Name: Xiao Li
ID: 904236938
STATS 170

Forecasting Average Duration of Unemployment in US by using Civilian Unemployment Rate, Number of Civilians Unemployed for 27 weeks and over and Civilian Labor Force Participation Rate.

I. Introduction:

In this paper, I forecast the average duration of unemployment[1] in US. I hypothesize that there is a positive relationship between number of civilians unemployed for 27 weeks and over[2] and average duration of unemployment in US. Also, there may be a positive relationship between civilian unemployment rate[3] and average duration of unemployment[1]. Moreover, I expect that the lower the civilian labor force participation rate[4], the shorter the average duration of unemployment[1]. A seasonal effect is expected since when school year ends, more students are going to find jobs which increasing the labor force, hence increase the unemployment rate.

In order to construct the model, I collected data from https://fred.stlouisfed.org/. In this https://fred.stlouisfed.org/, we can see that it says that the data are not seasonally adjusted. The following section describes the variables selected to answer the research question. So I have four variables: average duration of unemployment[1], civilian unemployment rate3) , number of civilians unemployed for 27 weeks and over[2], and civilian labor force participation rate[4].

II. The Data:

Here, I describe the variables used for the analysis. I give a short name to each of the variables used.

The variable **AD[1]** measures average duration of unemployment in US[1] . It was obtained from U.S. Bureau of Labor Statistics, https://fred.stlouisfed.org/series/LNU03008275.

The variable **N27[2]** measures number of civilians unemployed for 27 weeks and over in US[2] . It was obtained from U.S. Bureau of Labor Statistics. https://fred.stlouisfed.org/series/LNU03008636.

The variable **CUR[3]** measures civilian unemployment rate in US[3] . It was obtained from U.S. Bureau of Labor Statistics. https://fred.stlouisfed.org/series/UNRATENSA.

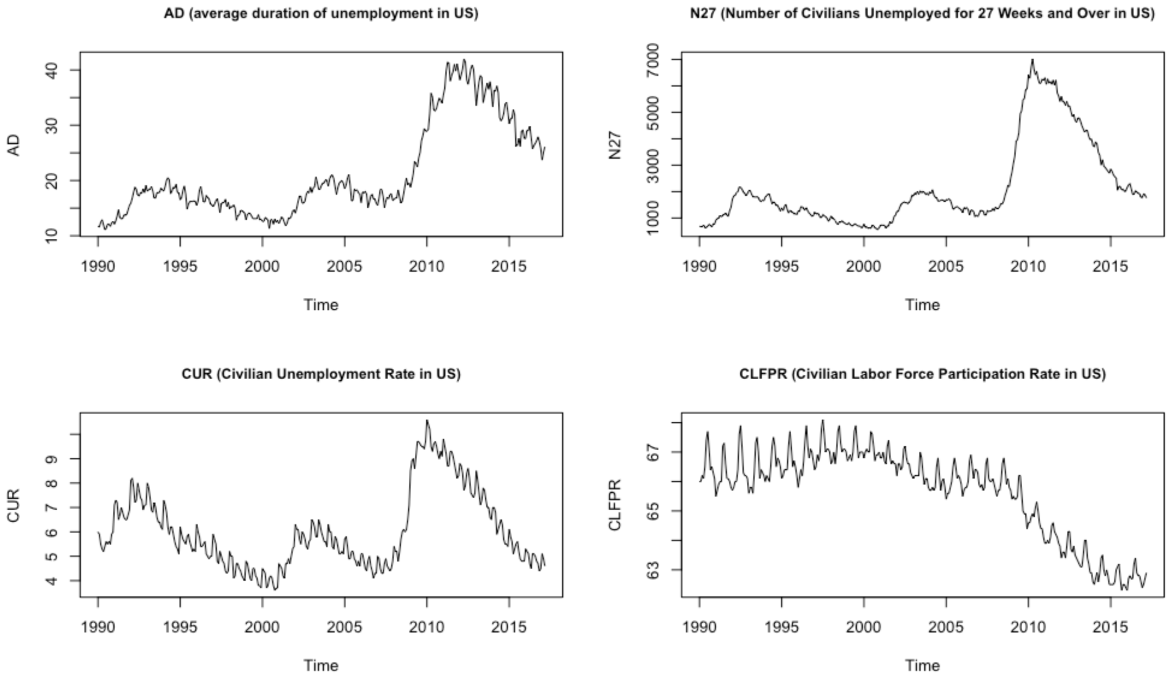
The variable **CLFPR[4]** measures civilian labor force participation rate in US[4] . It was obtained from U.S. Bureau of Labor Statistics. https://fred.stlouisfed.org/series/LNU01300000.

Table 1: Summary Statistics for All Variables

Short Name	Description	Period	Mean	Standard Deviation
AD	average duration of unemployment in US(in weeks)	1990.1-2017.3 (327 Months)	21.002	8.528
N27	Number of Civilians Unemployed for 27 Weeks and Over in US(in Thousands of persons)	1990.1-2017.3 (327 Months)	2182.924	1662.835
CUR	Civilian Unemployment Rate in US(in Percentage)	1990.1-2017.3 (327 Months)	6.086	1.583
CLFPR	Civilian Labor Force Participation Rate in US(in Percentage)	1990.1-2017.3 (327 Months)	65.753	1.423

