69B^2) * e_t}$$

3.4. Forecast for the next 12 month

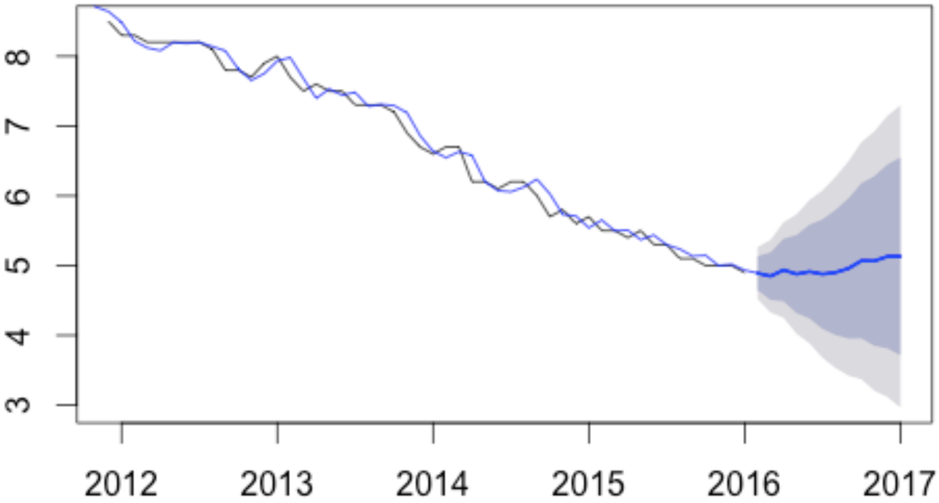
We make a forecast into future 12 month using the model selected, m2. The result is as in figure 9, blue line is the forecasted value and solid black line is the observed value. We could see the two lines are highly overlapped, which means our choice of model is very decent.

Figure 9: Forecasts from year 1948 to 2017 using model m2



By a closer look in the forecast for the next 12-month from February of 2016 to January of 2017 in the Figure 10, we could say the unemployment rate in year 2016 is showing a slightly upward trend.

Figure 10: Forecasts from Year 2012 to 2017.

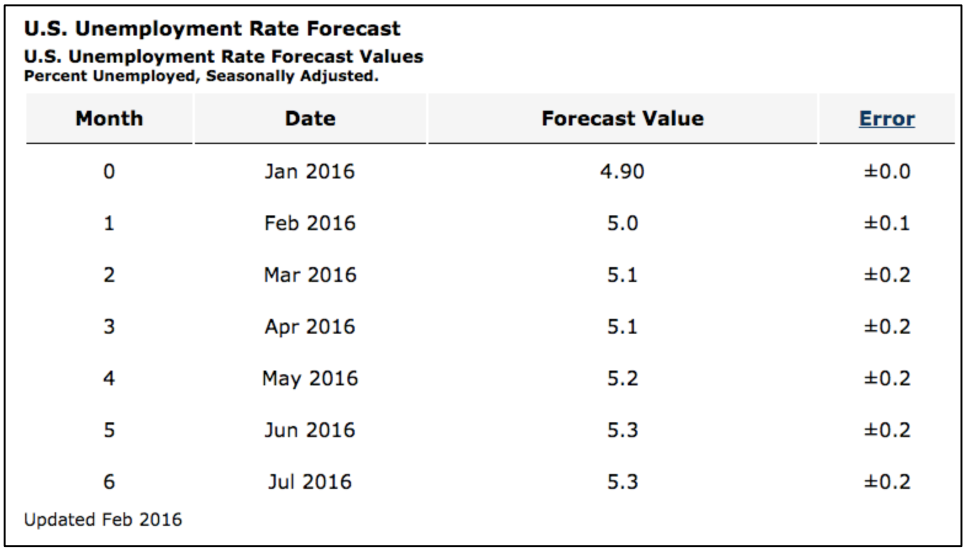


The forecasted unemployment rate for the next 12-month in original scale:

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2016		4.87	4.84	4.89	4.85	4.89	4.86	4.89	4.92	4.99	4.97	5.01
2017	5.01											

From the forecasted value of the unemployment rate, we would suggest anyone who is graduating try to find a job as soon as possible. It seems next year the overall unemployment rate is increasing.

