Statistics Assignment

Analysis of Time series Data

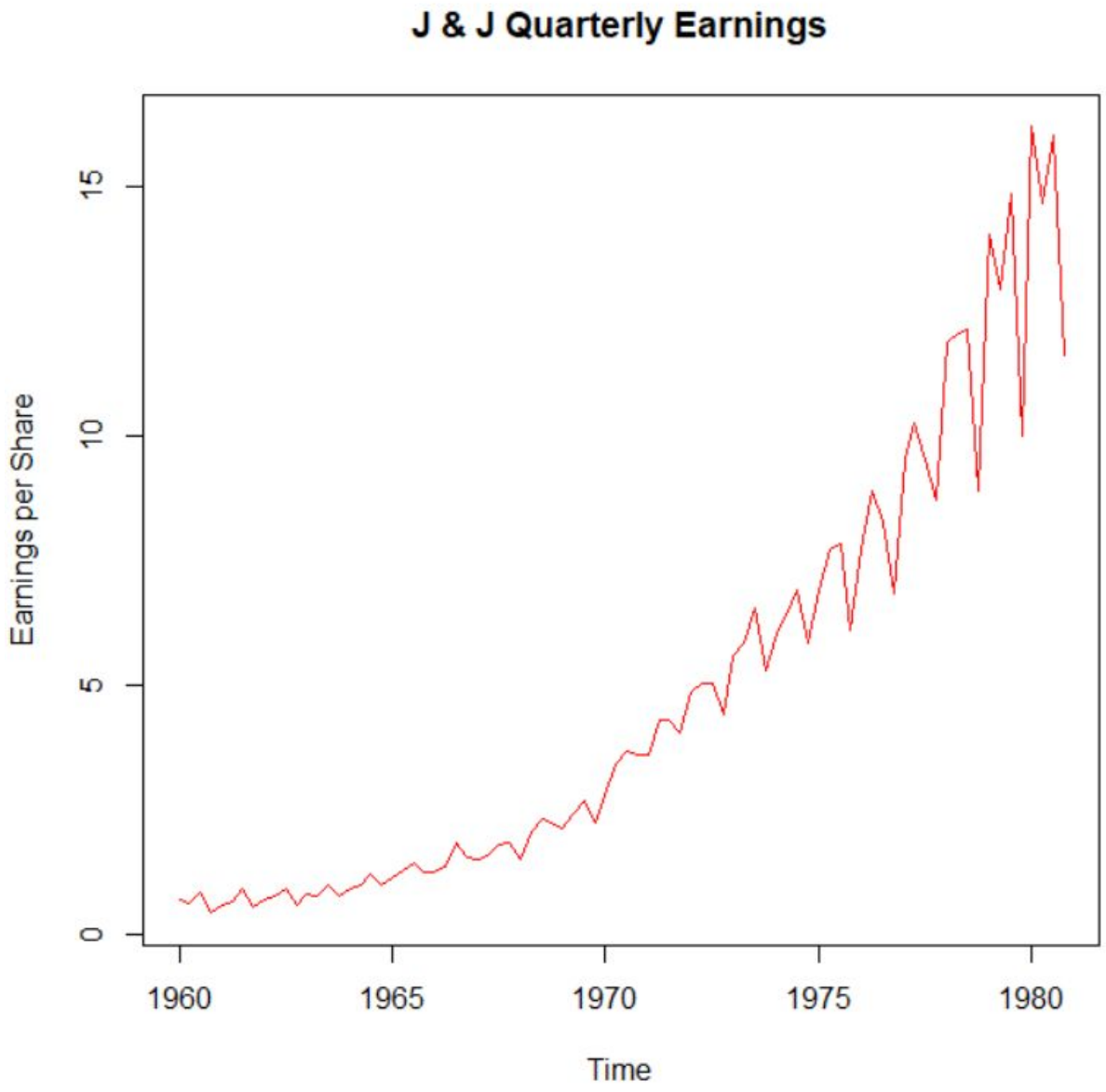
By : Prasun De
Roll No : BS1826

We are going to look at the dataset : **Quarterly Earnings per Johnson and Johnson share from 1960 to 1980**

We shall fit an appropriate SARIMA model to this data .