III. Description of the Data:

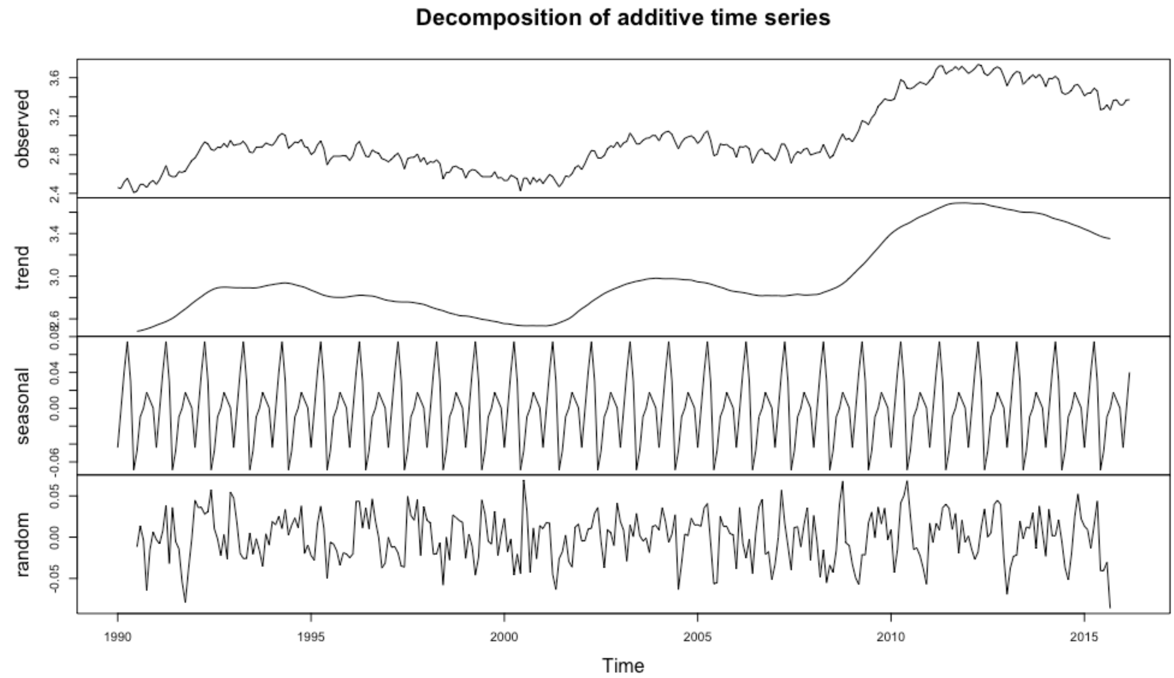
Plot 1: Time Plots of All Variables



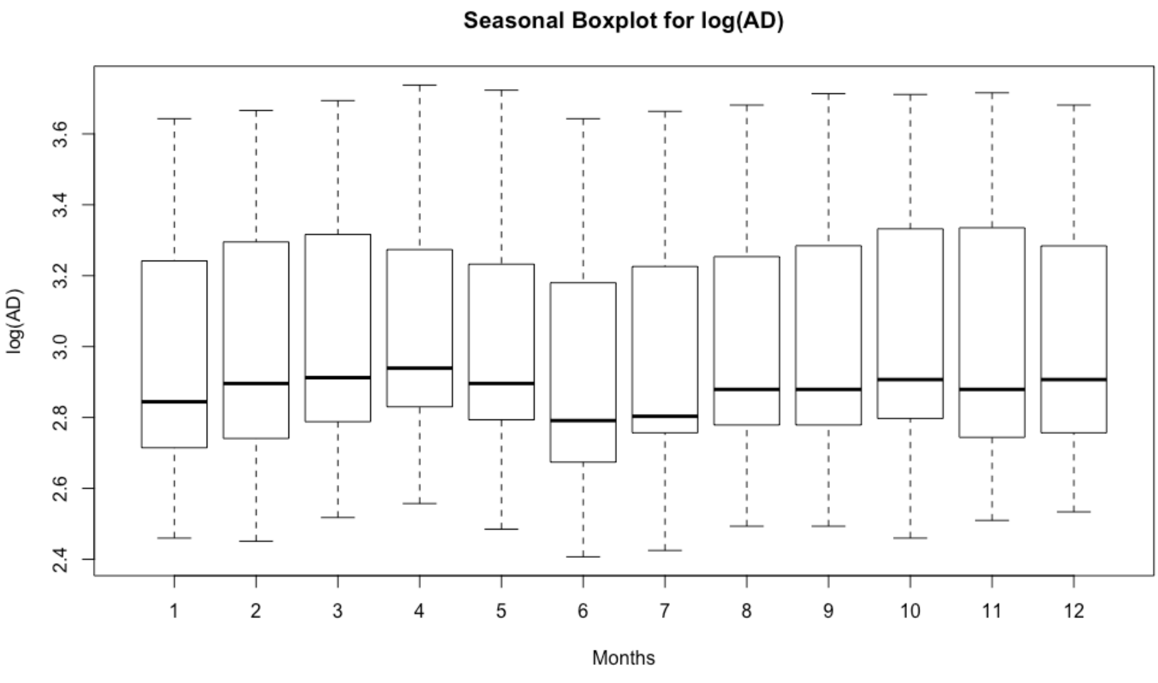
From the Plot 1, the graphs of AD, N27 and CUR show that there is slight decrease in first ten years. Then there is fluctuation. Around 2008, a steep increase takes place. After two years, a decreasing trend takes place. However, for graph of CLFPR, at the beginning, it is fluctuating. But after 2000, it shows a decreasing trend.

Furthermore, I observe that the variance seems to be proportional to the mean. Hence, the log transformation is needed. I also observe that there is seasonality in the plots.

Plot 2: Decomposition of Additive Time Series of log(AD)



