STAT 672: Covid-19 Time Series Analysis of Bangladesh

Sharmin Hossain- Student ID: 1337949

16 Dec 2022

Introduction

Coronavirus disease 2019 (COVID-19) is a contagious disease caused by a virus, the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2).On 8 March, Bangladesh confirmed 3 laboratories tested coronavirus cases for the very first time. Bangladesh is one of the most densely populated globally. For this reason, the transmission rate of COVID-19 was increasing day by day.To reduce the transmission rate in Bangladesh, the government declared a lockdown throughout the nation from March 23, 2020, to various lengths. The vaccination activity started from 27 January, 2021. Till this day, more than 2 million people were recorded to be infected by this virus and among the whole population, 75% are recorded to be vaccinated so far.
As i wanted to work on something related to my country, i choose this topic for the availability of dataset and relevancy of present world we live in.

Objective

The main objective of this project to predict the daily new infected cases of Bangladesh. Additionally, i wanted to know if there is any change in the number of new infected cases after the vaccination started.

Dataset Collection & Preparation

The dataset was collected from "World Health Organization"'s official website.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	iso_code	continent	location	date	total_case	new_case	total_dea	new_deat	total_case	new_cases_per_million	Vaccination	total_dea	new_deat tc
2	BGD	Asia	Banglades	08-03-20	3	3			0.018	0.018	0		
3	BGD	Asia	Banglades	09-03-20	3	0			0.018	0	0		
4	BGD	Asia	Banglades	10-03-20	3	0			0.018	0	0		
5	BGD	Asia	Banglades	11-03-20	3	0			0.018	0	0		
6	BGD	Asia	Banglades	12-03-20	3	0			0.018	0	0		
7	BGD	Asia	Banglades	13-03-20	3	0			0.018	0	0		
8	BGD	Asia	Banglades	14-03-20	3	0			0.018	0	0		
9	BGD	Asia	Banglades	15-03-20	5	2			0.029	0.012	0		
10	BGD	Asia	Banglades	16-03-20	8	3			0.047	0.018	0		

The csv file consisted of different variables such as new cases, new vaccination, new deaths etc. For the purpose of this project, i used the new_cases_per_million column and date column.

Initially, I had to add another column called "Vaccination" because i wanted to understand the effect of vaccination on the new infected cases. I made the qualitative column in excel based on the "new vaccinations" column. The first day, i had a number there, i put "1" for that row and afterwards. Everything else was put as "0" for "vaccination" column.

After this, i had checked "date" column in my dataset which was shown as "character". That's why i had to convert it to "date" so that i can work on the time series.

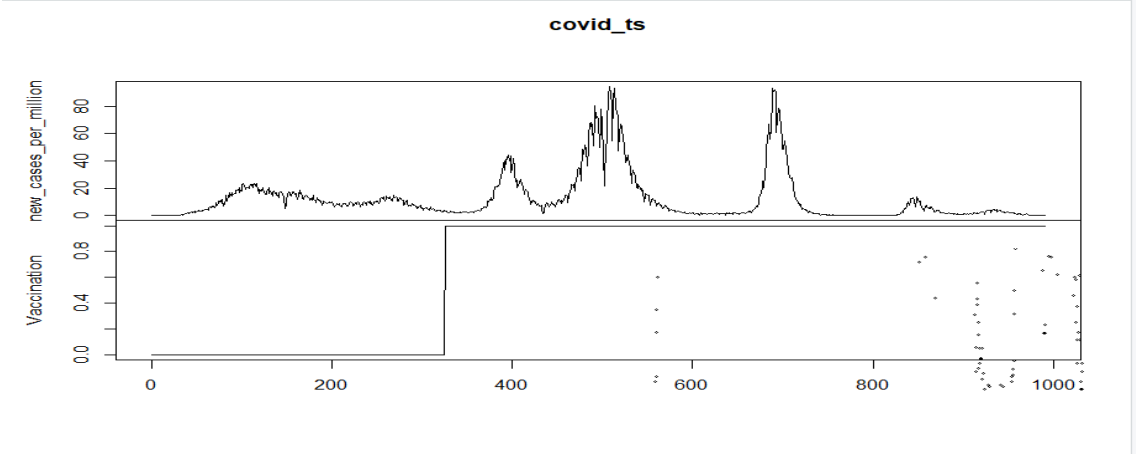
Also, i have divided the dataset into two groups named as training and testing data set to make prediction and compare it afterwards. I took the last 7 days data as the "testing" set.