Fitting SARIMA model

Step 1 : Load the data and then plot the time series



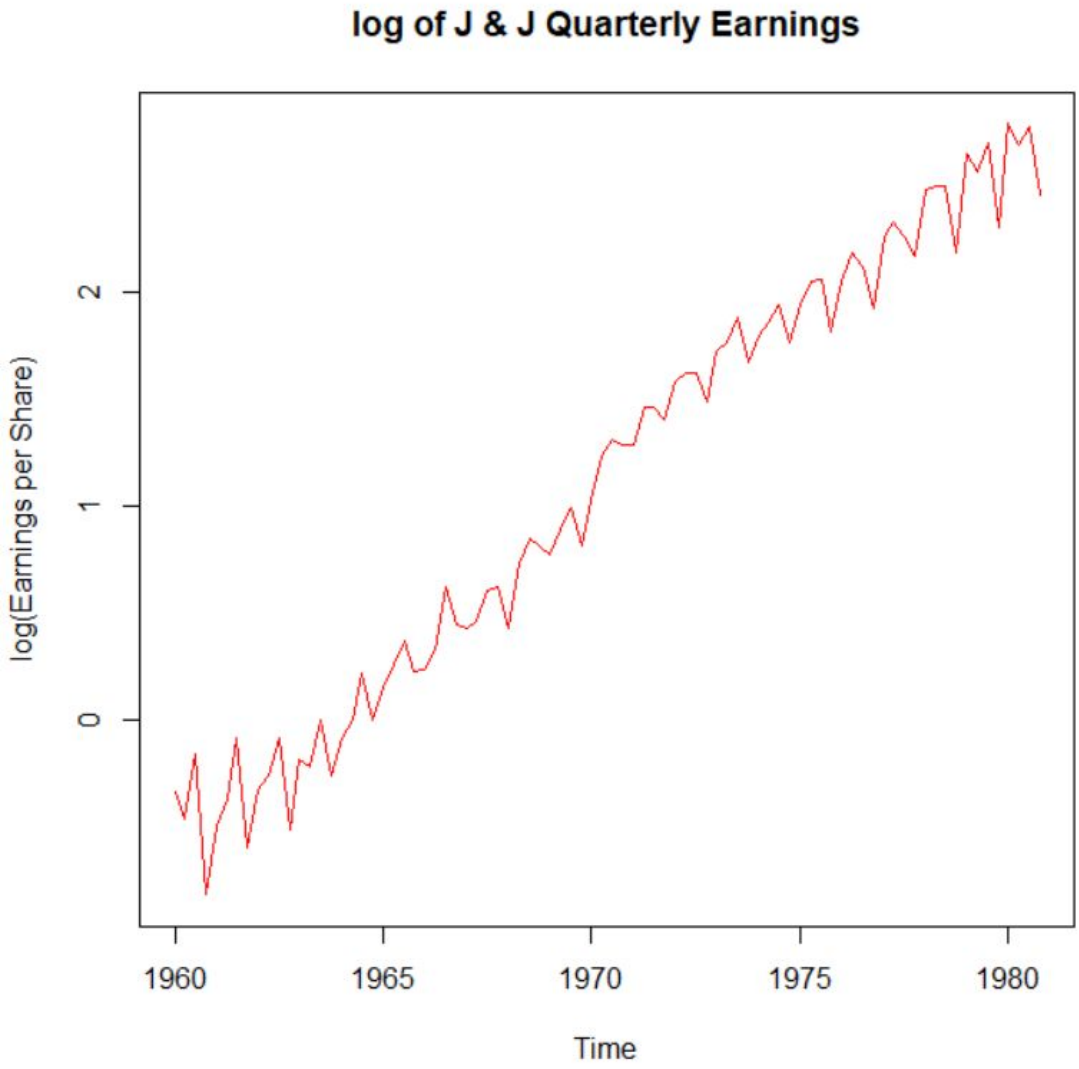
We can immediately observe that is an increasing trend in this plot . Also , the variance seems to be increasing with time .

We can also expect to have seasonality in this quarterly data . We will find some seasonal behaviour every quarter .

We will address the issue of increasing variance by doing a variance stabilizing transformation :

We will consider log(Xt) for the time series {X1 , X2 , ….} .

