STAT 153 - Midterm II

Huidi Wang

April 19, 2016

Introduction

In this report, I will take Q5train.csv series dataset as an example to do time series data analysis. This dataset is of length 536 and gives the google trends data (downloaded on 11 April, 2015) for a particular query from the first week of January, 2004 to the week of 06 April– 12 April, 2014, weekly data. My goal is to predict the next 104 observations of the series, which are for next two years.

I will process this dataset from the following steps.

(1) Exploratory data

(2) Data Transformation

(3) Deal with trend and seasonality

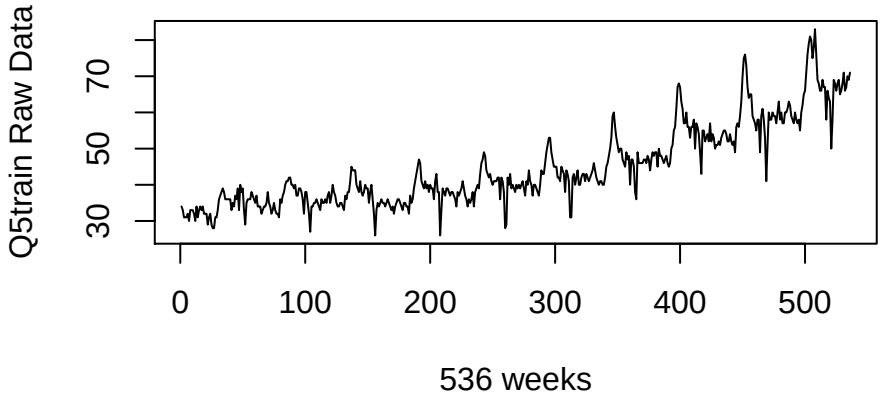
(4) Fit an ARMA model to the residuals after removing trend and sesonality

(5) Model Diagnostics (check if the fitted ARMA is adequate.)

(6) Forecast

(1) Exploratory data

We have total 536 weekly data in Q5train with min=26, max=83, and mean=44.71 The dataset is plotted below. We can clearly see the data existing with trend and seasonality.

