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.

To construct the model, I will use different time series method to make the forecast about the future. The time series methods I will use are ARIMA, vector autoregreesion, exponential smoothing, and time series regression. I will first see the time plot of the data to determine what transformation I need to conduct. Furthermore, I will decide whether to do differencing or not based on stationarity. After applying different models, I will also verify the validity by checking the residuals. Finally, I will pick the best model based on the performance of difference time series method.

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** measures average duration of unemployment in US[1].

The variable **N27** measures number of civilians unemployed for 27 weeks and over in US[2].

The variable **CUR** measures civilian unemployment rate in US[3].

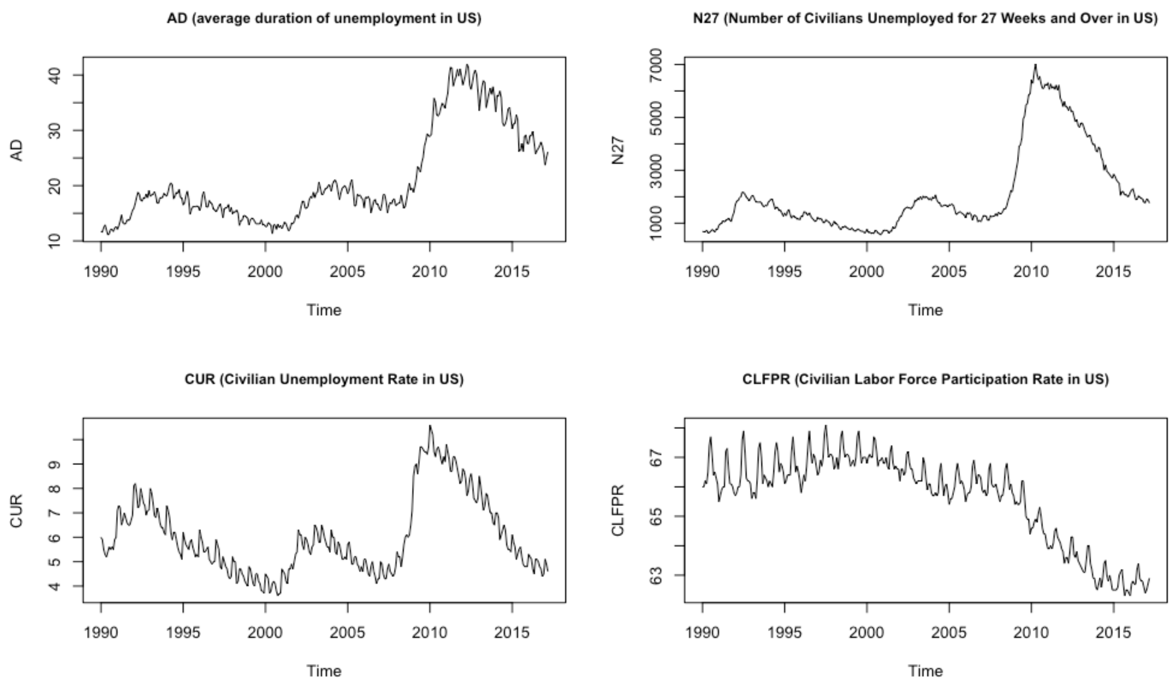
The variable **CLFPR** measures civilian labor force participation rate in US[4].

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:

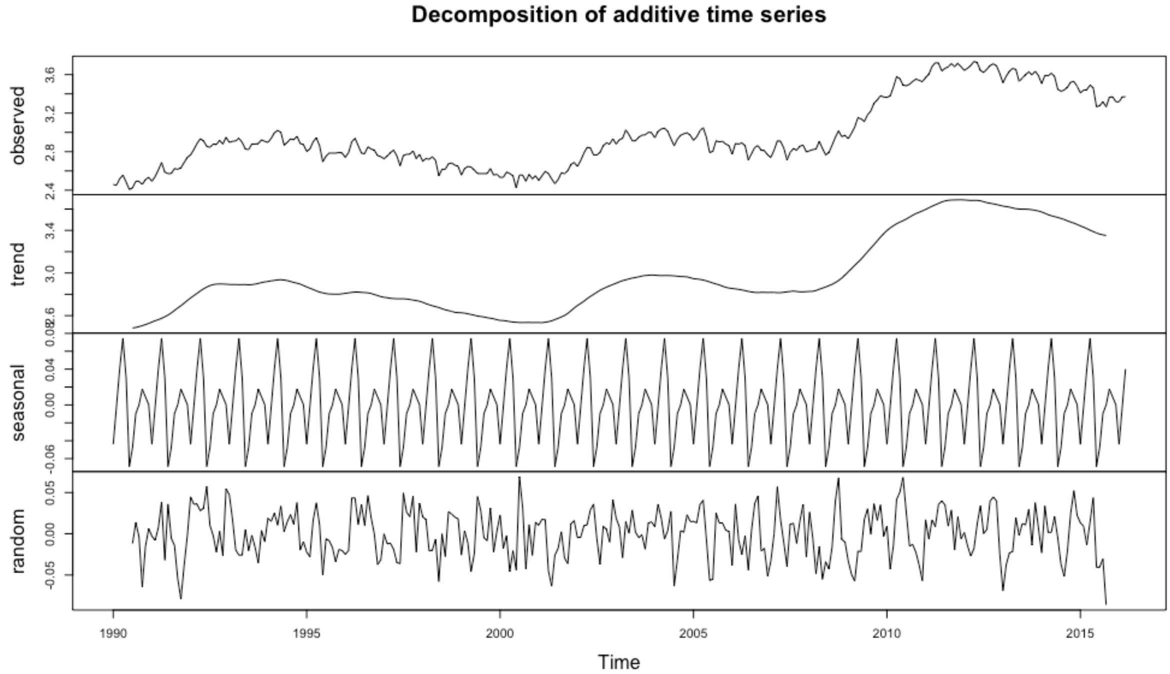
Plot 1: Time Plots of All Variables



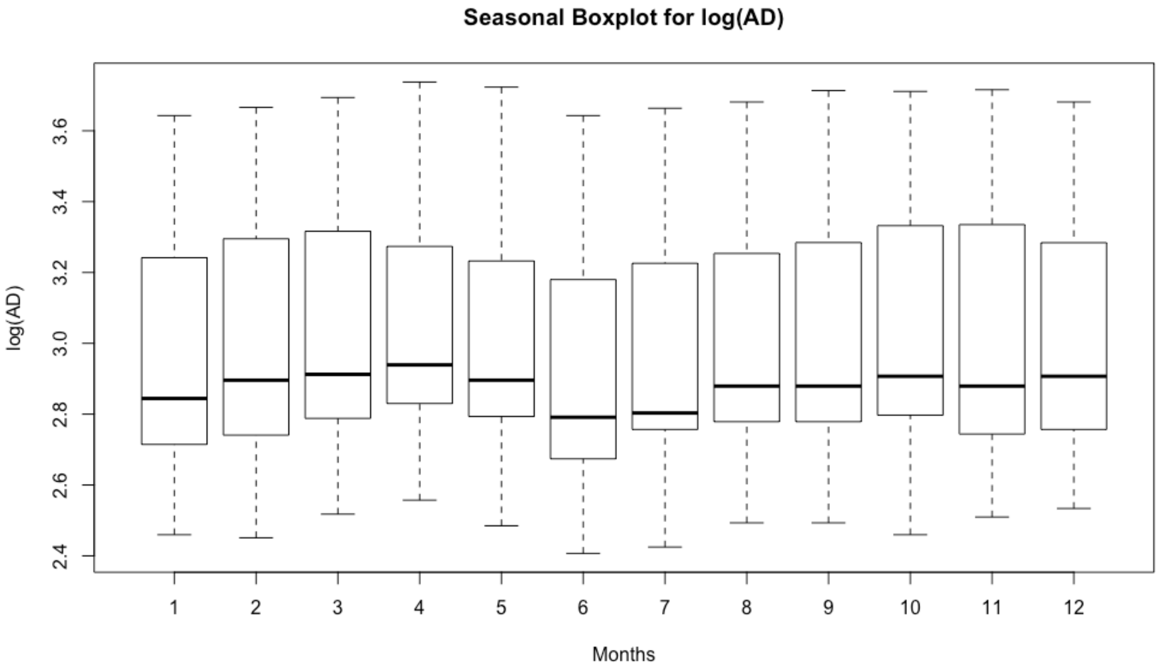
From the Plot 1, the graphs of AD, N27 and CUR show that there is slight decrease in first ten years. Then there is fluctuation. Around 2008, a steep increase takes place. After two years, a decreasing trend takes place. However, for graph of CLFPR, at the beginning, it is fluctuating. But after 2000,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



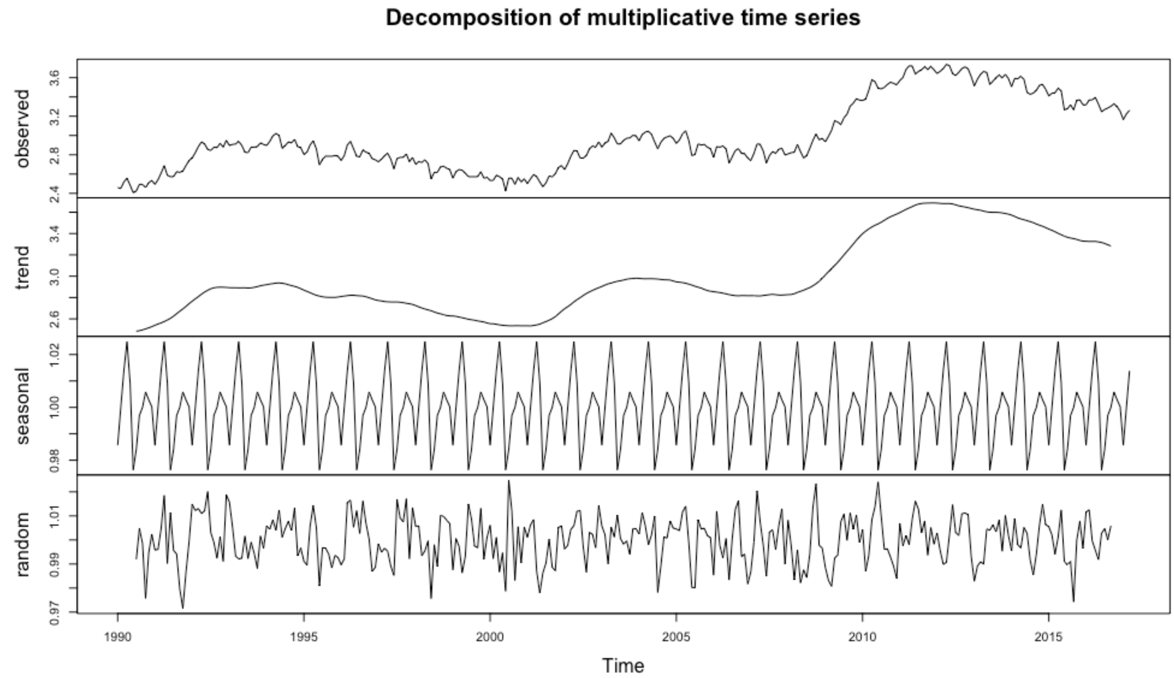
Plot 3: Seasonal Plot of log of AD



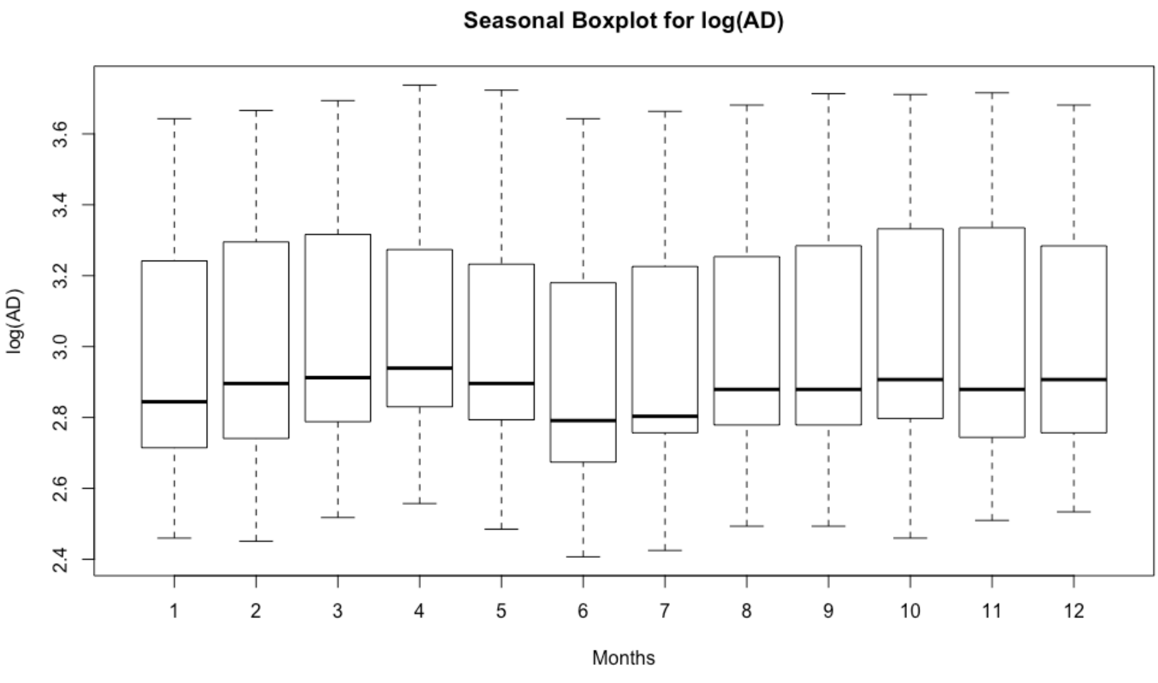
From plot 2, we can see that the trend has slight fluctuation at the beginning. But around 2008, there is a steep increase in the trend then following a downward trend after 2010.There is superimposed seasonal effect exists. But for the random components, we can see that the variance of it is decreasing 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)