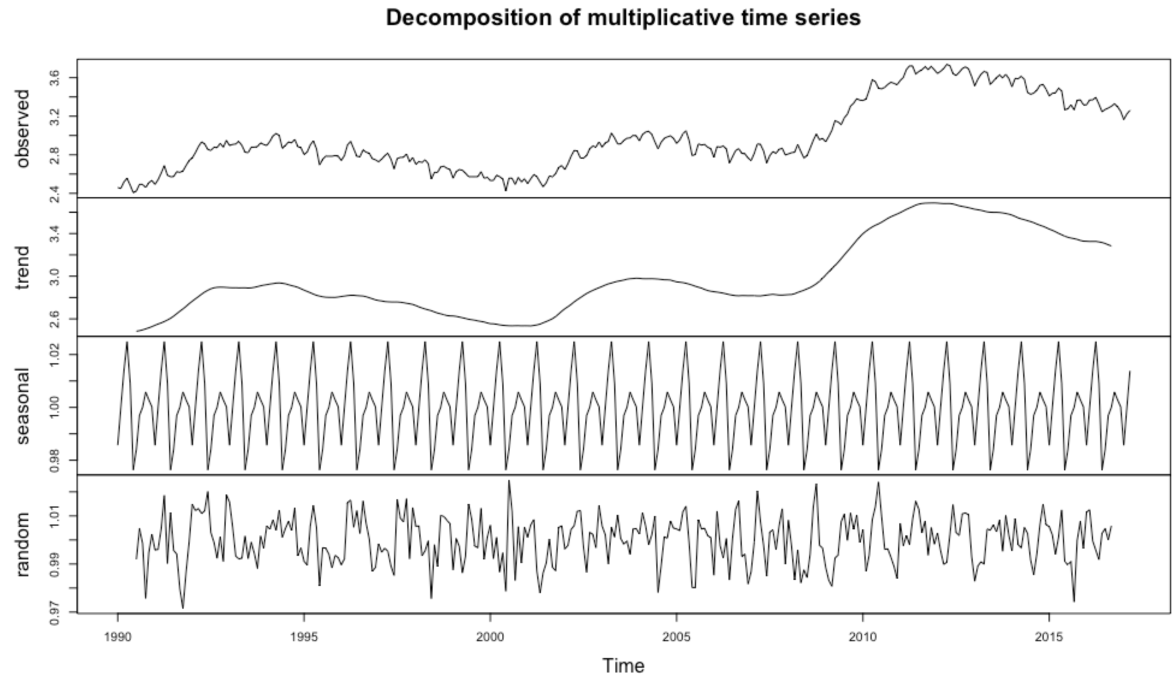
Plot 3:



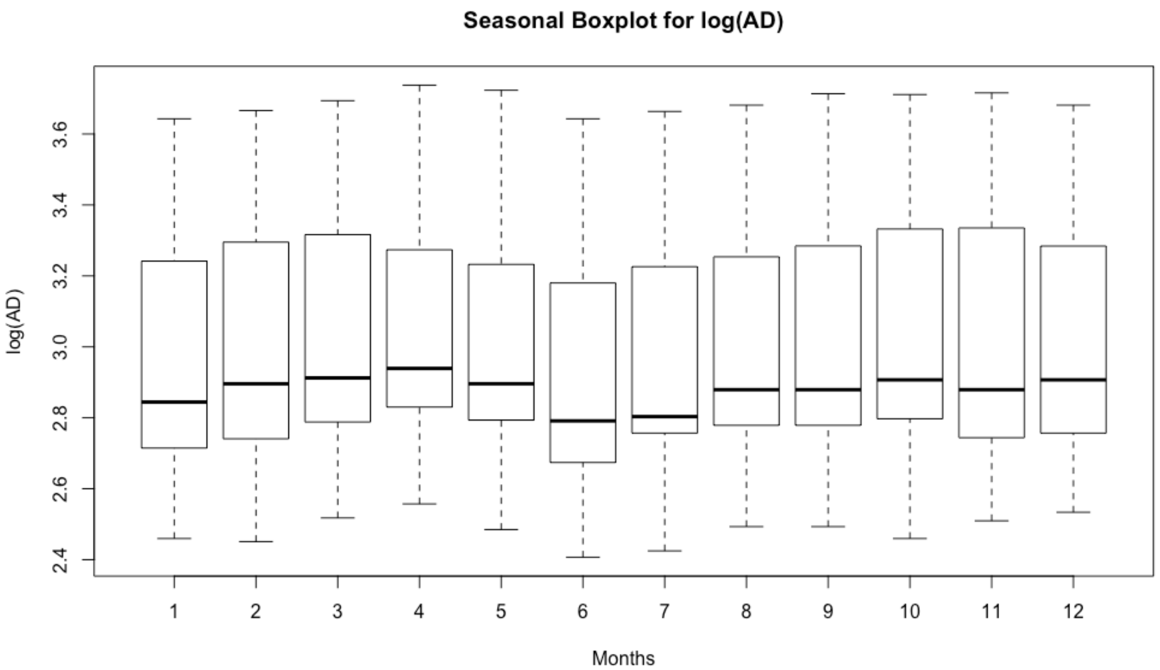
From plot 2, we can see that the trend has slight fluctuation at the beginning. But around 2008, there is a steep increase in the trend then following a downward trend after 2010.There is superimposed seasonal effect exists. But for the random components, we can see that the variance of it is decreasing then increasing. Hence, additive decomposition is not suitable here.

From plot 3, we can see that in the winter and summer, the log of average duration of unemployment has a smaller value than that in spring and autumn. This may because more companies are willing to recruit in January, June, and July.

Plot 4: Decomposition of Multiplicative Time Series of log(AD)



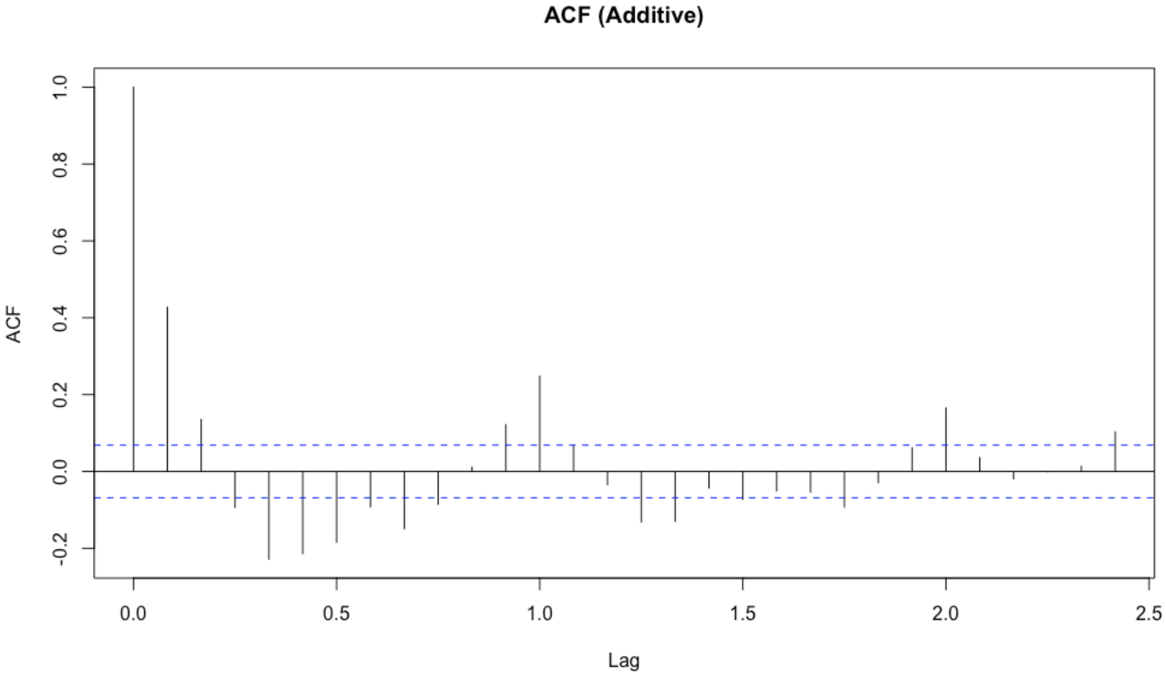
Plot 5:



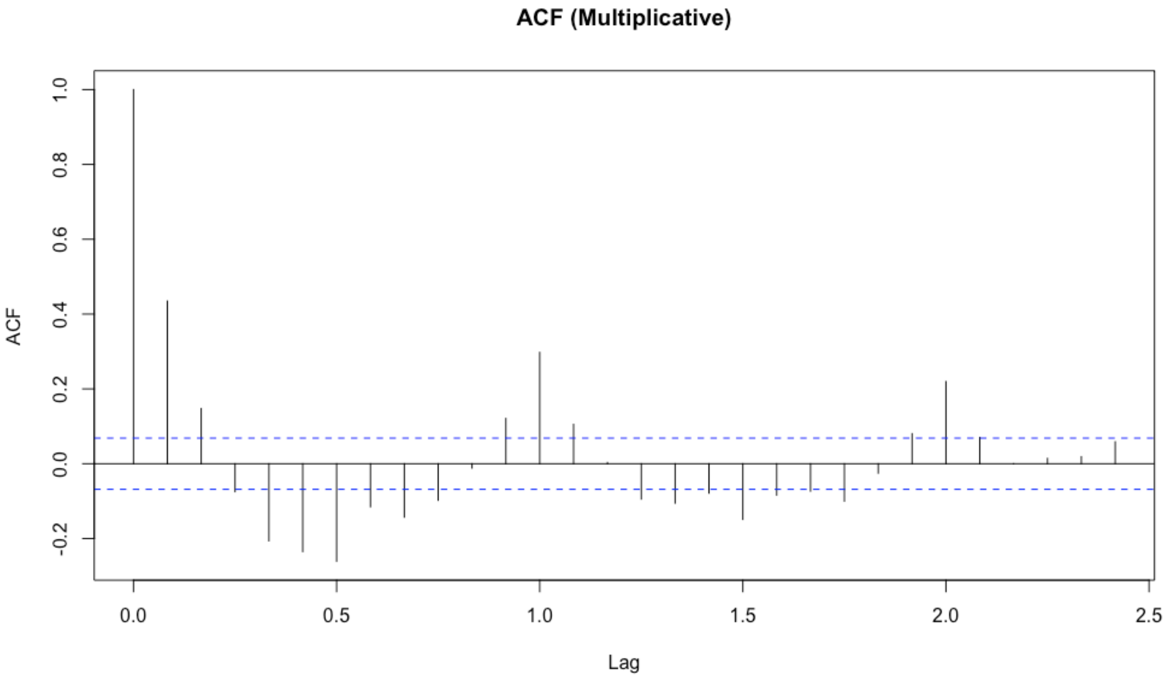
From plot 4, we can see that the trend increases at the beginning then gradually decreases. Around 2001, same pattern takes place. But around 2008, there is a steep increase in the trend then following a downward trend after 2010. There is superimposed seasonal effect exists. For random component, it is approximately stationary in mean. Hence, multiplicative decomposition is more suitable than additive decomposition.

From plot 5, we can see that in the winter and summer, the log of average duration of unemployment has a smaller value than that in spring and autumn. This may because more companies are willing to recruit in January, June, and July.

Plot 6: ACF of Additive Decomposition



Plot 7: ACF of Multiplicative Decomposition



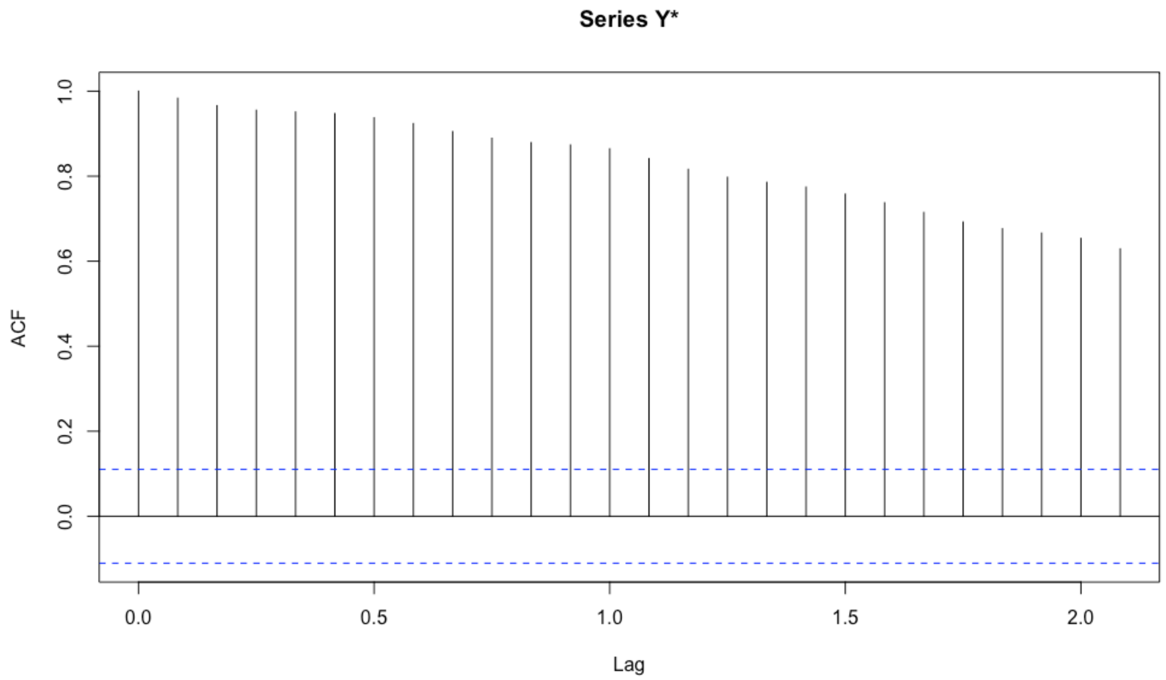
From Plot 6 and Plot 7, both ACF are dies down quickly.. Meanwhile, we can observe that at lag =12, there is a significant seasonality. Most of r_k values are significant. But ACF of Multiplicative Decomposition has a better performance than additive decomposition since after lag = 2, r_k seems to be not significant for all. In the future, we need to use differencing method or other advanced methods to improve. Perhaps, lag = 12 and differences =1 will be a good way to solve this problem.