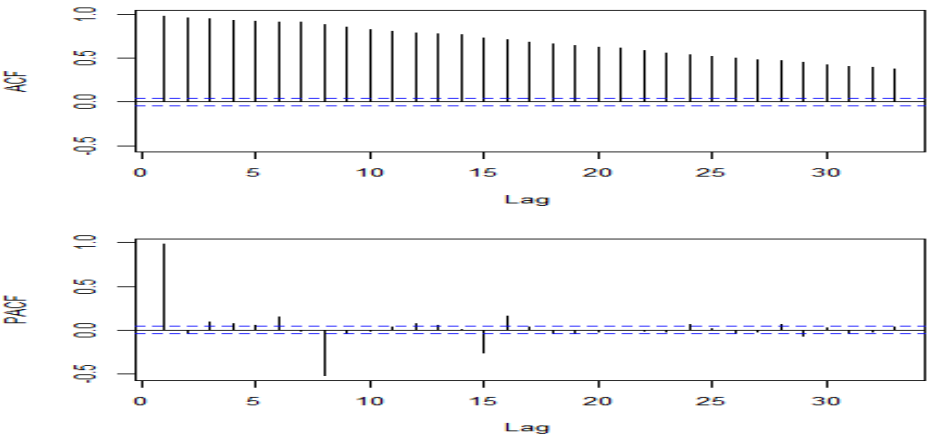
```
> covid_test <- tail(covid[,c("new_cases_per_million","Vaccination","date")], 7)   #get 7 last days
> covid_test
    new_cases_per_million Vaccination       date
992                 0.117           1 2022-11-24
993                 0.111           1 2022-11-25
994                 0.134           1 2022-11-26
995                 0.093           1 2022-11-27
996                 0.169           1 2022-11-28
997                 0.064           1 2022-11-29
998                 0.105           1 2022-11-30
> covid_train <- head(covid[,c("new_cases_per_million","Vaccination","date")], nrow(covid) - nrow(covid_test)) #get th
e rest data
> tail(covid_train)
    new_cases_per_million Vaccination       date
986                 0.111           1 2022-11-18
987                 0.105           1 2022-11-19
988                 0.140           1 2022-11-20
989                 0.152           1 2022-11-21
990                 0.134           1 2022-11-22
991                 0.193           1 2022-11-23
```

Stationarity Checking

At first, I had made a time series object with my training data taking only two columns ("new_cases_per_millions","vaccination"). Then i have plotted it to observe the general ACF, PACF and raw data.

From ACF, it can be seen the data is tailing off pretty slowly whereas from PACF, the data seems to be cut of after 17 lags.





After that, i wanted to check the stationarity of the time series.
Thats why i took the augmented Dickey–Fuller test individually for those two variables. Where vaccination showed the sign of non-stationarity.