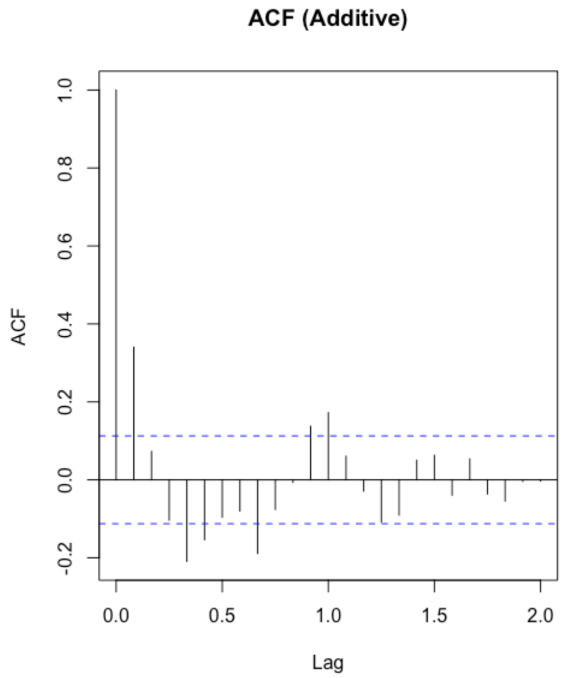
Plot 5: Seasonal Plot of log of AD



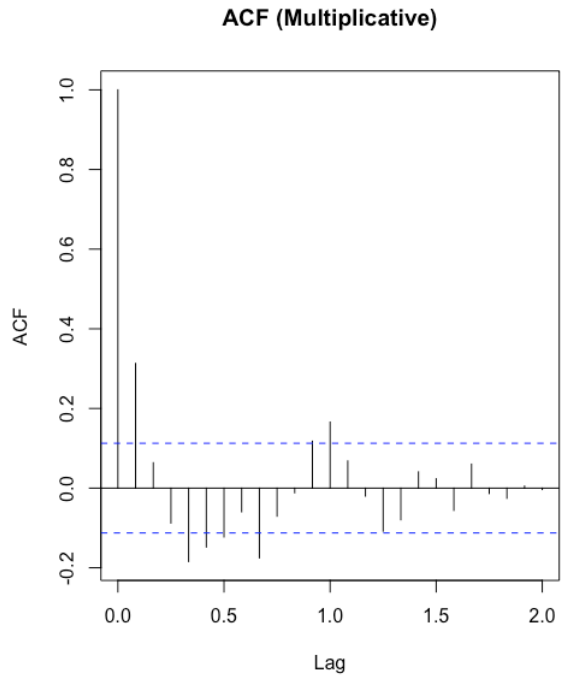
From plot 4, we can see that the trend increases at the beginning then gradually decreases. Around 2001, same pattern takes place. But around 2008, there is a steep increase in the trend then following a downward trend after 2010. There is superimposed seasonal effect exists. For random component, it 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 random terms in Multiplicative Decomposition has a better performance than ACF of random terms in Additive Decomposition. ACF dies down quickly in both plots. 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.

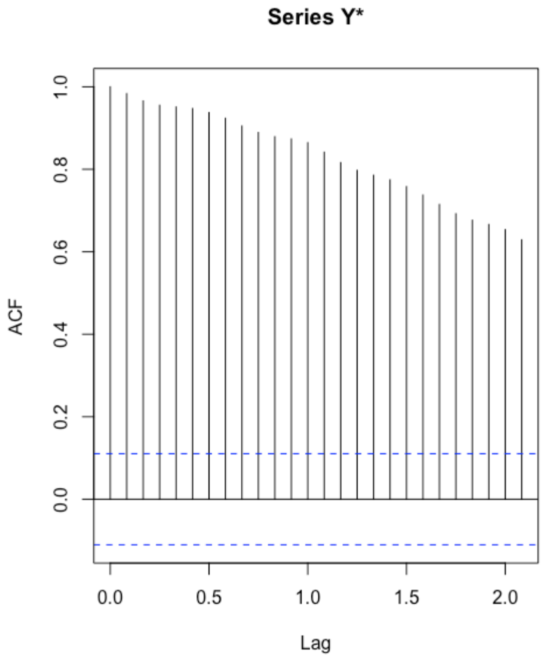
IV. ARIMA Modelling

Table 2: New Notations for All Variables

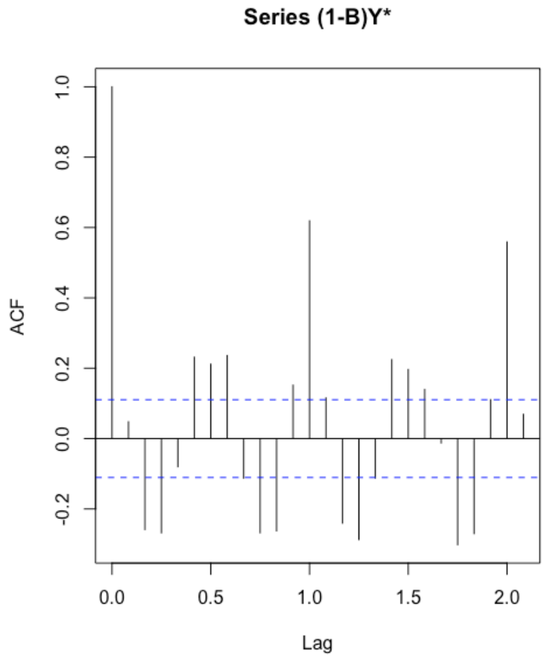
Short Name	Description	Period	Mean	Standard Deviation
Y* (Y_t)	Log of average duration of unemployment in US(in weeks)	1990.1-2016.3 (315 Months)	2.974	0.362
x1	Log of Number of Civilians Unemployed for 27 Weeks and Over in US(in Thousands of persons)	1990.1-2016.3 (315 Months)	7.452	0.663
x2	Log of Civilian Unemployment Rate in US(in Percentage)	1990.1-2016.3 (315 Months)	1.775	0.248
x3	Log of Civilian Labor Force Participation Rate in US(in Percentage)	1990.1-2016.3 (315 Months)	4.186	0.022