IV. ARIMA Modelling

(a):

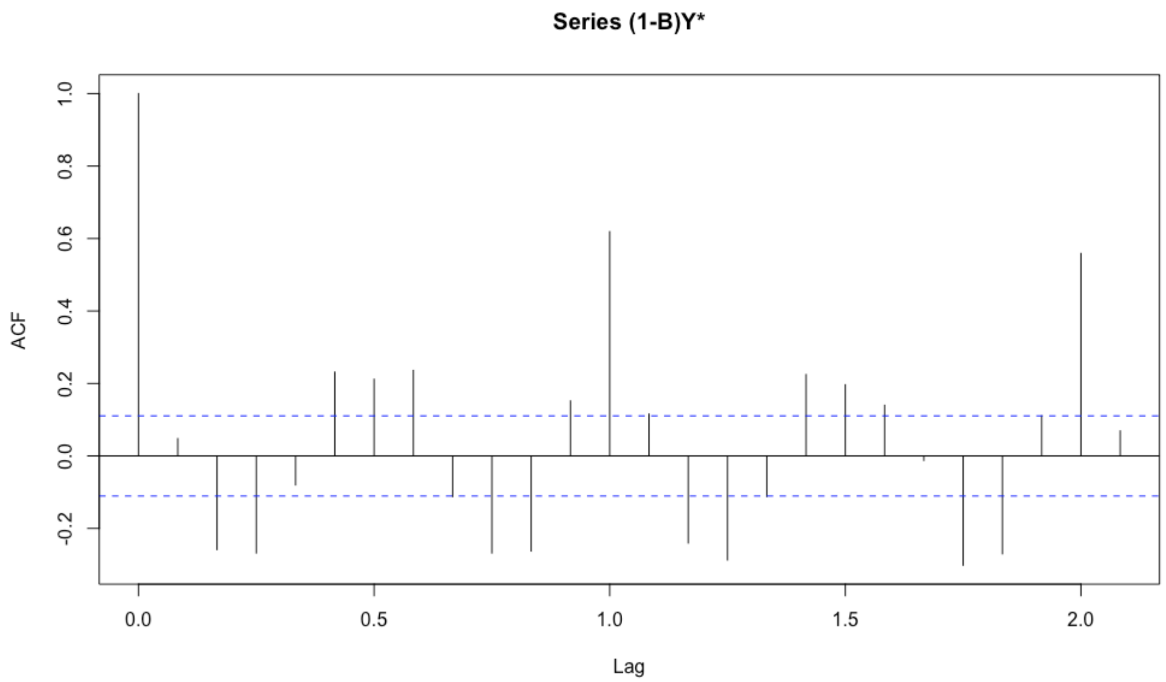
We have Yt = the average duration of unemployment in US between Jan 1990 and March 2016. According to Section III, I decided to log the data so Y* is the logged time series. In section III, by using decomposition, I found that my data has a trend and a seasonal effect. In this section IV, I will show it by using ACF plot.

Plot 8: ACF of Series Y*



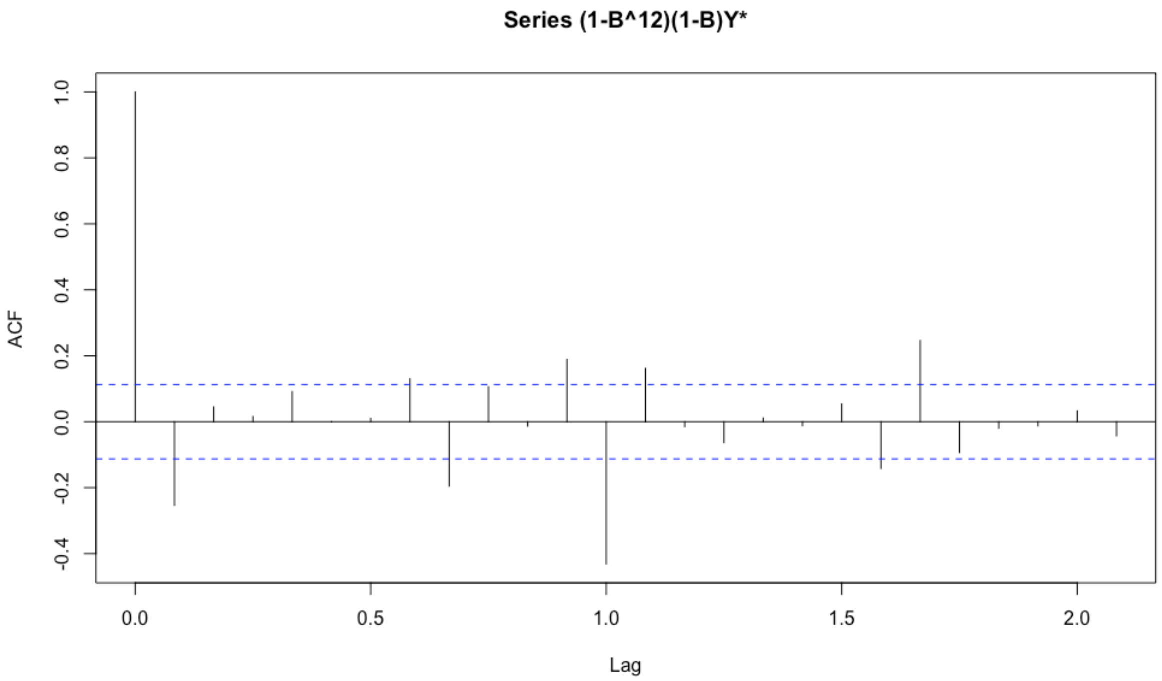
From Plot 8, we can see that there is a visible trend in the ACF. Hence, I decide to regular difference the data.