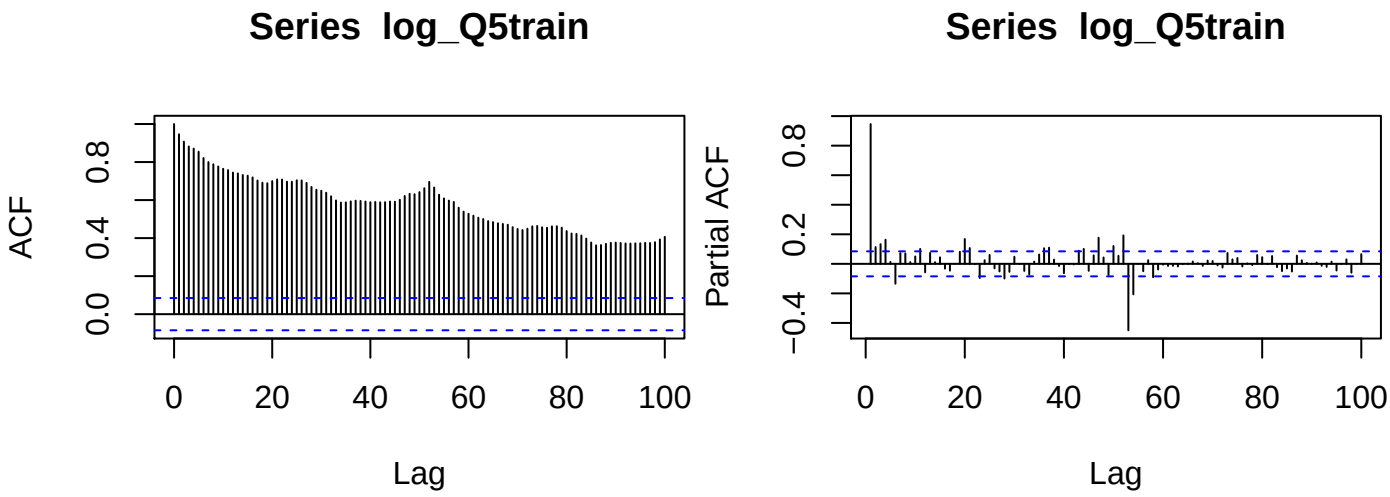



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   26.00   36.00   41.00   44.71   52.00   83.00
```

(2) Data Transformation

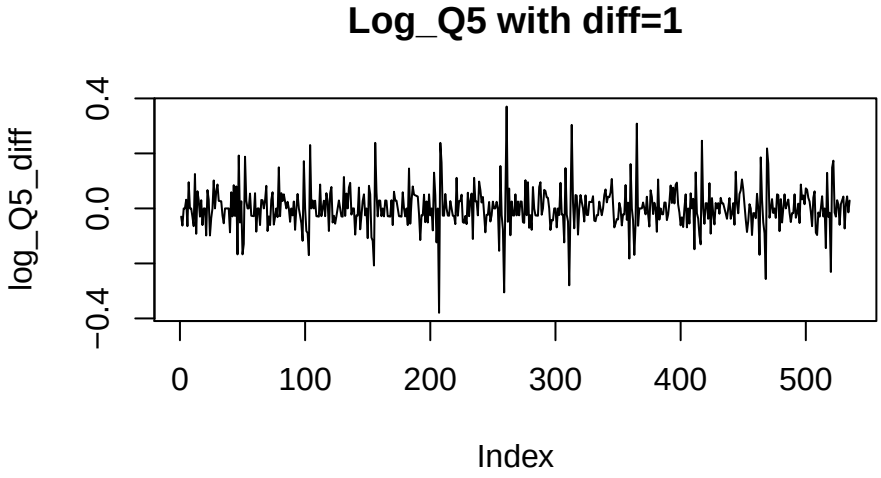
Since log transformation has nicer properties, like log-log relationship represents the proportion of two variables which is more stationary than the absolute differencecs, and by observing data can find out noise is getting bigger as y value goes bigger. So I will take log of the raw data first, and then use the logged data to do the following steps of anlysis.

Here attached the ACF and pACF of Q5train after taking log.



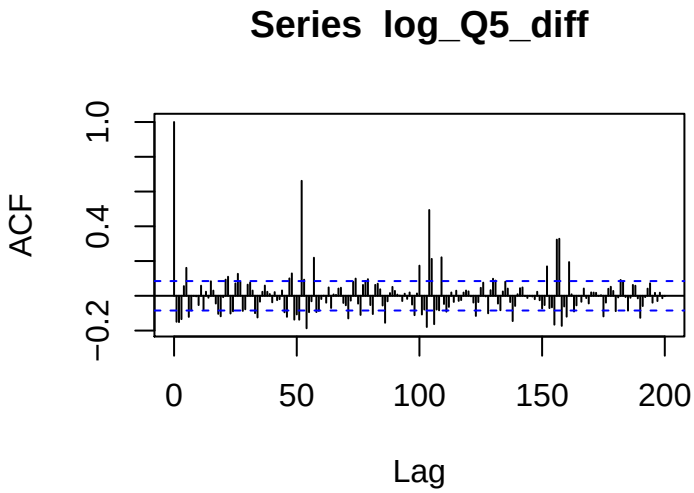
(3) Deal with Trend and Seasonality

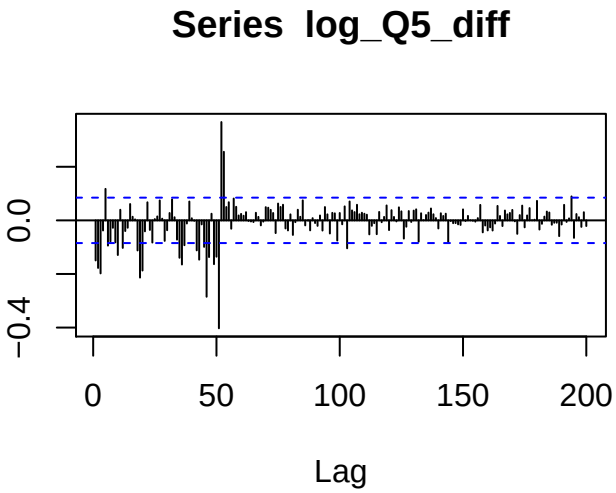