First, I subtract the latest 12 observations from the data to be the out-of-sample data to test the validity of my prediction. Hence, the in-sample data will contain 315 observations. 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 of the average duration of unemployment in US between Jan 1990 and March 2016. 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*



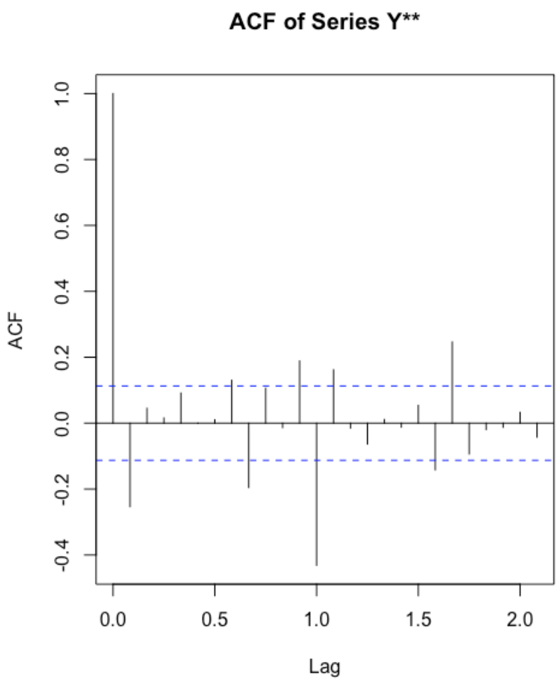
Plot 9: ACF of Series (1-B)Y*



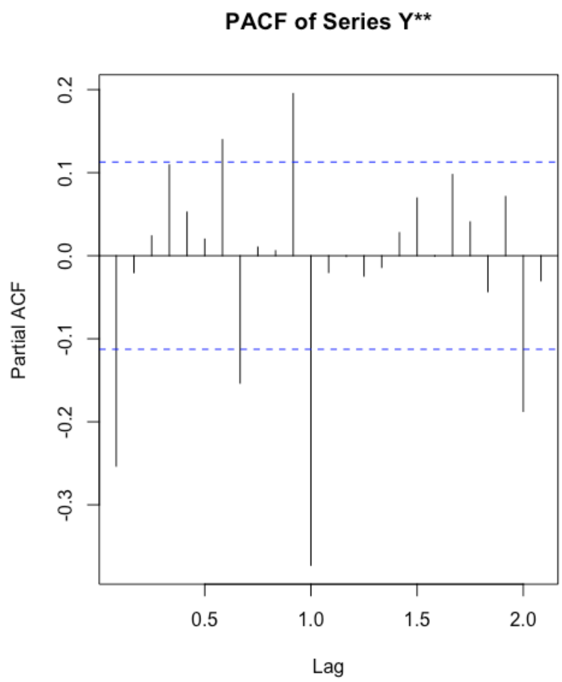
From Plot 8, we can see that there is a visible trend in the ACF. Hence, I decide to regular difference the data. And from plot 9, a significant seasonal lag happens at lag = 12. Hence, I decide to seasonally difference the series (1-B)Y*. After pre-transformation and differencing, I will begin my future work with series Y** :

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

Plot 10: ACF of Series Y**



Plot 11: PACF of Series 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. 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, I tried ARIMA(1,1,1)(1,1,1)$_{12}$, ARIMA(1,1,0)(1,1,1)$_{12}$, and ARIMA(0,1,1)(1,1,1)$_{12}$.

Plot 12: ACF of residuals of three different ARIMA Models

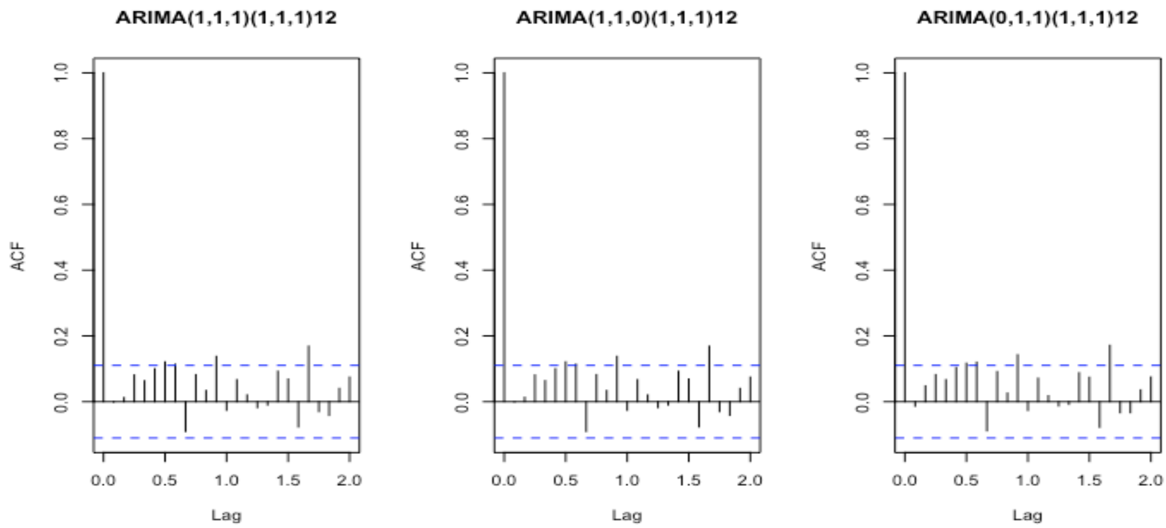
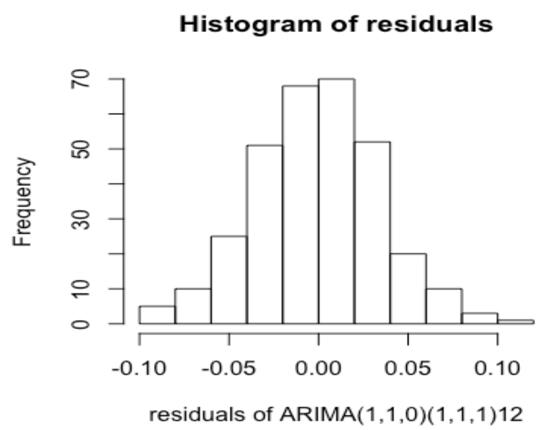