Let us now plot the transformed time-series:



Step 2 : Doing the necessary differencing

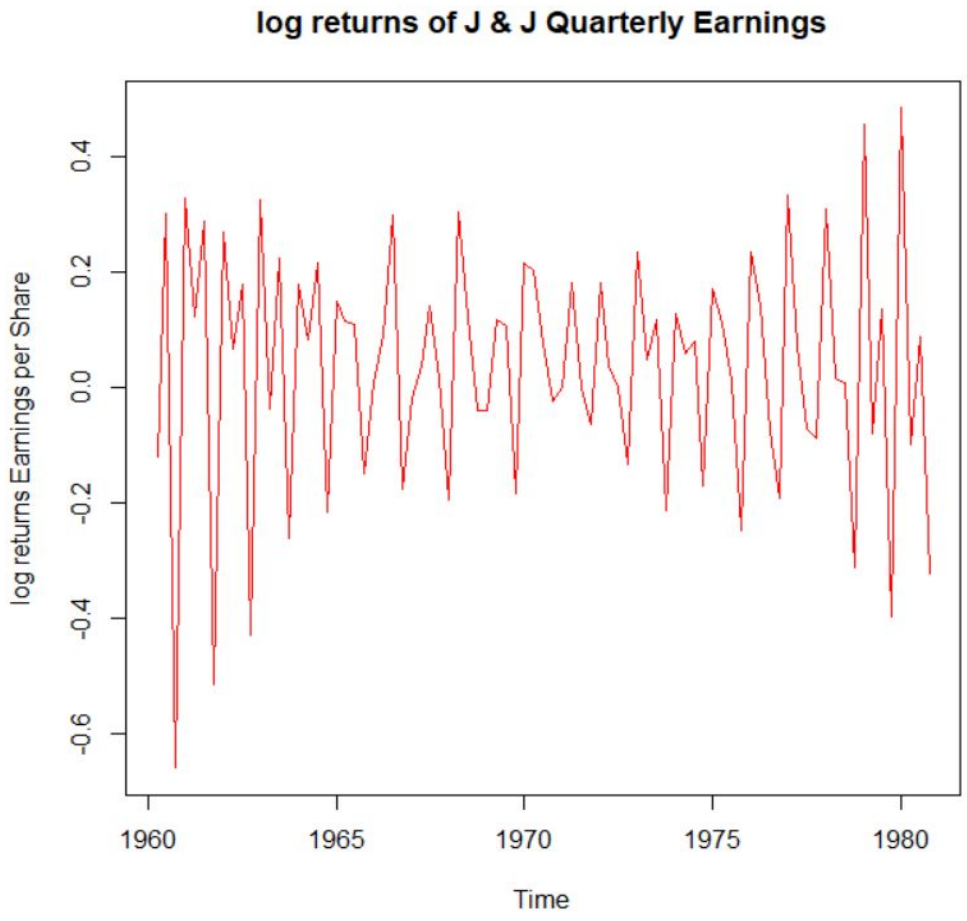
We observe that the variance seems quite stationary now . Also , there is a clear linear trend in this data .

To remove this linear trend , we consider the first order difference :