Through the three graphs above, we can clearly see the increasing trend of data and seasonal procress during time series.In this part, I will try to remove this trend and seasonality from log_Q5train in order to get a relative stationary data.



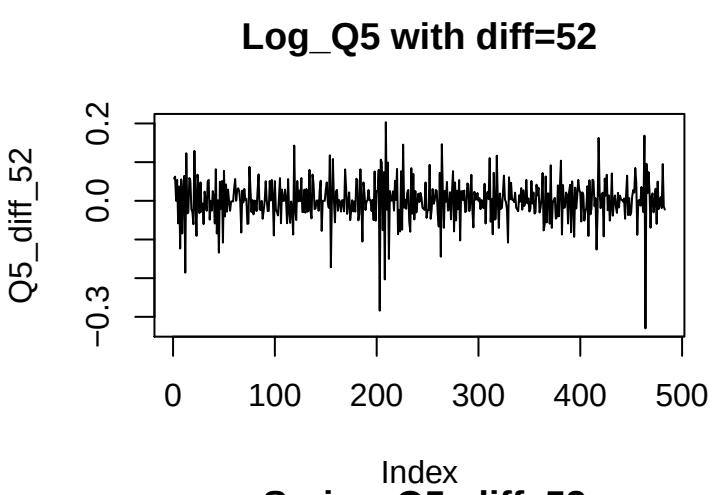
After taking diff=1, we can see the increasing trend has been removed.

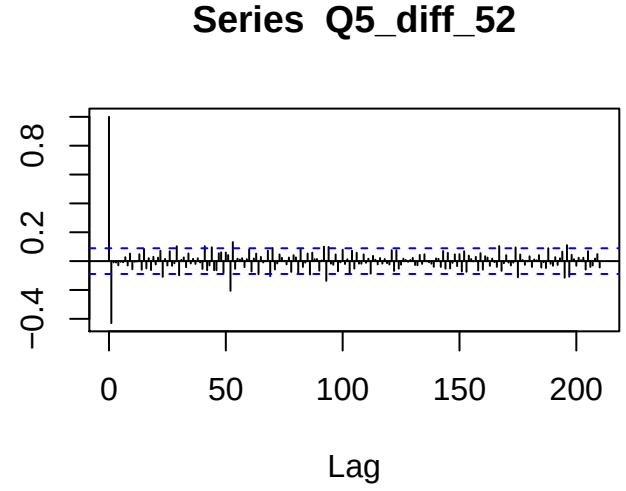
Let's take a look at ACF autocorrelation. It's not hard to find out at lag=52, 104, 156 ACFs are obvious larger than acf at other lags. And it looks like acf repeatedly over every 52 lags and considering this is weekly data. We can reasonaly believe the seasonal diff=52.

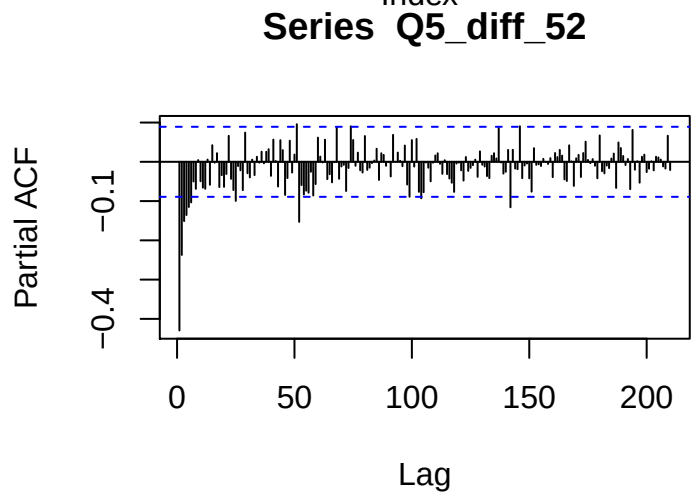




So next step, I will remove seasonality by differecing data log_Q5_diff at diff=52.







(4) Fit an ARMA model to the residuals after removing trend and sesonality