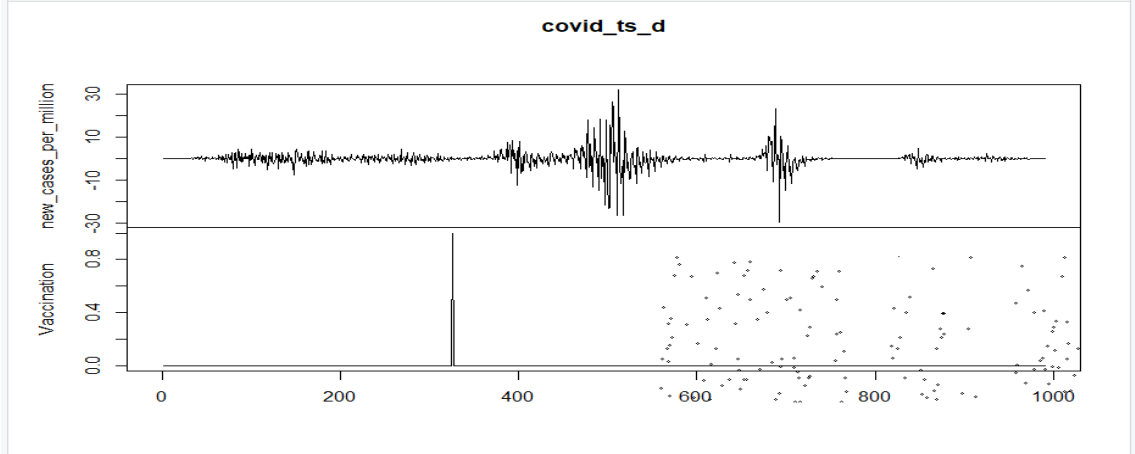
```
> adf.test(covid_ts[,c("new_cases_per_million")])

        Augmented Dickey-Fuller Test

data:  covid_ts[, c("new_cases_per_million")]
Dickey-Fuller = -4.3979, Lag order = 9, p-value = 0.01
alternative hypothesis: stationary

Warning message:
In adf.test(covid_ts[, c("new_cases_per_million")]) :
  p-value smaller than printed p-value
> adf.test(covid_ts[,c("Vaccination")])

        Augmented Dickey-Fuller Test

data:  covid_ts[, c("Vaccination")]
Dickey-Fuller = -1.6168, Lag order = 9, p-value = 0.7405
alternative hypothesis: stationary
```

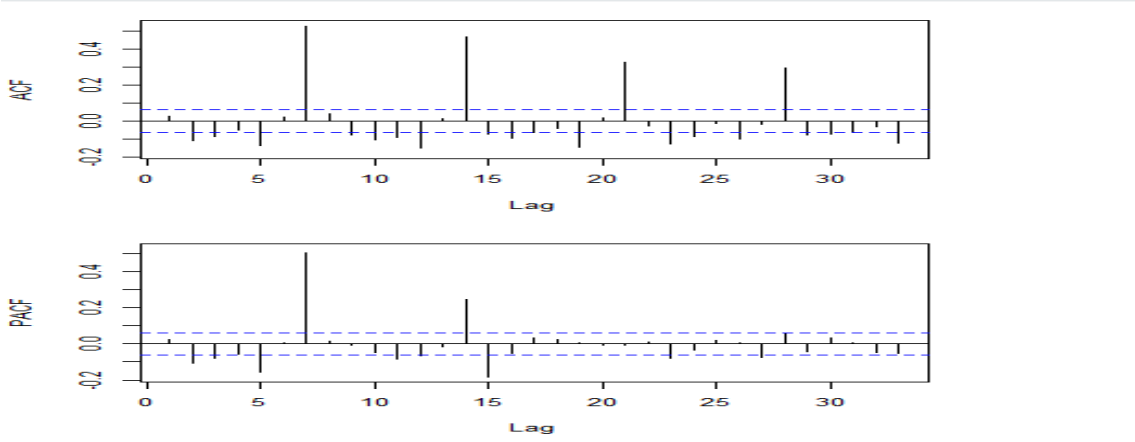
But as we know adf test does not work with multivariate time series properly.

I had to do the Johansen Test for co-integration to check if there is a co-integration between those two variables. But i could not reject the null hypothesis (no cointegration) for any number of rank (rank 0 and rank<= 1) for 95% confidence interval.

```
> jotest<-ca.jo(covid_ts[,c("new_cases_per_million","Vaccination")], type="trace", ecdet="none", spec="longrun")
> summary(jotest)

######################
# Johansen-Procedure #
######################

Test type: trace statistic , with linear trend

Eigenvalues (lambda):
[1] 0.012663316 0.002031447

Values of teststatistic and critical values of test:

          test 10pct  5pct  1pct
r <= 1 |  2.01  6.50  8.18 11.65
r = 0  | 14.62 15.66 17.95 23.52

Eigenvectors, normalised to first column:
(These are the cointegration relations)

                      new_cases_per_million.l2 Vaccination.l2
new_cases_per_million.l2                  1.00000        1.0000
Vaccination.l2                           -1.13627     -548.0437

Weights W:
(This is the loading matrix)

                     new_cases_per_million.l2 Vaccination.l2
new_cases_per_million.d           -2.426159e-02  -8.620235e-05
Vaccination.d                     -2.707162e-05   5.546531e-06
```

Then, i had to do one difference of the main time series.
Then i again have plotted the difference data, ACF and PACF for that. Here, ACF and PACF both seemed to be cut off after a certain lag.