Table 3: Table for AIC value of different ARIMA Model

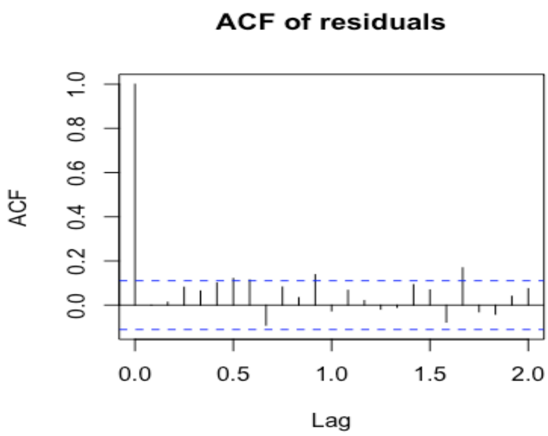
Model	AIC Value
ARIMA(1,1,1)(1,1,1)$_{12}$	-1122.919
ARIMA(1,1,0)(1,1,1)$_{12}$	-1124.919
ARIMA(0,1,1)(1,1,1)$_{12}$	-1124.175

According to Plot 12, ACF of residuals of three different ARIMA model have similar performance. And for all of them, the residuals are stationary. Hence, I decided to look at AIC values to choose the best model. In table 3, 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.

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



Plot 14: 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.

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

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

Expanding all the terms, I get the following equation:

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

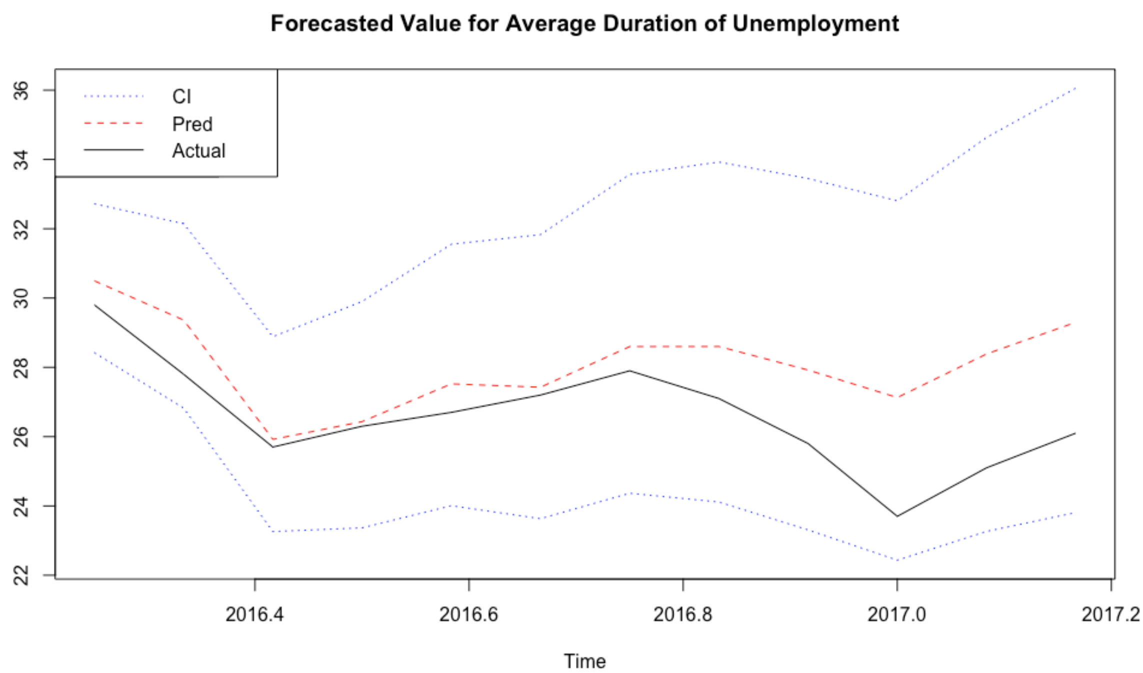
From the R output, I get a_1=-0.193, a_{12}=0.253 and b_{12}=-0.949. Hence,

$$\begin{aligned}\mathbf{Y_t^*} &\mathbf{= 0.807Y_{t-1}^* + 0.193Y_{t-2}^* + 1.253\, Y_{t-12}^* + 1.011\, Y_{t-13}^* + 0.06\, Y_{t-14}^*} \\ &\mathbf{- 0.253\, Y_{t-24}^* + 0.204\, Y_{t-25}^* + 0.049Y_{t-26}^* + w_t - 0.949w_{t-12}}\end{aligned}$$

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

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

Plot 15: Forecasted Value for Average Duration of Unemployment



In plot 15, 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.

Table 4: ARIMA Table

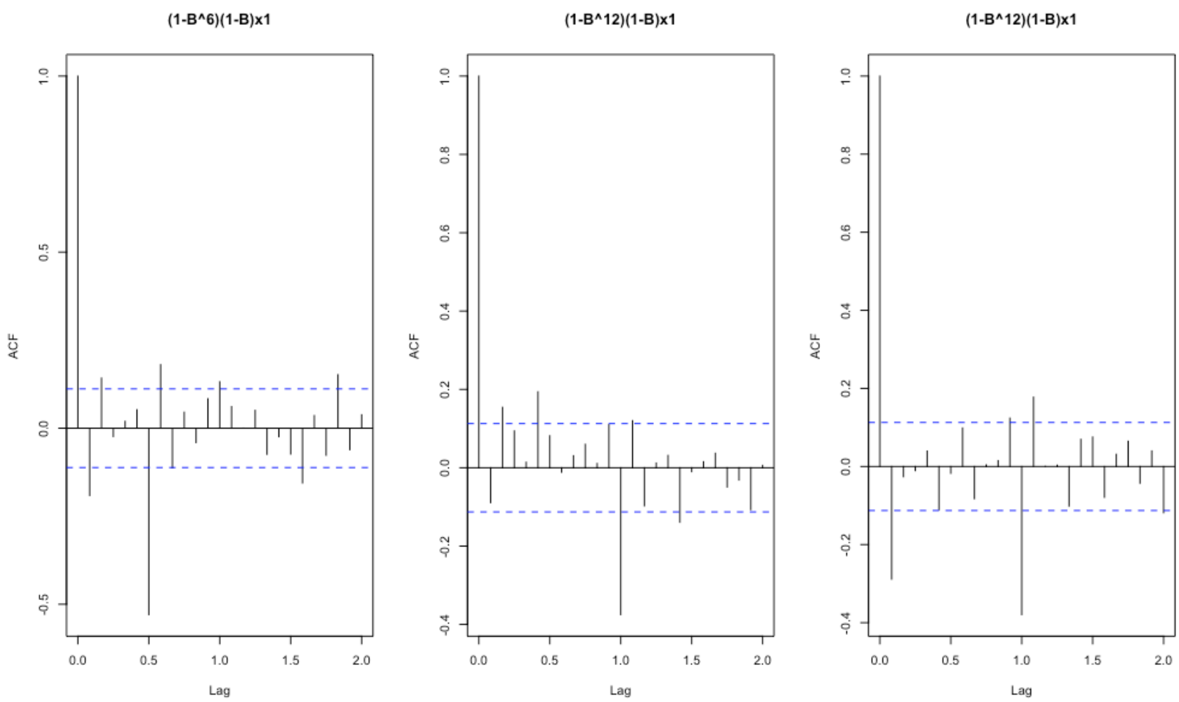
Date	Out-of-sample Actual Values	Forecasted Values
2016.4	29.8	30.5
2016.5	27.8	29.4
2016.6	25.7	25.9
2016.7	26.3	26.4
2016.8	26.7	27.5
2016.9	27.2	27.4
2016.10	27.9	28.6
2016.11	27.1	28.6
2016.12	25.8	27.9
2017.1	23.7	27.1
2017.2	25.1	28.4
2017.3	26.1	29.3
RMSE	1.9097	