$$\log(X_t) - \log(X_{t-1})$$

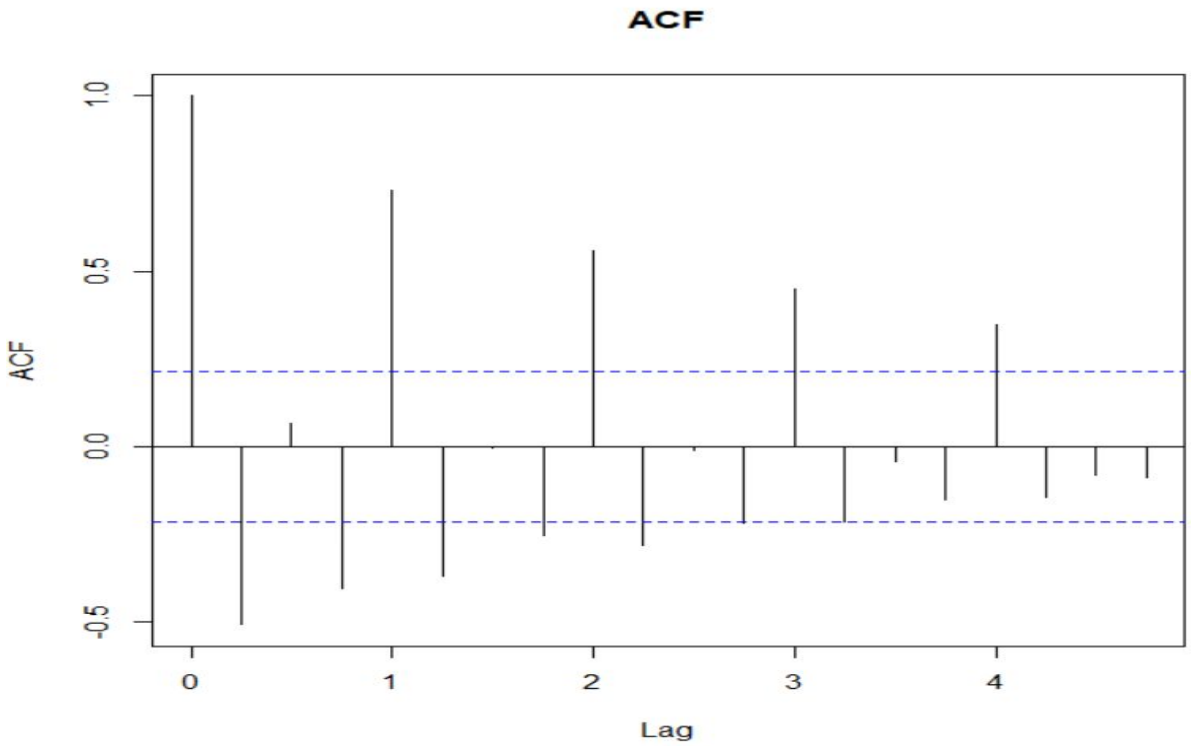
This is also known as log return

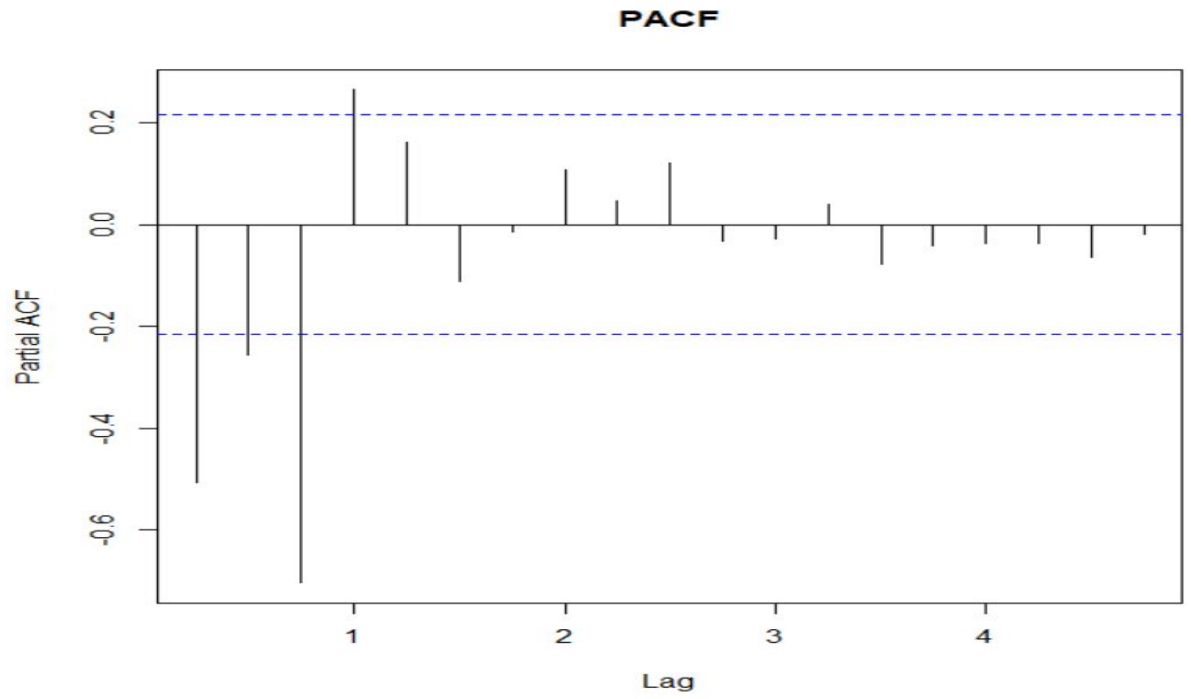
Applying this difference to the time series , we get :



The variance is approximately constant (though it does seem a little less in the middle) . There is clearly no trend in this transformed time series .