After performing Johansen test again, i was able to reject the null hypothesis and therefore conclude that the time series is stationary.

```
> summary(jotest_d)

######################
# Johansen-Procedure #
######################

Test type: trace statistic , with linear trend

Eigenvalues (lambda):
[1] 0.3768534 0.3340116

Values of teststatistic and critical values of test:

          test 10pct  5pct  1pct
r <= 1 | 401.61  6.50  8.18 11.65
r = 0  | 868.90 15.66 17.95 23.52

Eigenvectors, normalised to first column:
(These are the cointegration relations)

                        new_cases_per_million.l2 Vaccination.l2
new_cases_per_million.l2                1.0000000          1.00
Vaccination.l2                         -0.6576407      16123.22

Weights W:
(This is the loading matrix)

                       new_cases_per_million.l2 Vaccination.l2
new_cases_per_million.d            -1.079257e+00  -7.372923e-05
Vaccination.d                       7.915506e-05  -6.214355e-05
```

ARIMA Procedure

ARIMA, known as 'Auto Regressive Integrated Moving Average' is a class of models that 'explains' a given time series based on its own past values, that is, its own lags and the lagged forecast errors, so that equation can be used to forecast future values.

For my project, i started with ARIMA procedure as several studies of covid data from many countries have shown promising results with ARIMA models. That's why i started using it in the context of Bangladesh too.

At first, i have done a auto ARIMA on "new_cases_per_millions" using "Vaccination" as a xregressor. From there, it showed me the best ARIMA for my model would be ARIMA(5,0,5) It also showed model estimator for my external regressor variable (vaccination estimator=0.0679)

```
> covid_arima_auto <- auto.arima(y = covid_ts_d$new_cases_per_million,ic="aicc",trace=T,xreg = covid_ts_d$vaccination

 Fitting models using approximations to speed things up...

 ARIMA(2,0,2) with non-zero mean : 5330.613
 ARIMA(0,0,0) with non-zero mean : 5438.233
 ARIMA(1,0,0) with non-zero mean : 5440.604
 ARIMA(0,0,1) with non-zero mean : 5439.433
 ARIMA(0,0,0) with zero mean     : 5436.221
 ARIMA(1,0,2) with non-zero mean : 5426.748
 ARIMA(2,0,1) with non-zero mean : 5425.039
 ARIMA(3,0,2) with non-zero mean : 5320.076
 ARIMA(3,0,1) with non-zero mean : 5425.859
 ARIMA(4,0,2) with non-zero mean : 5286.876
 ARIMA(4,0,1) with non-zero mean : 5426.084
 ARIMA(5,0,2) with non-zero mean : 5277.295
 ARIMA(5,0,1) with non-zero mean : 5409.185
 ARIMA(5,0,3) with non-zero mean : 5270.667
 ARIMA(4,0,3) with non-zero mean : 5310.305
 ARIMA(5,0,4) with non-zero mean : 5157.74
 ARIMA(4,0,4) with non-zero mean : 5300.287
 ARIMA(5,0,5) with non-zero mean : 5157.049
 ARIMA(4,0,5) with non-zero mean : 5168.893
 ARIMA(5,0,5) with zero mean     : 5154.995
 ARIMA(4,0,5) with zero mean     : 5166.844
 ARIMA(5,0,4) with zero mean     : 5155.691
 ARIMA(4,0,4) with zero mean     : 5298.242

 Now re-fitting the best model(s) without approximations...

 ARIMA(5,0,5) with zero mean     : 5152.264

 Best model: Regression with ARIMA(5,0,5) errors

> covid_arima_auto
Series: covid_ts_d$new_cases_per_million
Regression with ARIMA(5,0,5) errors

Coefficients:
         ar1      ar2     ar3      ar4      ar5      ma1     ma2      ma3     ma4     ma5  vaccination
      0.2420  -1.0127  0.0499  -0.5661  -0.4245  -0.2815  1.0394  -0.1833  0.5748  0.1295       0.0679
s.e.  0.0833   0.0644  0.1142   0.0609   0.0792   0.0871  0.0662   0.1086  0.0576  0.0750       3.0023

sigma^2 estimated as 10.49:  log likelihood=-2563.97
AIC=5151.94   AICc=5152.26   BIC=5210.72
>
```

Model Diagnostics

I wanted to check if my ARIMA model would be a good fit and also if the residuals satisfy the normality assumption. I did these two tests to check these.