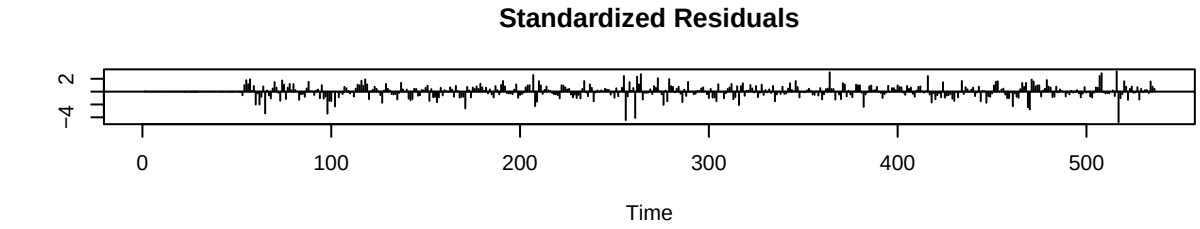
Guess: Looking at the seasonal lags (1s, 2s, 3s, ... with s = 52), the ACF seems to cut off after 1s, and the PACF tails off after 1s, or probably after 2s which is not too obvious, in the seasonal lags, suggesting an MA(1)6 as the seasonal component at first. Looking within the seasonal lags (h = 1, 2,...), the ACF seems to cut off after lag 1 and the PACF seems to tail off, suggesting an MA(1) as the non-seasonal component. So this is fitting an MA(1) × MA(1)6 model.

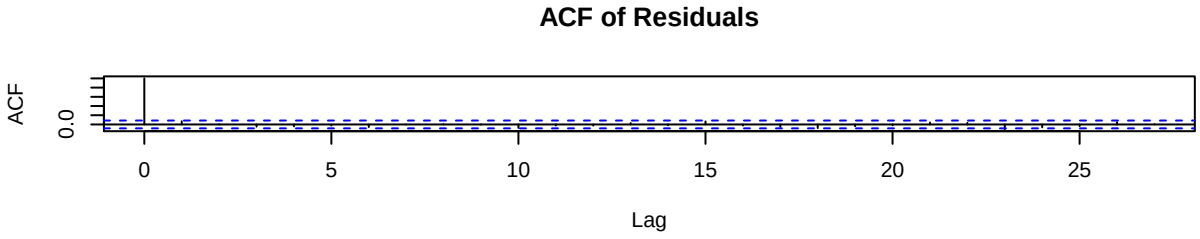
(5) Model Diagnostics (check if the fitted ARMA is adequate.)

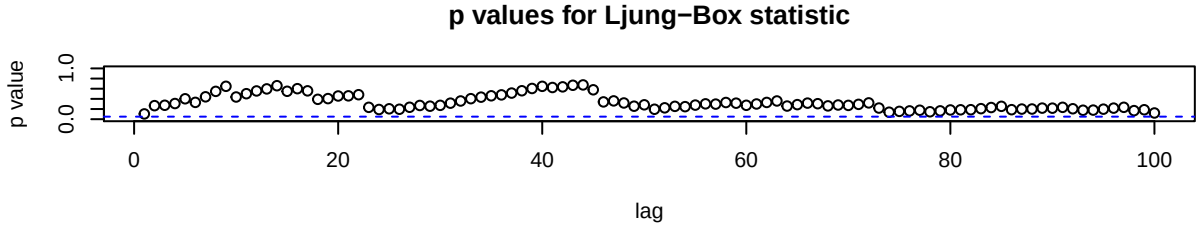
1)Residuals and P-value

Let's show fitted model MA(1) × MA(1)6 through arima function. The result is shown as follow.