V. Vector Autoregression

V.1 Cross-Correlations

To make all the time series of independent variables stationary, I did both regular differencing and seasonal differencing. For variable x1, I regular differenced by differences = 1 and seasonal differenced by lag = 6. For variables x2 and x3, I both regular differenced by differences =1 and seasonal differenced by lag = 12.

Plot 16: ACF of time series of three independent variables

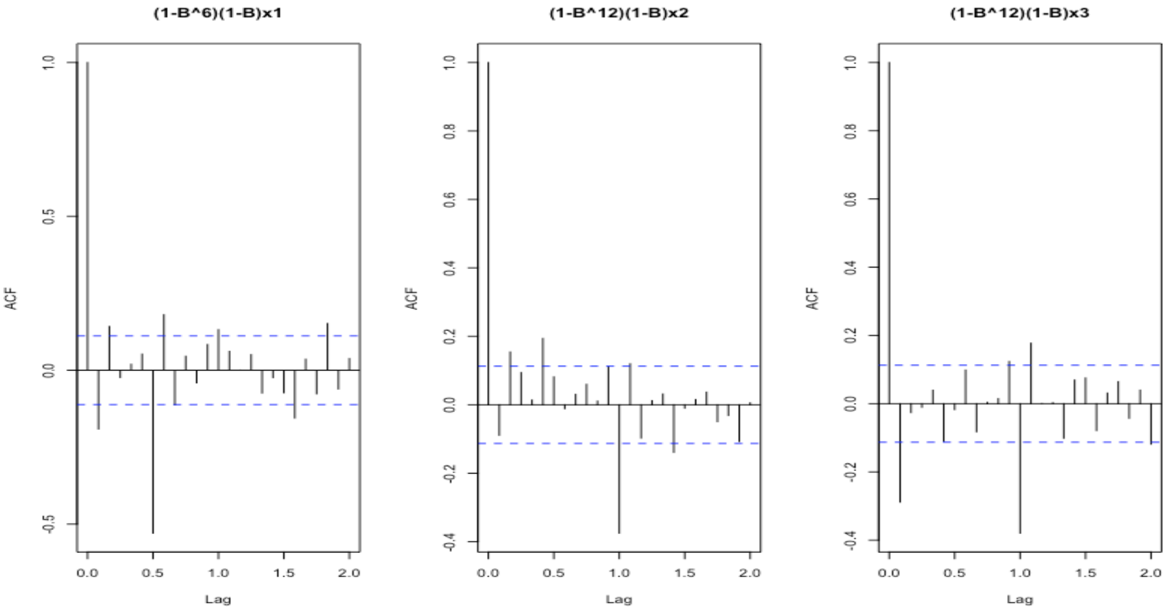


According to plot 16, we can see that all ACF dies down quickly and there are only a few r_k values are significant. Hence, we can say that all three time series are stationary.

Table 5: Polynomial Notations

Original	New	Polynomial Notation
Y*	**Y****	**$(1-B^{12})(1-B)Y^*$**
x1	**x1****	**$(1-B^{6})(1-B)x1$**
x2	**x2****	**$(1-B^{12})(1-B)x2$**
x3	**x3****	**$(1-B^{12})(1-B)x3$**

Plot 17: CCF plot of Target Variables and Independent Variables



According to plot 17, for ccf plot of Y** and x1**, we can see that the significant spike is at positive lags, it means that x1** leads so Y** depends on x1**. For ccf plot of Y** and x2**, we can see that the significant spike is at positive lags, it means that x2** leads so Y** depends on x2**. for ccf plot of Y** and x3**, we can see that the significant spike is at negative lags, it means that Y** leads so x3** depends on Y**.

V.2 Unit root tests and cointegration tests

Hypothesis Testing for Unit Root Test:
H_o: Time series variable is non-stationary and have a unit root
H_a: Stationary

Hypothesis Testing for Cointegration Test:
H_o: Two variables are not cointegrated
H_a: Two variables are cointegrated