Box-Ljung Test

The Ljung-Box test, named after statisticians Greta M. Ljung and George E.P. Box, is a statistical test that checks if auto-correlation exists in a time series.

The null hypothesis H_0:No auto-correlation; The residuals are independently distributed. that means the model does not exhibit lack of fit.

The alternative hypothesis H_A: There is an auto-correlation;The residuals are not independently distributed. The model exhibits lack of fit.

From this test, we can see we can not reject the null hypothesis based on the p value

$$0.8256 > 0.05$$

That means the model is a good fit.

```
> Box.test(covid_arima_auto$residuals, type = "Ljung-Box")

        Box-Ljung test

data:  covid_arima_auto$residuals
X-squared = 0.048568, df = 1, p-value = 0.8256
```

Shapiro-Wilk Normality Test

The Shapiro–Wilk test can be used to decide whether or not a sample fits a normal distribution, and it is commonly used for small samples.

The null hypothesis H_0 : residuals are normally distributed

The alternative hypothesis H_A : residuals are not normally distributed

From the test, we can see that we have to reject the null hypothesis based on the p value

$$2.2e - 16 < 0.05$$

That means residuals are not normally distributed.

```
> shapiro.test(covid_arima_auto$residuals)

        Shapiro-Wilk normality test

data:  covid_arima_auto$residuals
W = 0.66462, p-value < 2.2e-16
```

GARCH Procedure

An ARCH (autoregressive conditionally heteroscedastic) model is a model for the variance of a time series. ARCH models are used to describe a changing, possibly volatile variance. A GARCH (generalized autoregressive conditionally heteroscedastic) model uses values of the past squared observations and past variances to model the variance at time t.

As the ARIMA model errors did not appear to be normal, i had to use GARCH model for my dataset.

At first i checked if there is any ARCH effect in my model.

```
> ArchTest(covid_ts_d_del)

        ARCH LM-test; Null hypothesis: no ARCH effects

data:  covid_ts_d_del
Chi-squared = 830.31, df = 12, p-value < 2.2e-16
```

From my ARCH test, it was shown that the hypothesis of having no ARCH effect was rejected based on the p-value

$$2.2e-16 < 0.05$$

That means, my model has an ARCH effect.

Then i checked what garch order can be used for my model. It appears to be GARCH(1,1) can be a good model for my data.

```
> garch(covid_ts_d_del[,c("new_cases_per_million")],grad="numerical",trace=FALSE)

Call:
garch(x = covid_ts_d_del[, c("new_cases_per_million")], grad = "numerical",     trace = FALSE)

Coefficient(s):
       a0        a1        b1
0.0006705  0.4626402  0.6946445
```