To validate our conclusion, the following is the forecast from the financial forecast center [3]. We could see even though the value is slightly different from our dataset, but the overall trend of unemployment rate for this year is going to increase.



U.S. Unemployment Rate Forecast

U.S. Unemployment Rate Forecast Values

Percent Unemployed, Seasonally Adjusted.

Month	Date	Forecast Value	Error
0	Jan 2016	4.90	±0.0
1	Feb 2016	5.0	±0.1
2	Mar 2016	5.1	±0.2
3	Apr 2016	5.1	±0.2
4	May 2016	5.2	±0.2
5	Jun 2016	5.3	±0.2
6	Jul 2016	5.3	±0.2

Updated Feb 2016

4. Conclusions

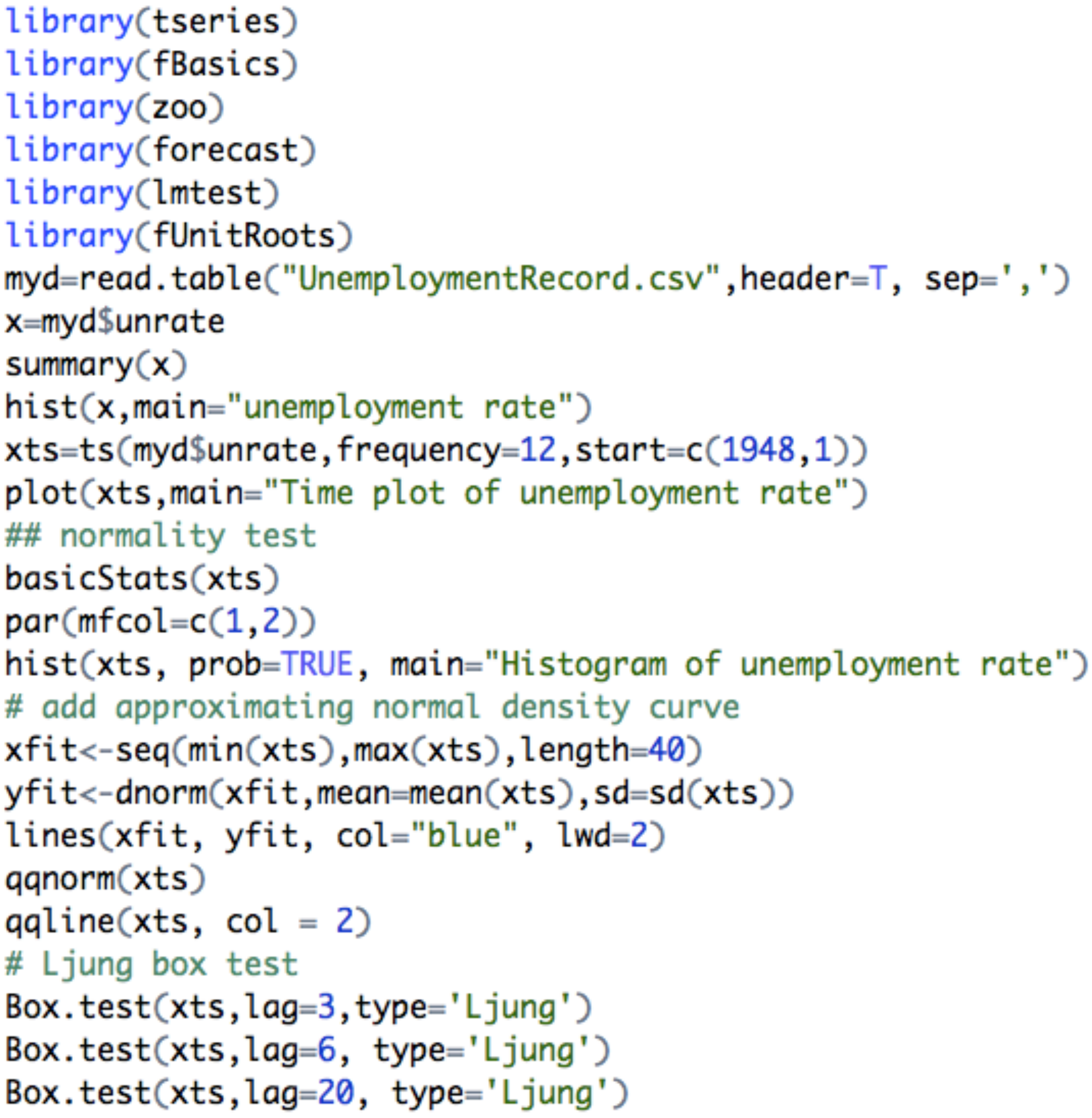
At this point, we have achieved our goal of finding a match for our data and make a forecast using the model. From our selected model m2, ARIMA (2,1,2)(2,0,0)[12] with logarithm transformation of the original data, it seems the unemployment rate is going to increase the following months, not good news to students who are graduating this summer. But the good news is the overall forecast is going to merge to the mean in the long run. So timing of finding a job is very important.