Plot 9: ACF of Series (1-B)Y*



From plot 9, a significant seasonal lag happens at lag = 12. Hence, I decide to seasonally difference the series (1-B)Y*.

Plot 10: ACF of Series $(1-B^{12})(1-B)Y^*$



From plot 10, ACF dies down quickly. There are only a few significant rk, (r1, r7, r8, r11, r12, r13, r19 and r20), left in the plot. We can say that this series is stationary now.

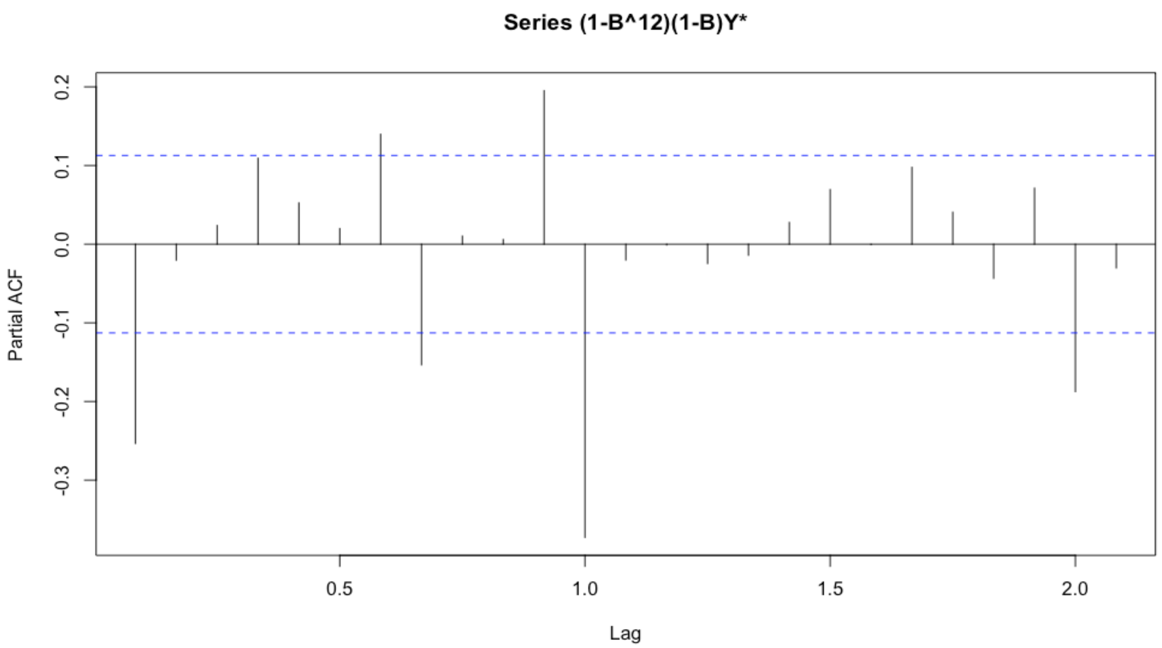
(B):

After pre-transformation and differencing, I will begin my future work with series Y** :

$$Y^{**} = (1-B^{12})(1-B)Y^*$$

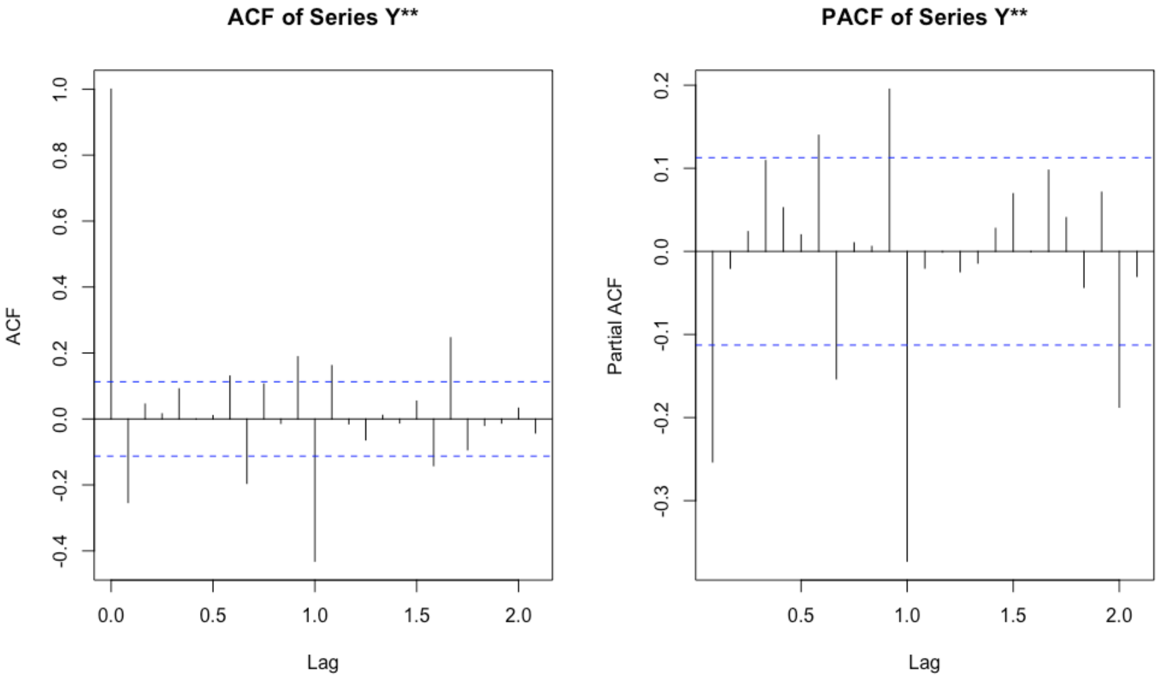
(C):