```
##
## Call:
## arima(x = log_Q5train, order = c(0, 1, 1), seasonal = list(order = c(0, 1, 1),
##     period = 52))
##
## Coefficients:
##           ma1     sma1
##       -0.7262  -0.2807
## s.e.   0.0380   0.0468
##
## sigma^2 estimated as 0.002021:  log likelihood = 810.46,  aic = -1614.93
```







To observe residuals from this model, I will focus on plots of "standardized residuals" and "ACF of Residuals". As we learned from class, One usually standardizes residuals by dividing by the square-root of the corresponding prediction error. If the model fits well, the standardized residuals should behave as an iid sequence with mean zero and variance one.

Check with the first two plots about residuals, it looks like residuals are around mean 0, but in correlogram, at lag=1,10, 15, 23 acfs are very close to blue bond. Besides, the Ljung-Box test shows when we look at lag=1-100, p-values are all above blue bond, which means residuals' mean equals zero at lag=1 to 100 should not be rejected.Considering still have a lot of p-values are close to blue bond and ACF doesnt look like an obvious white noise, I will overfit this model and hope to get smaller AIC. (Here, AIC=-1614.93 for Q5_0.)

AIC stands for Akaike's Information Criterion. It is a model selection criterion that recommends choosing a model for which: AIC = -2 log(maximum likelihood) + 2k is the smallest.

In this model, ma1 and sma1 are both statistical significant since t=ma1/s.e. are both over than 1.96 shows statistical significant.

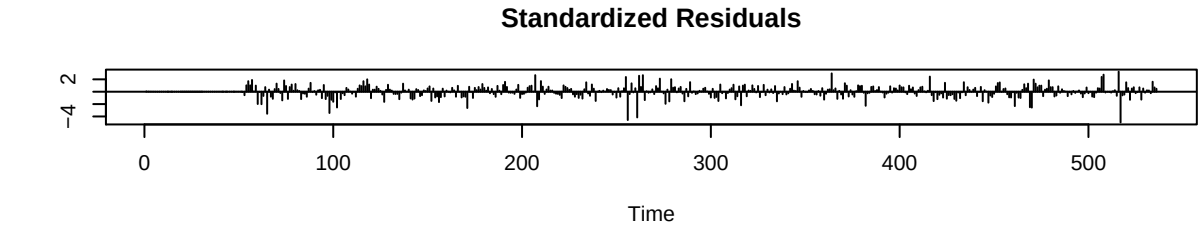
Let's try some combinations based on residuals' performance above.

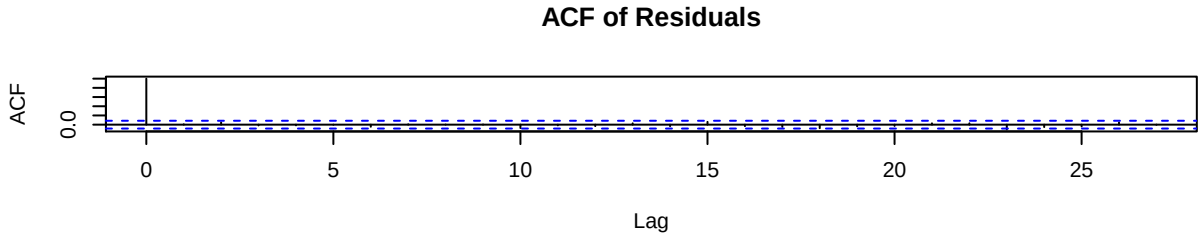
Q5_1=c(0, 1, 1)c(1, 1, 1)52

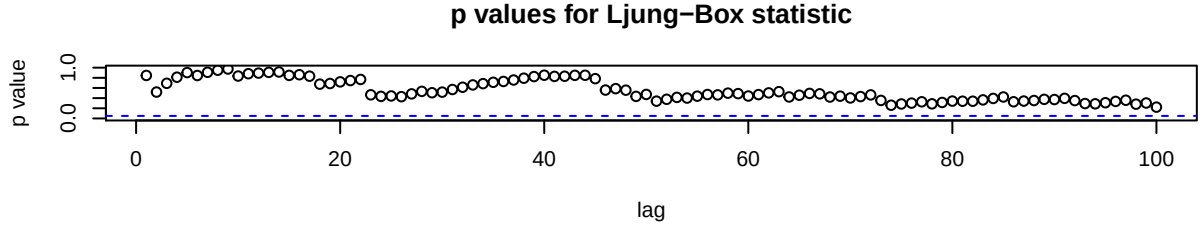
Q5_2=c(0, 1, 1)c(0, 1, 2)52

Q5_3=c(0, 1, 2)c(0, 1, 1)52