Let us plot the ACF and PACF of this differenced dat



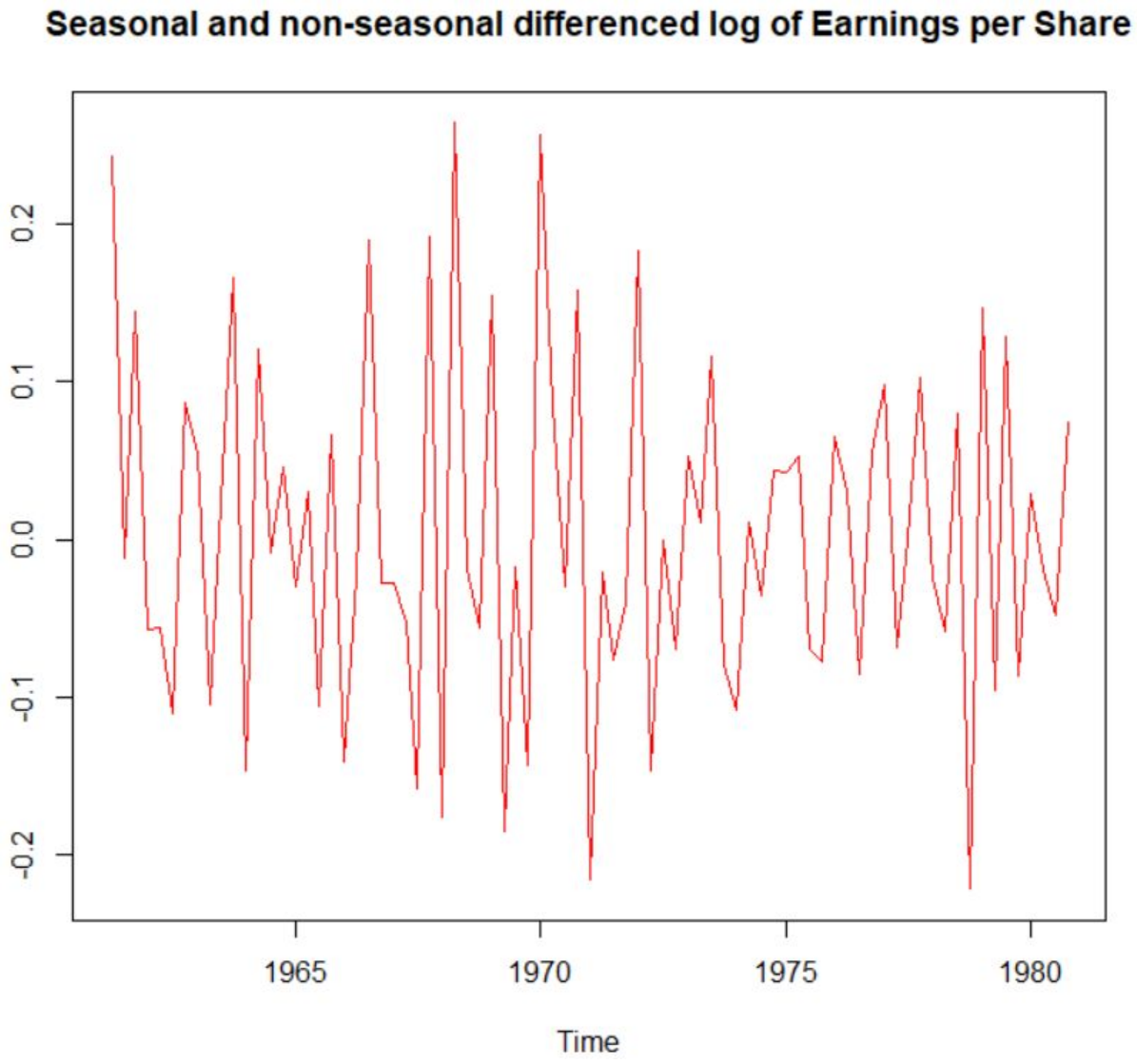


We have very strong Auto correlation for lags in multiple of 4 (or integral lags in this

graph) . Note that the data is quarterly so we have decimal lags because frequency of time series = 4 .