References

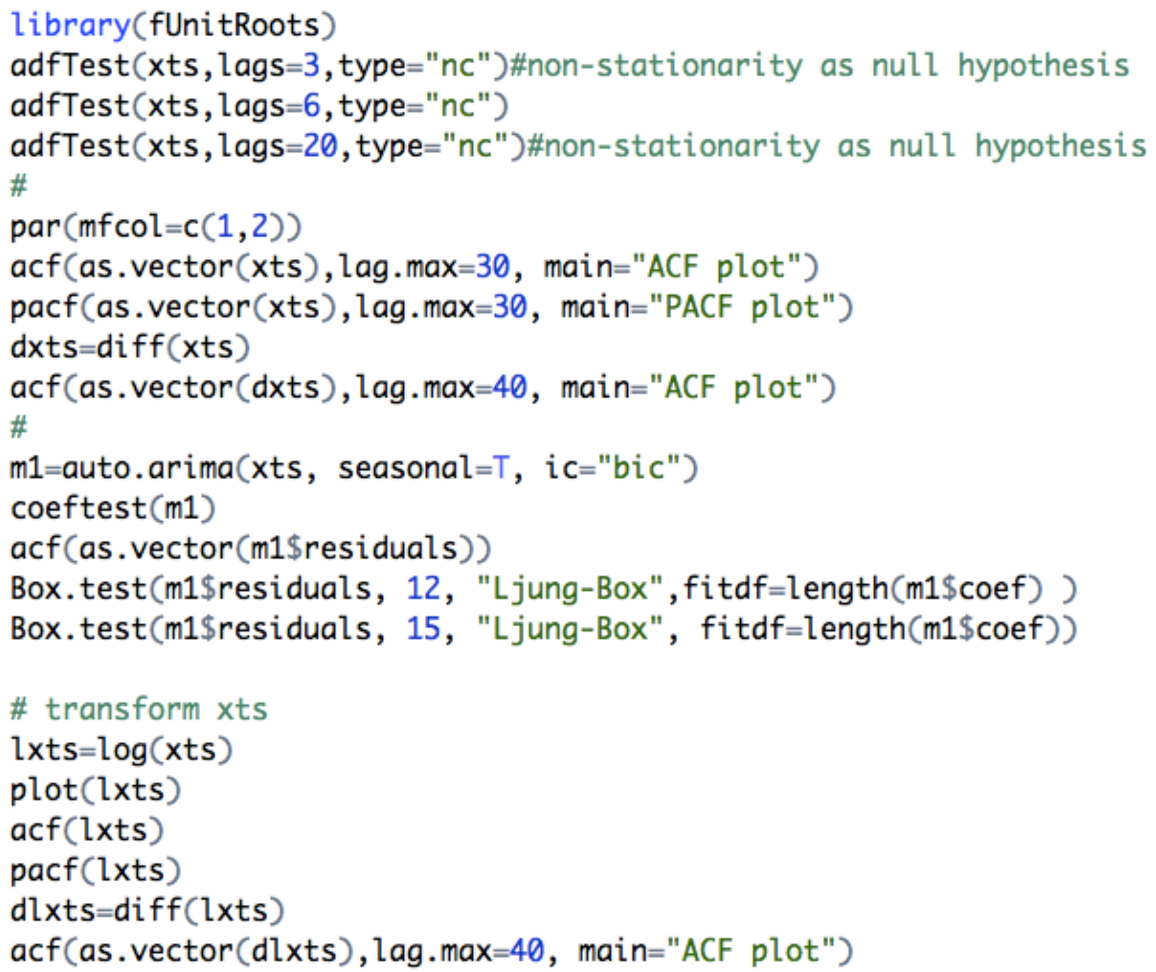
1. The Federal Reserve bank of St. Louis, https://research.stlouisfed.org/fred2/series/UNRATE?catbc=1&utm_expid=19978471-2.Y0NpAPxIQfK_8K7-O4DTQg.1&utm_referrer=https%3A%2F%2Fresearch.stlouisfed.org%2Ffred2%2Frelease%3Frid%3D50#
2. CNN news, http://money.cnn.com/2016/02/06/news/economy/obama-us-jobs/
3. The financial forecast center http://www.forecasts.org/unemploy.htm