Then i input all my specification to the ugarchfit and finally found all the parameters for my data. I have also added "Vaccination" as a external regressor in my model fit. I have used my previously found ARIMA(5,0,5) model here too.

```
*----------------------------------*
*          GARCH Model Fit         *
*----------------------------------*

Conditional Variance Dynamics
-----------------------------------
GARCH Model     : apARCH(1,1)
Mean Model      : ARFIMA(5,0,5)
Distribution    : std

Optimal Parameters
------------------------------------
        Estimate  Std. Error    t value Pr(>|t|)
mu      0.000000    0.000007    0.00000 1.000000
ar1     0.319570    0.002537  125.94917 0.000000
ar2    -0.995927    0.004059 -245.38350 0.000000
ar3     0.112894    0.007393   15.27046 0.000000
ar4    -0.524127    0.002163 -242.36499 0.000000
ar5    -0.361173    0.000866 -417.15492 0.000000
ma1    -0.534781    0.003808 -140.41921 0.000000
ma2     1.035748    0.030413   34.05592 0.000000
ma3    -0.388317    0.005686  -68.28936 0.000000
ma4     0.550661    0.007182   76.67092 0.000000
ma5     0.052356    0.005626    9.30663 0.000000
mxreg1  0.027038    0.028652    0.94366 0.345344
omega   0.005803    0.002803    2.07023 0.038430
alpha1  0.152042    0.004362   34.85871 0.000000
beta1   0.862334    0.007914  108.96525 0.000000
gamma1 -0.642766    0.048335  -13.29827 0.000000
delta   0.147429    0.050119    2.94160 0.003265
vxreg1  0.000000    0.089043    0.00000 1.000000
shape   3.821248    0.402373    9.49678 0.000000

Robust Standard Errors:
        Estimate  Std. Error    t value Pr(>|t|)
mu      0.000000    0.001540   0.000000 1.000000
ar1     0.319570    0.038200   8.365664 0.000000
ar2    -0.995927    0.019261 -51.705821 0.000000
ar3     0.112894    0.125351   0.900619 0.367791
ar4    -0.524127    0.009255 -56.630688 0.000000
ar5    -0.361173    0.009168 -39.397023 0.000000
ma1    -0.534781    0.070767  -7.556908 0.000000
ma2     1.035748    0.586002   1.767482 0.077148
ma3    -0.388317    0.063226  -6.141733 0.000000
ma4     0.550661    0.081037   6.795171 0.000000
ma5     0.052356    0.090538   0.578281 0.563075
mxreg1  0.027038    0.538253   0.050232 0.959937
omega   0.005803    0.026359   0.220161 0.825746
alpha1  0.152042    0.099561   1.527128 0.126729
```

```
LogLikelihood : -1285.632

Information Criteria
------------------------------------

Akaike       2.6356
Bayes        2.7296
Shibata      2.6349
Hannan-Quinn 2.6714

Weighted Ljung-Box Test on Standardized Residuals
------------------------------------
                         statistic p-value
Lag[1]                       2.213  0.1369
Lag[2*(p+q)+(p+q)-1][29]   149.144  0.0000
Lag[4*(p+q)+(p+q)-1][49]   187.404  0.0000
d.o.f=10
H0 : No serial correlation

Weighted Ljung-Box Test on Standardized Squared Residuals
------------------------------------
                        statistic p-value
Lag[1]                      105.1       0
Lag[2*(p+q)+(p+q)-1][5]     105.5       0
Lag[4*(p+q)+(p+q)-1][9]     106.8       0
d.o.f=2

Weighted ARCH LM Tests
------------------------------------
            Statistic Shape Scale P-Value
ARCH Lag[3]   0.01132 0.500 2.000  0.9153
ARCH Lag[5]   0.54432 1.440 1.667  0.8705
ARCH Lag[7]   1.69584 2.315 1.543  0.7813

Nyblom stability test
------------------------------------
Joint Statistic:  17.0586
Individual Statistics:
mu     4.029779
ar1    0.035954
ar2    0.102742
ar3    0.051320
ar4    0.108234
ar5    0.059287
ma1    0.020976
ma2    0.051503
ma3    0.004146
ma4    0.223010
ma5    0.012807
mxreg1 0.702850
```

Overall Result

From Pearson goodness of fit test, we can easily see that we can not reject the null hypothesis. That means there is no significant difference between the observed and the expected value.

```
Asymptotic Critical Values (10% 5% 1%)
Joint Statistic:          4.03 4.33 4.92
Individual Statistic:     0.35 0.47 0.75

Sign Bias Test
------------------------------------
                    t-value    prob sig
Sign Bias            2.4753 0.01348  **
Negative Sign Bias   0.4148 0.67837
Positive Sign Bias   0.6522 0.51442
Joint Effect         8.7543 0.03274  **


Adjusted Pearson Goodness-of-Fit Test:
------------------------------------
  group statistic p-value(g-1)
1    20     16.06       0.65323
2    30     30.30       0.39899
3    40     44.67       0.24584
4    50     62.53       0.09283


Elapsed time : 3.384429
```

Also, from GARCH model outputs p values for mxreg1 vxreg1, we can see there are no significance of external variable (Vaccination). So we can easily ignore those for our fitted model.