```
##
## Call:
## arima(x = log_Q5train, order = c(0, 1, 2), seasonal = list(order = c(0, 1, 1),
##     period = 52))
##
## Coefficients:
##           ma1      ma2     sma1
##       -0.6697  -0.0905  -0.2784
## s.e.   0.0441   0.0451   0.0468
##
## sigma^2 estimated as 0.002004:  log likelihood = 812.47,  aic = -1616.93
```







Please check Appendix R to see all graphs. Briefly statement: (1)Q5_3 model shows relatively white noise of residuals and obviously higher p-value in average comparing to other Q5-0, Q5-1, and Q5-2. (2) When

we check the new-added parameter ma2=-0.0905 and s.e.=0.0451. t=ma2/s.e.=2>1.96 so we could reject ma2=0 at 95% Confidence level. This model gives a stronger predict. AIC=-1616.93< -1614.93 from Q5-0

Back to Q5_1 and Q5_2, they have similar residuals in plots as Q5_0. But for each Q5_1 and Q5_2, their new-added parameters are actually not statistical significant. We can prove it still by checking their t-values.

So far, we believe MA(2)_MA(1)52 is more fittable to our dataset. To further accurate our prediction, I prefer to overfit again based on MA(2)MA(1)52 model. Let's try more combination.

Q5_4=c(0, 1, 2)c(1, 1, 1)52

Q5_5=c(0, 1, 2)c(0, 1, 2)52

See diagrams in R Appendix. If we take a look on these two models, their corresponding graphs are very similar to MA(2)MA(1)52. All p-values are obviously higher than blue bondary, and acf looks whit noise. Resideuals' mean approches to zero. When we check every new coeffients in details, we'll find out MA(2)MA(1)52 is still the one that has most statistical significant at 5% significant level on the model's coefficients , like ma1, ma2, and sma1. As a result, I would like to fit MA(2)MA(1)52 as the dataset's ARMA model. (If we consider AIC at this moment, this model also has the smallest AIC which is also good.)

Use ARMAacf to chechk our result matching with raw data or not. %%%%%%%%%%%

2)AIC and BIC

Compare all the models' AICs and BICs from above. ALways choose the one with smallest AIC or smallest BIC.