Appendix

R code.

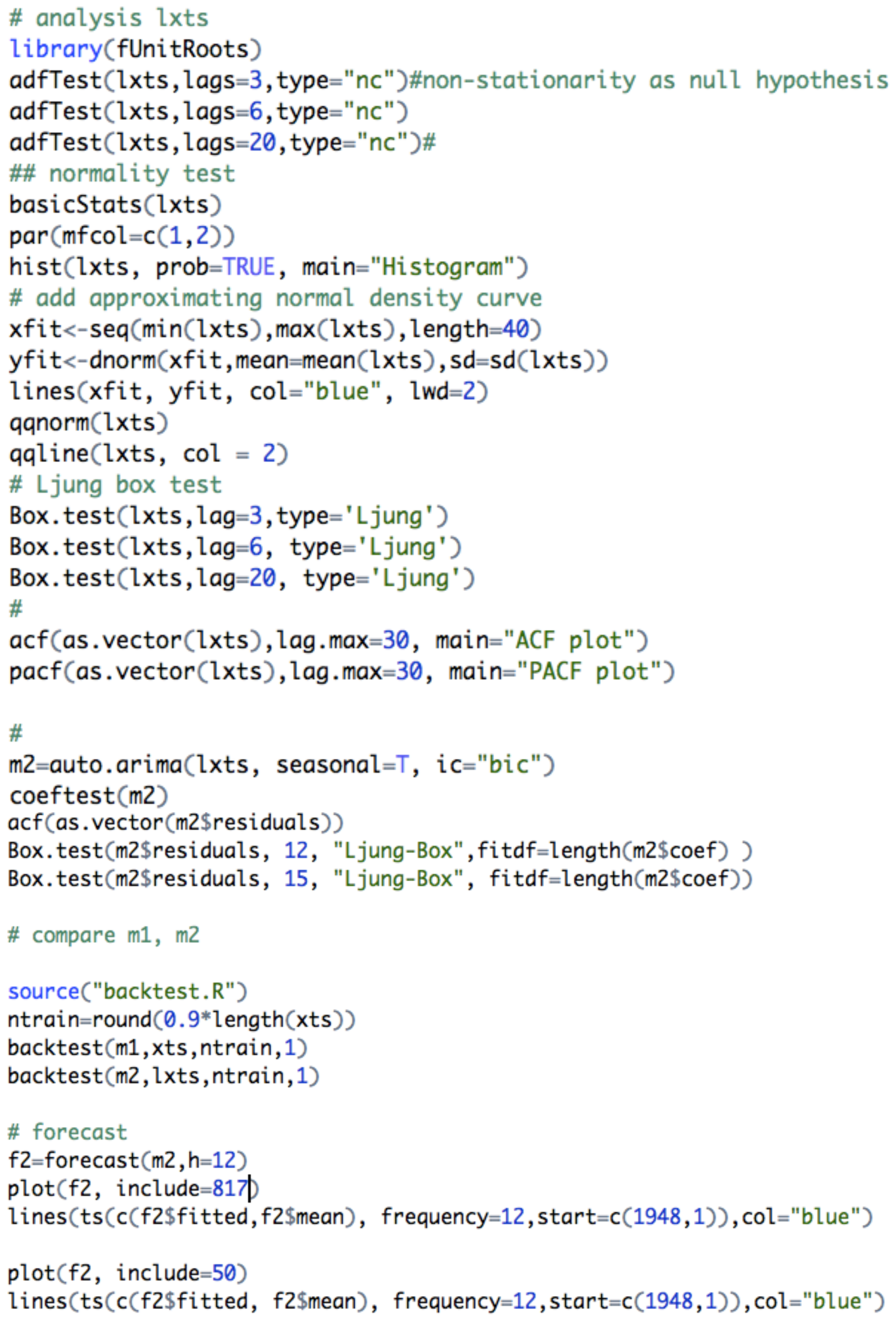


```
library(tseries)
library(fBasics)
library(zoo)
library(forecast)
library(lmtest)
library(fUnitRoots)
myd=read.table("UnemploymentRecord.csv",header=T, sep=',')
x=myd$unrate
summary(x)
hist(x,main="unemployment rate")
xts=ts(myd$unrate,frequency=12,start=c(1948,1))
plot(xts,main="Time plot of unemployment rate")
## normality test
basicStats(xts)
par(mfcol=c(1,2))
hist(xts, prob=TRUE, main="Histogram of unemployment rate")
# add approximating normal density curve
xfit<-seq(min(xts),max(xts),length=40)
yfit<-dnorm(xfit,mean=mean(xts),sd=sd(xts))
lines(xfit, yfit, col="blue", lwd=2)
qqnorm(xts)
qqline(xts, col = 2)
# Ljung box test
Box.test(xts,lag=3,type='Ljung')
Box.test(xts,lag=6, type='Ljung')
Box.test(xts,lag=20, type='Ljung')
```



```r
library(fUnitRoots)
adfTest(xts,lags=3,type="nc")#non-stationarity as null hypothesis
adfTest(xts,lags=6,type="nc")
adfTest(xts,lags=20,type="nc")#non-stationarity as null hypothesis
#
par(mfcol=c(1,2))
acf(as.vector(xts),lag.max=30, main="ACF plot")
pacf(as.vector(xts),lag.max=30, main="PACF plot")
dxts=diff(xts)
acf(as.vector(dxts),lag.max=40, main="ACF plot")
#
m1=auto.arima(xts, seasonal=T, ic="bic")
coeftest(m1)
acf(as.vector(m1$residuals))
Box.test(m1$residuals, 12, "Ljung-Box",fitdf=length(m1$coef) )
Box.test(m1$residuals, 15, "Ljung-Box", fitdf=length(m1$coef))