Plot 11: PACF of Series $(1-B^{12})(1-B)Y^*$



(d):

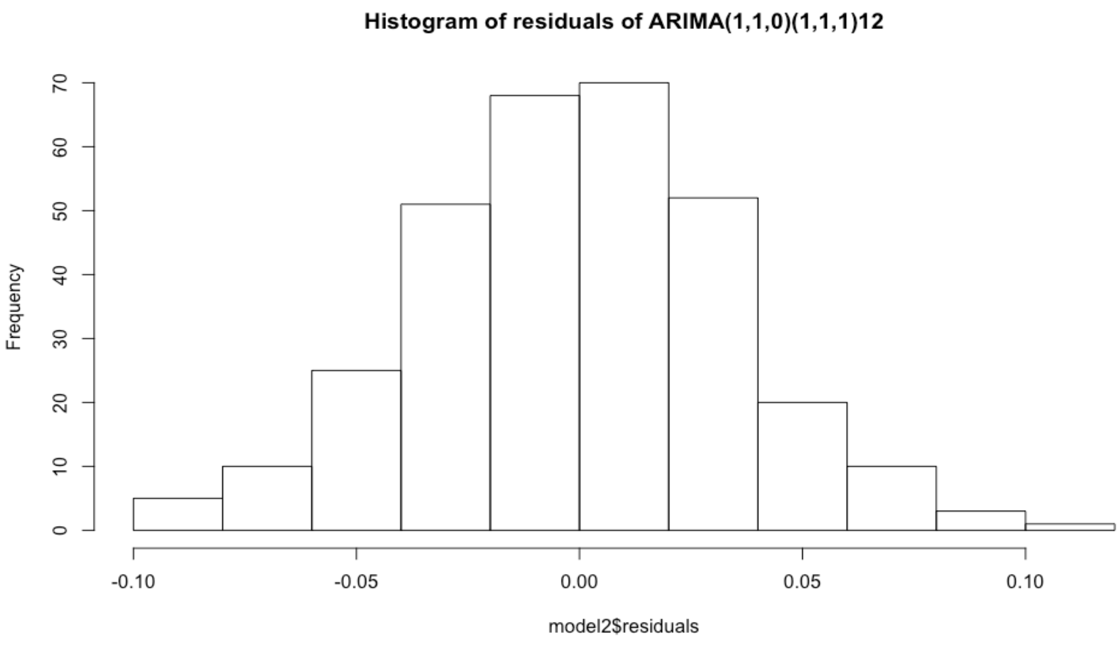
Plot 12: ACF and PACF of Y**



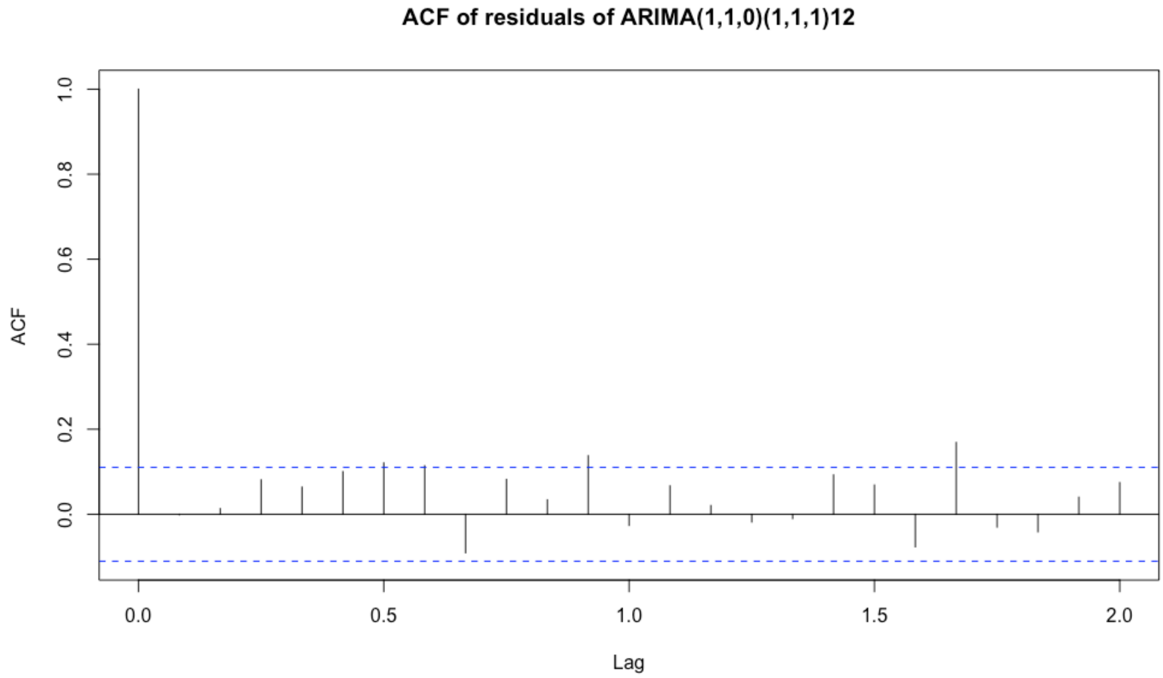
By looking at both ACF and PACF together. I cannot say exactly which one is dies down quickly or cuts off. Both ACF and PACF seem to cut off after lag = 1 or die down quickly. To better improve my model, I decide to try AR and MA both to find the best one.
By applying three different ARIMA models, ARIMA(1,1,1)(1,1,1)$_{12}$, ARIMA(1,1,1)(1,1,1)$_{12}$, and ARIMA(1,1,1)(1,1,1)$_{12}$, I get AIC scores to choose the best one. The AIC value for ARIMA(1,1,0)(1,1,1)$_{12}$ is the smallest. Hence, I choose ARIMA(1,1,0)(1,1,1)$_{12}$ as my forecast model.