```
#AIC(Q5_0)  aic = -1614.93                              BIC(Q5_0)  -1602.389 #chose by smallesdt BIC -16
#AIC(Q5_1)  aic = -1614.1                               BIC(Q5_1)  -1597.38
#AIC(Q5_2)  aic = -1614.07                              BIC(Q5_2)  -1593.166
#AIC(Q5_3)  aic = -1616.93 chose for smallest AIC       BIC(Q5_3)  -1600.211
#AIC(Q5_4)  aic = -1616.4                               BIC(Q5_4)  -1595.496
#AIC(Q5_5)  aic = -1617.09 chose for smallest AIC       BIC(Q5_5)  -1596.188
#Note:AIC for Q5_3 and Q5_5 are pretty much same small.
```

3)Cross Validation

Briefly introduce how to calculate CV

a. Fix a model M5-0=c(0, 1, 1)*c(1,1,0)_52 as my prediction before. Fix k from 5 to 9 because we have total 10 years and sixteen weeks with prediction for the next two years. 5<10

b. Fit the model M5-0 to the data from the first k years plus sixteen weeks, then use the fitted model to predict the data for the next 52 weeks.

c. Calculate the sum of squares of errors of prediction for the 52 weeks.

d. Repeat these steps for k=5,6,7,8,9

e. Average the sum of squares of errors of prediction over k = 5,6,7,8,9 Denote this value by CV5-1=mean(MSE)

f. Calculate CV for each estimated model, and choose the model with the smallest Cross-Validation score.

Overfitting Models by CV

As shown in R Appendix, I use computeCVmse function to calculate MSE. Start from my prediction as before MA(1) × AR(1)52.

MSE5_0 = computeCVmse(c(0, 1, 1), c(0,1,1)) #CV0=6.501581

This looks not bad.To compare it with other simple ARMA:

MSE5_1 = computeCVmse(c(1, 1, 0), c(0,1,1)) #7.74596

MSE5_2 = computeCVmse(c(1, 1, 0), c(1,1,0)) #7.613845

MSE5_3 = computeCVmse(c(0, 1, 1), c(1,1,0)) #6.516456

CV0 is the smallest, so we start from MSE5_0 to overfit models.

MSE5_4 = computeCVmse(c(0, 1, 1), c(1,1,1)) #6.50228

MSE5_5 = computeCVmse(c(0, 1, 1), c(0,1,2)) #6.418859

MSE5_6 = computeCVmse(c(0, 1, 2), c(0,1,1)) #6.507531

MSE5_7 = computeCVmse(c(0, 1, 2), c(0,1,3)) #6.284072

MSE5_8 = computeCVmse(c(0, 1, 1), c(0,1,3)) #6.277156

MSE5_9 = computeCVmse(c(1, 1, 1), c(0,1,3)) #6.289948

MSE5_10 = computeCVmse(c(0, 1, 1), c(2,1,1)) #6.490148

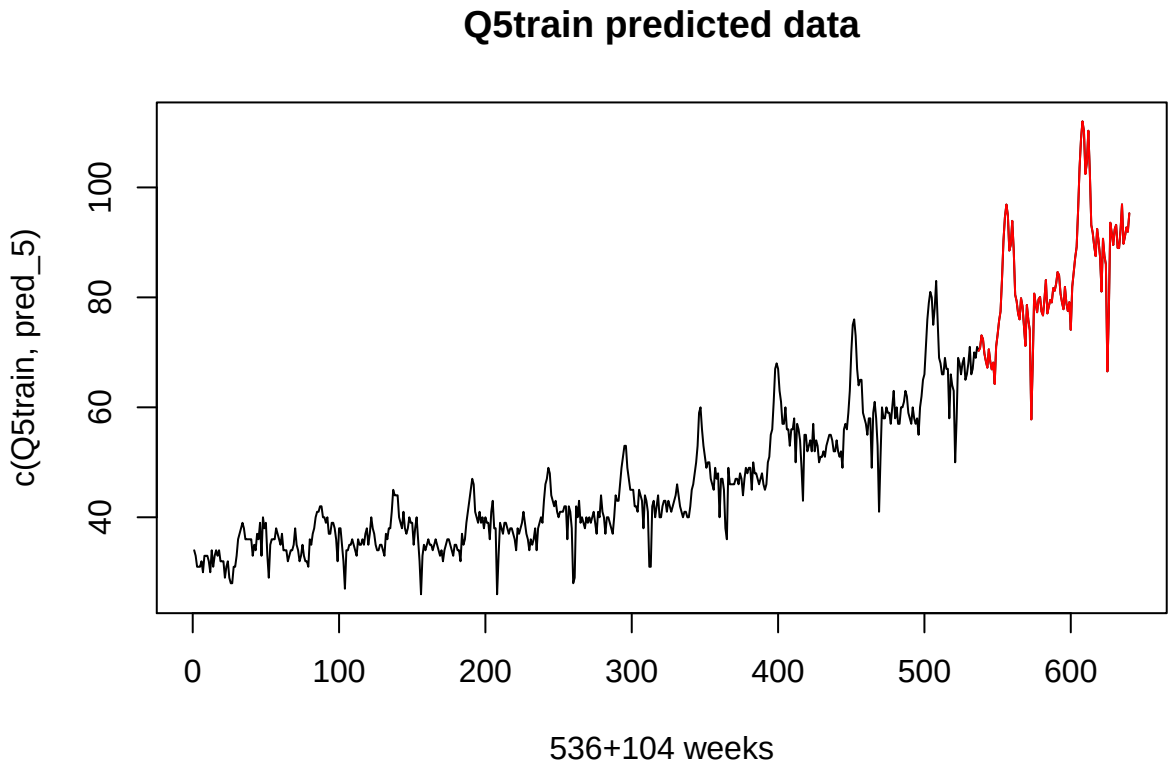
So far, by playing around with various combinations, we can find out MSE5_7 and MSE5_8 have smallest CV. Please see all CV calculations in R Appendix. Then I'm curious if their residuals make sense too. Q5_6<- arima(log_Q5train, order = c(0, 1, 2), seasonal = list(order = c(0, 1, 3), period = 52))

Q5_7<- arima(log_Q5train, order = c(0, 1, 1), seasonal = list(order = c(0, 1, 3), period = 52))

After checking with armia function, we can see (from R Appendix) in general c(0, 1, 2)c(0, 1, 3)52 has relatively higher p-values and smaller AIC than c(0, 1, 1)c(0, 1, 3)52. So finally, I will choose c(0, 1, 2)*c(0, 1, 3)52 as fitted model under Cross Vlidation method.

(6) Forecast and Conclusion

In above, we have two fitted moedls:c(0, 1, 2)c(0, 1, 3)52 from CV method, c(0, 1, 2)c(0, 1, 1)52 from Residuals, p-value and AIC method, and c(0,1,1)c(0,1,1)52 from BIC method. Combining all their performance under each method, finally I will choose c(0, 1, 2)c(0, 1, 3)52 to predict the next 104 values. I mainly focus on Cross-Validation since this one works better on prediction. Once get the fitted model, I use predict function in R to predict 104 ahead numbers. (See R Appendix) I'll plot predicted numbers with raw data as below to see if they look like correctly.


Q5train predicted data
c(Q5train, pred_5)
40
60
80
100
0
100
200
300
400
500
600
536+104 weeks