Fitted Model

$Y_t = 0.319570 * Y_{t-1} - 0.995927 * Y_{t-2} + 0.112894 * Y_{t-3} - 0.524127 * Y_{t-4} - 0.361173 * Y_{t-5} + e_t$
$-0.534781 * e_{t-1} + 1.035748 * e_{t-2} - 0.388317 * e_{t-3} + 0.550661 * e_{t-4} + 0.052356 ** e_{t-5}$

$$e_t = \sigma_t * \epsilon_t$$

$\sigma_t^{0.147429} = 0.005803 + 0.152042(|Y_{t-1}| + |Y_{t-2}| + |Y_{t-3}| + |Y_{t-4}| + |Y_{t-5}|$
$-(-0.642766) * (Y_{t-1} + Y_{t-2} + Y_{t-3} + Y_{t-4} + Y_{t-5})^{0.147429} + 0.862334 * \sigma_{t-1}^{0.147429}$

here,

$$\mu = 0$$

$$\delta > 0$$

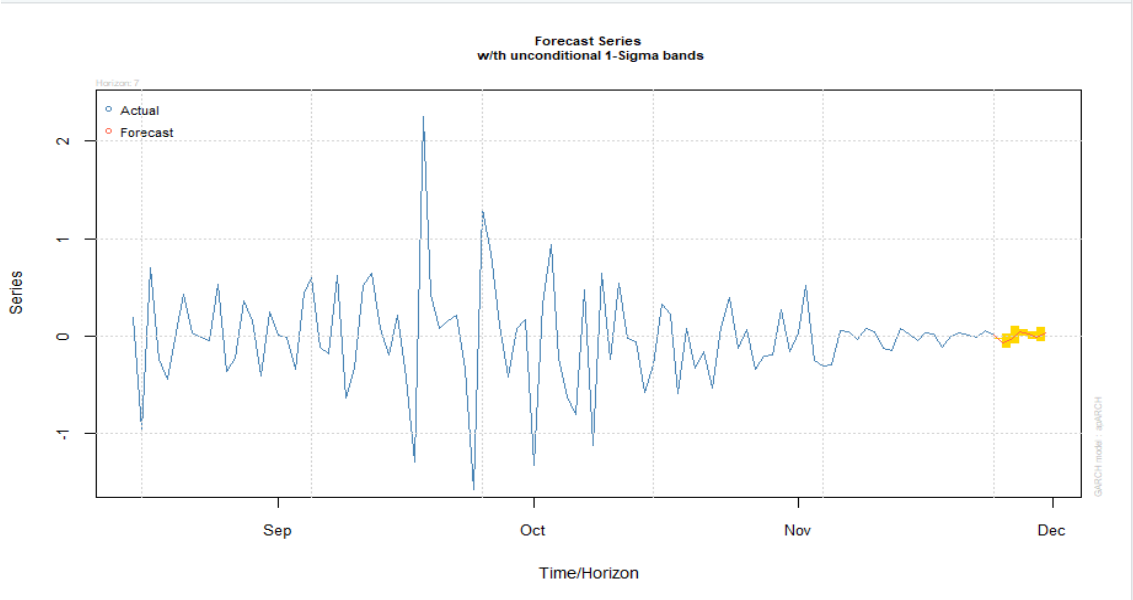
and

$$|\gamma| < 1$$

Prediction

After fitting my model, i have found out the next 7 days prediction using ugarchforecast. Overall, ARIMA(5,0,5) GARCH(1,1) model seems to be a good fit for my dataset.

```
> prediction<-ugarchforecast(garch_fit,data=covid_ts_d_del$new_cases_per_million,n.ahead=7)
> prediction

*----------------------------------*
*       GARCH Model Forecast       *
*----------------------------------*
Model: apARCH
Horizon: 7
Roll Steps: 0
Out of Sample: 0

0-roll forecast [T0=2022-11-23]:
       Series   Sigma
T+1  0.007803 0.04950
T+2 -0.072350 0.05043
T+3 -0.032744 0.05137
T+4  0.046722 0.05232
T+5  0.014850 0.05328
T+6 -0.010380 0.05425
T+7  0.030460 0.05523
```



References

1. https://covid19.who.int/region/searo/country/bd
2. https://online.stat.psu.edu/stat510/lesson/11/11.1
3. https://www.quantstart.com/articles/Johansen-Test-for-Cointegrating-Time-Series-Analysis-in-R/
4. https://rpubs.com/sdkshihsoj/ATSA: :text=In