Table 6: Unit root tests and cointegration tests results

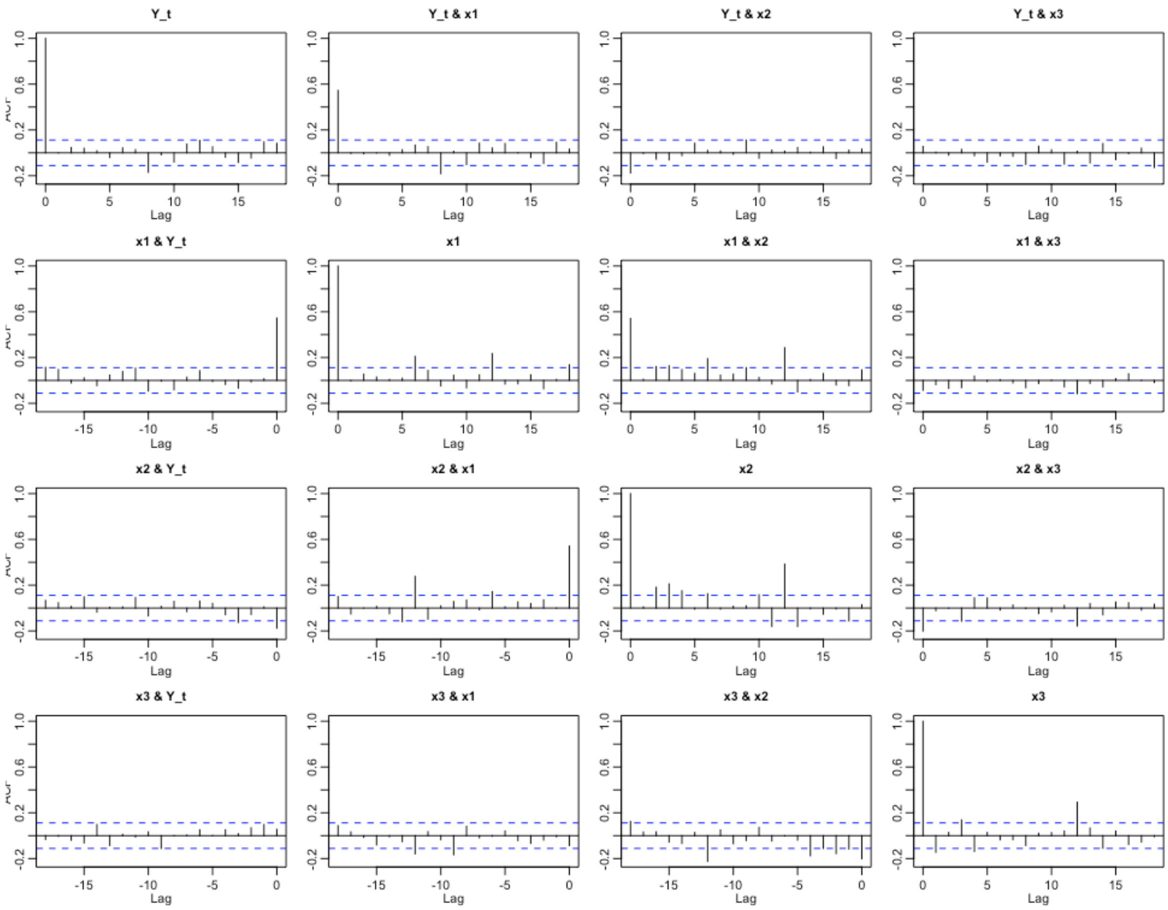
Unit Root Tests		Cointegration Tests	
Variables	P-value	Variables	P-value
Y*	0.8424	Y* & x1	0.1378
x1	0.4868	Y* & x2	0.15
x2	0.8574	Y* & x3	0.01347*
x3	0.8074		

In table 6, in unit root tests, p-value for all variables are not significant, so we fail to reject the null and conclude that all variables are stationary. In cointegration tests, only p-value for Y^* and x3 is significant under 5% significant level. So Y^* is only cointegrated with x3. Hence, by taking both results into account, I will not apply differencing to my data. The logged data will be used in further VAR model fitting.

V.3 Fitting of VAR

By applying ar command, R suggests using order=13 to fit the model. However, such order is so huge that most of the coefficients are insignificant. Hence, by trying different values of order, I find order=6 is sufficient to fit the model.

Plot 18: ACF of residuals in VAR(6)



According to plot 18, we can see that ACF of residuals of Y^* cuts off after lag=1. And for other three variables, there are only a few significant values. Moreover, the cross-ACFs only have very few significant values. Hence, we can say that cross-ACFs and ACFs are white noise.

Final Model:

$$Y^*_t = 3.398^{**} + 0.664Y^*_{t-1}{}^{**} + 0.019x1_{t-1} + 0.216x2_{t-1}{}^{**} - 0.746x3_{t-1} - 0.021Y^*_{t-2} + 0.010x1_{t-2} - 0.010x2_{t-2} + 1.000x3_{t-2} + 0.002Y^*_{t-3} + 0.065x1_{t-3} + 0.033x2_{t-3} - 3.599x3_{t-3}{}^{**} - 0.195Y^*_{t-4} + 0.101x1_{t-4} - 0.226x2_{t-4}{}^{*} + 3.156x3_{t-4}{}^{**} + 0.265Y^*_{t-5}{}^{*} - 0.078x1_{t-5} - 0.217x2_{t-5}{}^{*} + 1.594x3_{t-5} + 0.094Y^*_{t-6} - 0.033x1_{t-6} + 0.218x2_{t-6}{}^{**} - 2.234x3_{t-6}{}^{**}$$

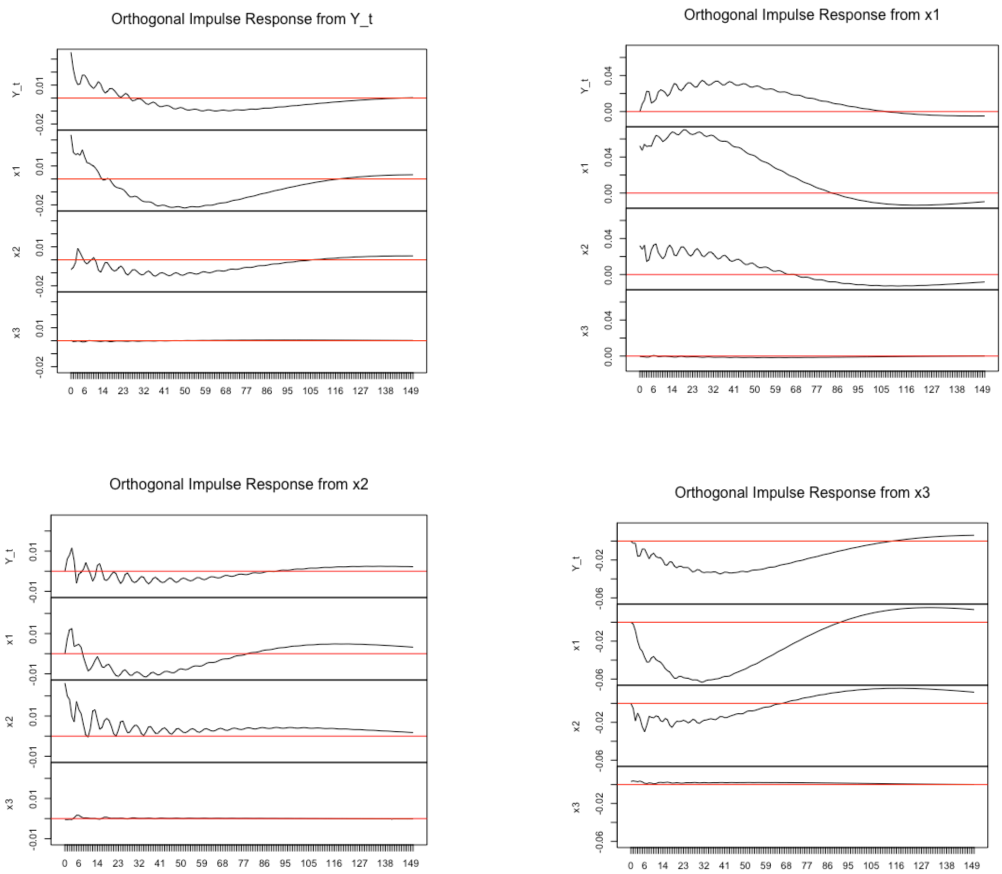
$$x1_t = 8.570^{**} - 0.064Y^*_{t-1} + 0.734x1_{t-1}^{**} + 0.272x2_{t-1}^{*} - 0.550x3_{t-1} - 0.011Y^*_{t-2} + 0.206x1_{t-2} + 0.017x2_{t-2} - 1.363x3_{t-2} + 0.104Y^*_{t-3} - 0.071x1_{t-3} - 0.133x2_{t-3} - 0.350x3_{t-3} - 0.460Y^*_{t-4}{}^{*} + 0.285x1_{t-4}^{*} - 0.336x2_{t-4}^{*} - 0.623x3_{t-4} + 0.272Y^*_{t-5} - 0.082x1_{t-5} + 0.243x2_{t-5} + 0.390x3_{t-5} - 0.180Y^*_{t-6} + 0.035x1_{t-6} - 0.047x2_{t-6} + 0.493x3_{t-6}$$