We also have spike in PACF for lag 4 (ie; corresponding to lag 1 acc to the graph)

These patterns in the ACF and PACF suggest that there is seasonality in the data . We should do seasonal differencing with lag 4 to get rid of this seasonality .



We have a stationary time series now .

To test that the time series is stationary , we look at the Ljung-Box test for stationarity in data :

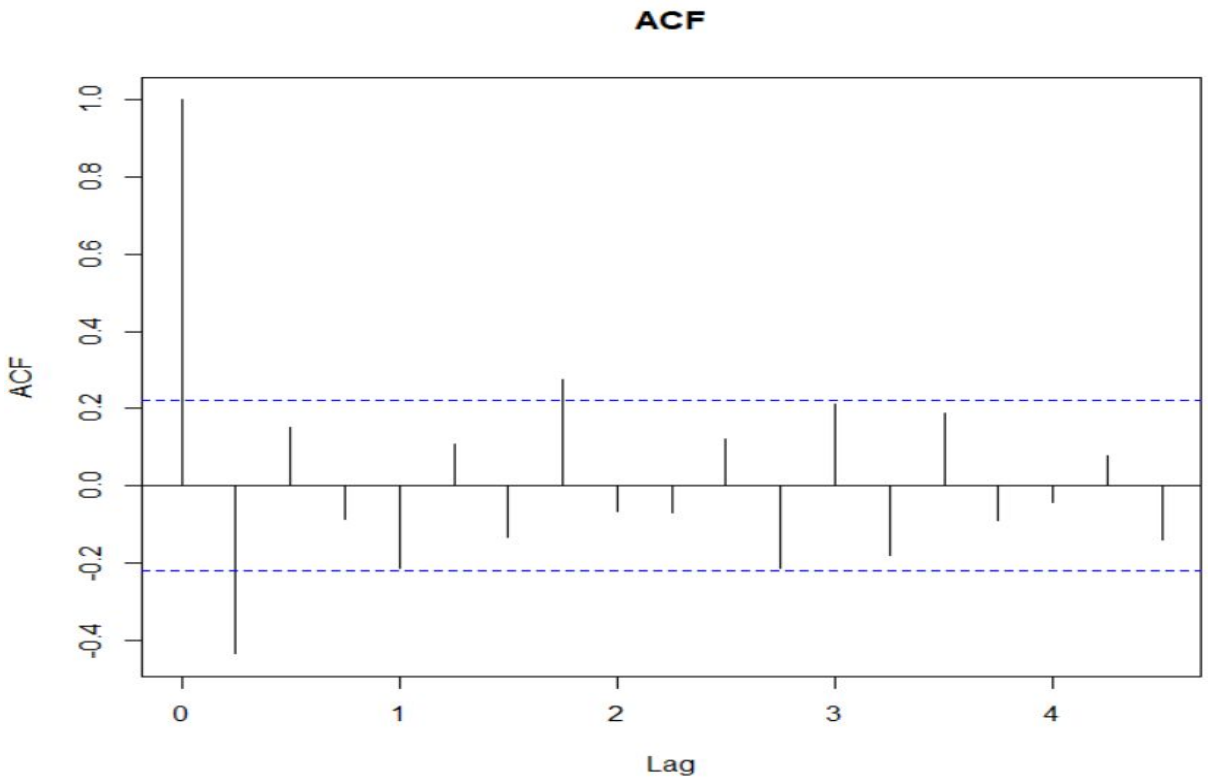
We use the Box.test() function in R to do this test .