# transform xts
lxts=log(xts)
plot(lxts)
acf(lxts)
pacf(lxts)
dlxts=diff(lxts)
acf(as.vector(dlxts),lag.max=40, main="ACF plot")
```



```
# analysis lxts
library(fUnitRoots)
adfTest(lxts,lags=3,type="nc")#non-stationarity as null hypothesis
adfTest(lxts,lags=6,type="nc")
adfTest(lxts,lags=20,type="nc")#
## normality test
basicStats(lxts)
par(mfcol=c(1,2))
hist(lxts, prob=TRUE, main="Histogram")
# add approximating normal density curve
xfit<-seq(min(lxts),max(lxts),length=40)
yfit<-dnorm(xfit,mean=mean(lxts),sd=sd(lxts))
lines(xfit, yfit, col="blue", lwd=2)
qqnorm(lxts)
qqline(lxts, col = 2)
# Ljung box test
Box.test(lxts,lag=3,type='Ljung')
Box.test(lxts,lag=6, type='Ljung')
Box.test(lxts,lag=20, type='Ljung')
#
acf(as.vector(lxts),lag.max=30, main="ACF plot")
pacf(as.vector(lxts),lag.max=30, main="PACF plot")

#
m2=auto.arima(lxts, seasonal=T, ic="bic")
coeftest(m2)
acf(as.vector(m2$residuals))
Box.test(m2$residuals, 12, "Ljung-Box",fitdf=length(m2$coef) )
Box.test(m2$residuals, 15, "Ljung-Box", fitdf=length(m2$coef))

# compare m1, m2

source("backtest.R")
ntrain=round(0.9*length(xts))
backtest(m1,xts,ntrain,1)
backtest(m2,lxts,ntrain,1)

# forecast
f2=forecast(m2,h=12)
plot(f2, include=817)
lines(ts(c(f2$fitted,f2$mean), frequency=12,start=c(1948,1)),col="blue")

plot(f2, include=50)
lines(ts(c(f2$fitted, f2$mean), frequency=12,start=c(1948,1)),col="blue")
```