$$x2_t = 2.194 - 0.074Y^*_{t-1} + 0.072x1_{t-1} + 0.731x2_{t-1}^{**} - 1.594x3_{t-1}^{*} - 0.031Y^*_{t-2} + 0.102x1_{t-2} + 0.053x2_{t-2} - 2.702x3_{t-2}^{**} + 0.362Y^*_{t-3}{}^{**} - 0.253x1_{t-3}^{**} - 0.166x2_{t-3} + 6.047x3_{t-3}^{**} - 0.258Y^*_{t-4} + 0.155x1_{t-4} + 0.001x2_{t-4} - 4.735x3_{t-4}^{**} + 0.036Y^*_{t-5} - 0.037x1_{t-5} + 0.489x2_{t-5}^{**} - 1.286x3_{t-5} - 0.143Y^*_{t-6} + 0.002x1_{t-6} - 0.146x2_{t-6} + 3.770x3_{t-6}^{**}$$

$$x3_t = 0.225 - 0.020Y^*_{t-1}{}^{*} - 0.003x1_{t-1} + 0.001x2_{t-1} + 1.159x3_{t-1}^{**} + 0.027Y^*_{t-2}{}^{**} - 0.004x1_{t-2} + 0.007x2_{t-2} - 0.321x3_{t-2}^{**} - 0.001Y^*_{t-3} + 0.001x1_{t-3} - 0.008x2_{t-3} + 0.092x3_{t-3} + 0.008Y^*_{t-4} - 0.011x1_{t-4} + 0.039x2_{t-4}^{**} + 0.191x3_{t-4}^{*} - 0.002Y^*_{t-5} + 0.004x1_{t-5} + 0.009x2_{t-5} - 0.442x3_{t-5}^{**} - 0.014Y^*_{t-6} + 0.012x1_{t-6}^{*} - 0.047x2_{t-6}^{**} + 0.269x3_{t-6}^{**}$$

V.4 Impulse Response Functions

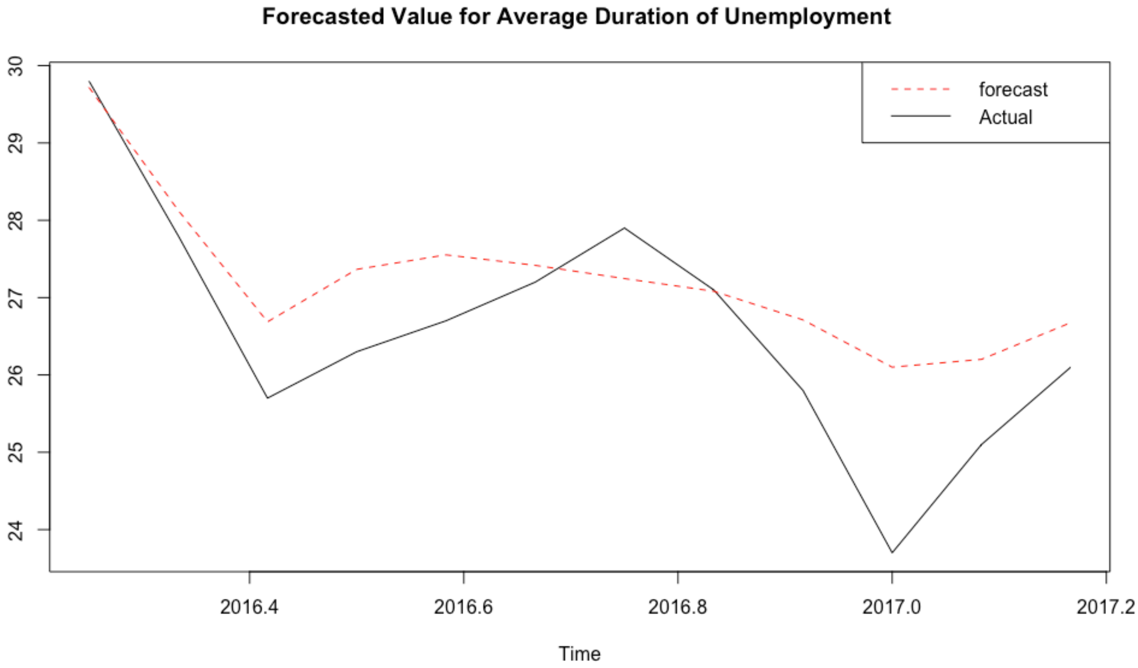
Plot 19: Impulse Response Functions for different variables



In plot 19, we can see the shocks of different variables. When there is a shock of target variable Y*, it takes more than 100 periods for Y*, x1 and x2 to rest. But x3 seems to rest after 20 periods. Moreover, when there is a shock of independent variable, it takes more than 100 periods for Y*, x1 and x2 to rest. But x3 will rest after 20 periods.

V.5 Forecast

Plot 20: Forecasted Value for Average Duration of Unemployment



According to plot 20, we can see that the forecasted value are close to actual value. We can see that VAR(6) is a good model to fit.

Table 7: VAR(6) Table

Date	Out-of-sample Actual Values	Forecasted Values
2016.4	29.8	29.7
2016.5	27.8	28.1
2016.6	25.7	26.7
2016.7	26.3	27.4
2016.8	26.7	27.6
2016.9	27.2	27.4
2016.10	27.9	27.2
2016.11	27.1	27.1
2016.12	25.8	26.7
2017.1	23.7	26.1
2017.2	25.1	26.2
2017.3	26.1	26.7
MSE	0.9817	

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