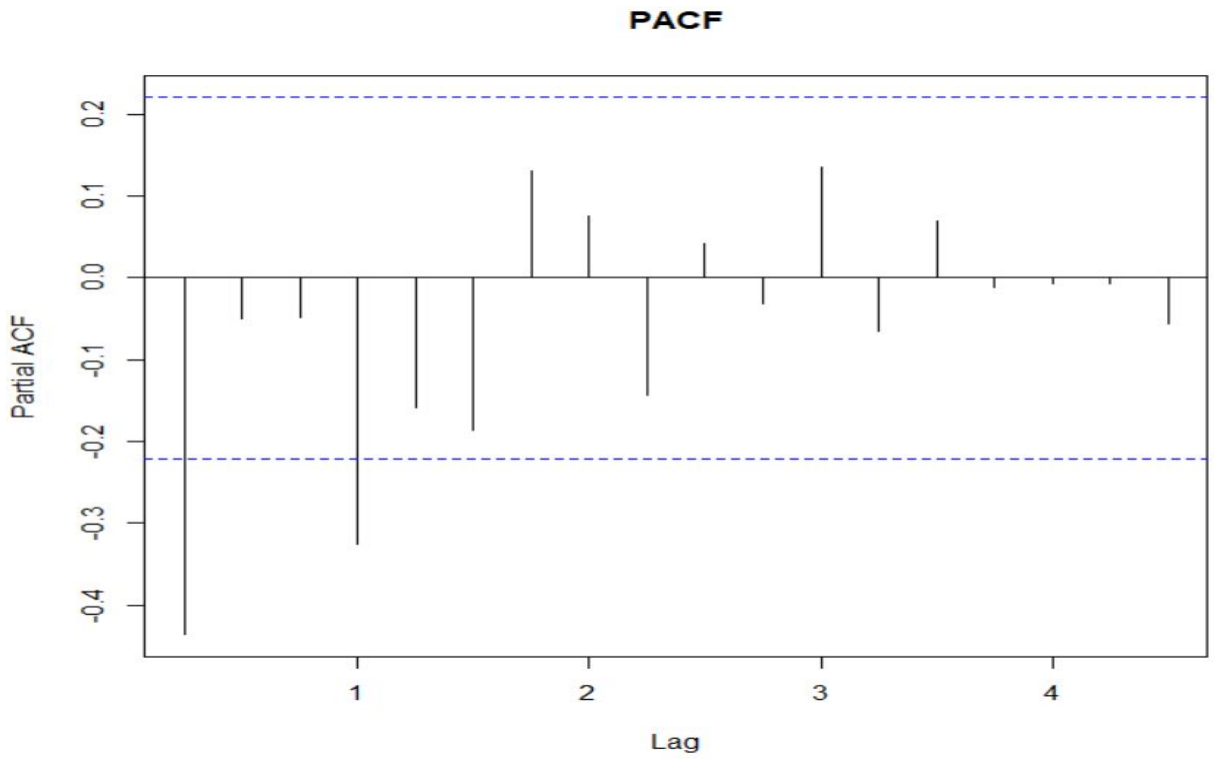
```
> data = diff(diff(log(jj)),4)
> Box.test(data , lag = log(length(data)))

        Box-Pierce test

data:  data
X-squared = 20.95, df = 4.3694, p-value = 0.0004658
```

We took lag as the length of the data which is quite common . The small p-value helps us to reject the null hypothesis that there is no autocorrelation between earlier lags of seasonal and non-seasonal differenced log of earnings per J&J share .

So , there is some autocorrelation between the previous lags . We will estimate that by plotting the ACF and PACF .

Step 3 : Examine the Acf and PACF of transformed data





So , we now analyze the ACF and PACF to get an idea of the parameters of the SARIMA model that we will fit to the data.

We will look at SARIMA(p,1,q,P,1,Q)4 models . we have 1st order non seasonal difference which contributes to the first ‘1’ and 1st order seasonal difference which contributes to the second ‘1’ . Seasonality

is of 4 months , so we have the 4 in the model .

We look at the ACF , PACF to estimate possible values of p,q,P,Q

There is a downward spike in the 1st lag in the ACF plot (corresponding to lag at 0.25 according to the plot because frequency of time series is 4) and then it dies down. This suggest a possible MA(1) term .

Now the value of acf at lag 4 (or lag 1 in the ACF plot as it is period 1) is ‘almost’ significant .We might have some seasonal correlation so this may hint at a possible seasonal MA(1) term .

So q = 0,1 and Q = 0,1 are possibilities

There is a downward spike in the 1st lag in the PACF plot (corresponding to lag at 0.25 according to the plot because frequency of time series is 4) and then it dies down. This suggest a possible AR(1) term .