(e)

Plot 13: Histogram of residuals of ARIMA(1,1,0)(1,1,1)$_{12}$



Plot 13: ACF of residuals of ARIMA(1,1,0)$(1,1,1)_{12}$



Plot 13 has shown that the distribution of residuals are very similar to normal distribution. Hence, the condition of normality of residuals is met. Plot 14 shows that the ACF of residuals cuts off after lag =1. Hence, the residuals are white noise. Also applying Ljung-box test, I get the p-value is $0.07536 > 0.05$. So we can say that residuals are independent. Also by applying t-test, I found that the absolute values of all t-values of three coefficients are greater than 2. Hence, we can reject the null and say that those coefficients are significant.

(f)
Since I have chosen ARIMA(1,1,0)$(1,1,1)_{12}$, the ARIMA polynomial notation is

$$(1-a_{12}B^{12})(1-a_1B)(1-B^{12})(1-B)Y_t^* = (1+b_{12}B^{12})w_t$$

Expanding all the terms, I get the following equation:

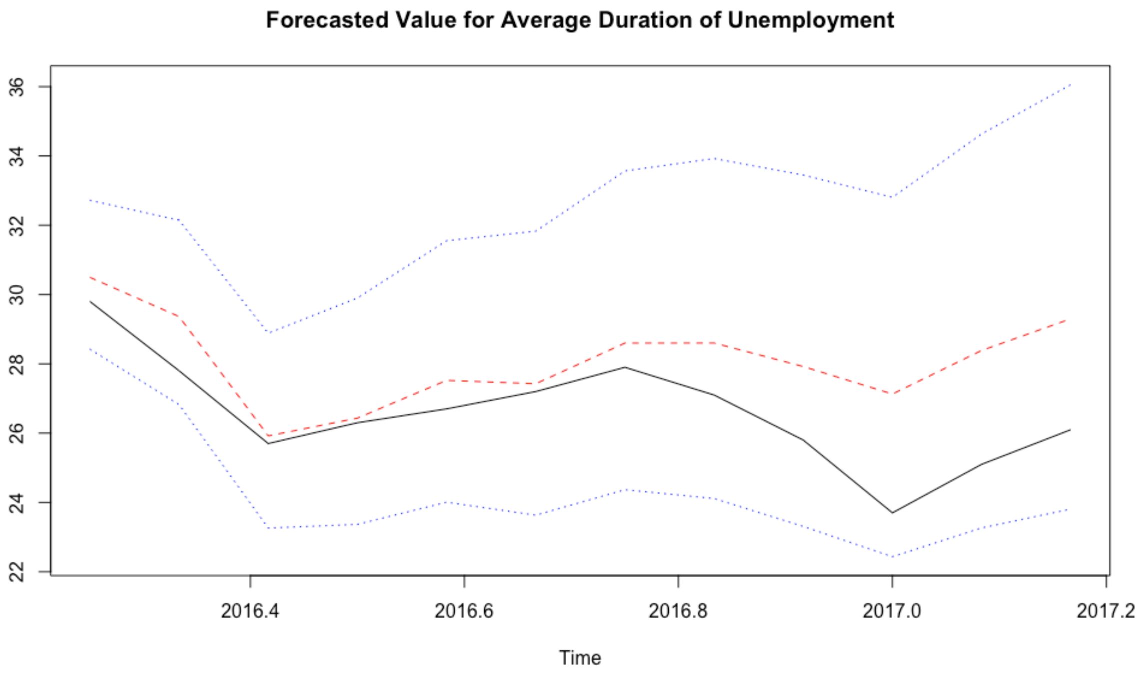
$$\begin{aligned}Y_t^* = {} & (1+a_1)Y_{t-1}^* - a_1Y_{t-2}^* + (1+a_{12})\, Y_{t-12}^* + (1+a_1+a_{12}+a_1a_{12})\, Y_{t-13}^* + (a_1+a_{12})\, Y_{t-14}^* \\ & - a_{12}\, Y_{t-24}^* + (a_{12}+a_1a_{12})\, Y_{t-25}^* - a_1a_{12}\, Y_{t-26}^* + w_t + b_{12}w_{t-12}\end{aligned}$$

Since the model used are logged data, so I need to unlog to get the prediction. So the final predicted values are:

$$Y_t = e^{Yt^*}$$

(g)

Plot 14: Forecasted Value for Average Duration of Unemployment



In plot 14, the red line represents forecasted value for average duration of unemployment. The black line represents the actual value for average duration of unemployment. And the blue lines are the 95% confidence intervals.

(h)

Table 2 : ARIMA Table

Date	Out-of-sample Actual Values	Forecasted Values
2016.4	29.8	30.5
2016.5	27.8	29.4
2016.6	25.7	25.9
2016.7	26.3	26.4
2016.8	26.7	27.5
2016.9	27.2	27.4
2016.10	27.9	28.6
2016.11	27.1	28.6
2016.12	25.8	27.9
2017.1	23.7	27.1
2017.2	25.1	28.4
2017.3	26.1	29.3
RMSE	1.9097	

V. Vector Autoregression

VI. Exponential Smoothing

VII. Time Series Regression

VIII. Final Results.

IX. References:

[1] https://fred.stlouisfed.org/series/LNU03008275

[2] https://fred.stlouisfed.org/series/LNU03008636

[3] https://fred.stlouisfed.org/series/UNRATENSA

[4] https://fred.stlouisfed.org/series/LNU01300000

[5] Cowpertwait, P.S.P. , Metcalfe, A.V. (2009). Introductory Time Series with R. Springer-Verlag

Appendix