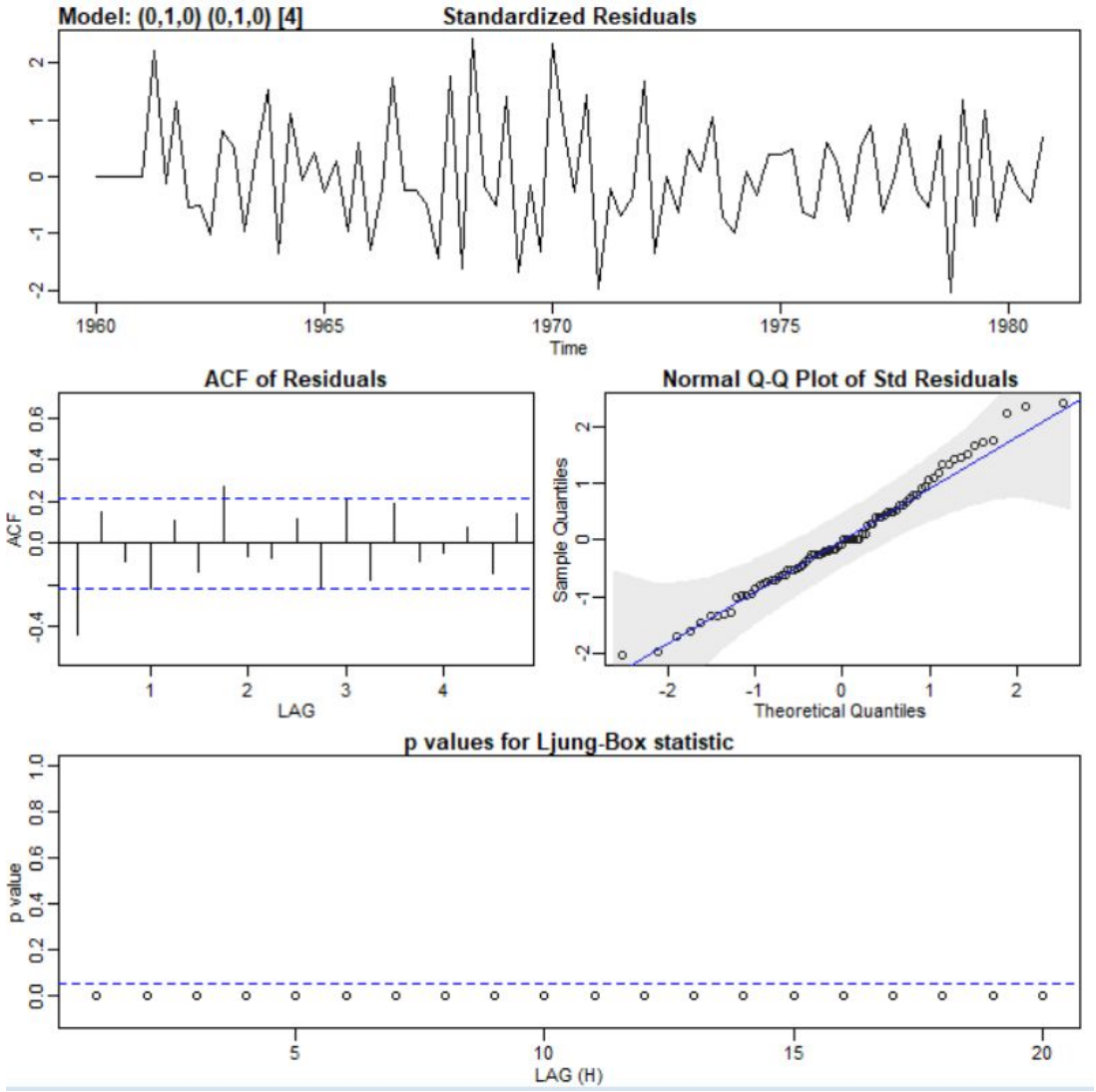
Now the value of pacf at lag 4 (or lag 1 in the PACF plot as it is period 1) is significant . This may hint at a possible seasonal AR(1) term .

So p = 0,1 and P = 0,1 are possibilities .

So we will look at SARIMA(p,1,q,P,1,Q)4 models for log(JJ data) where

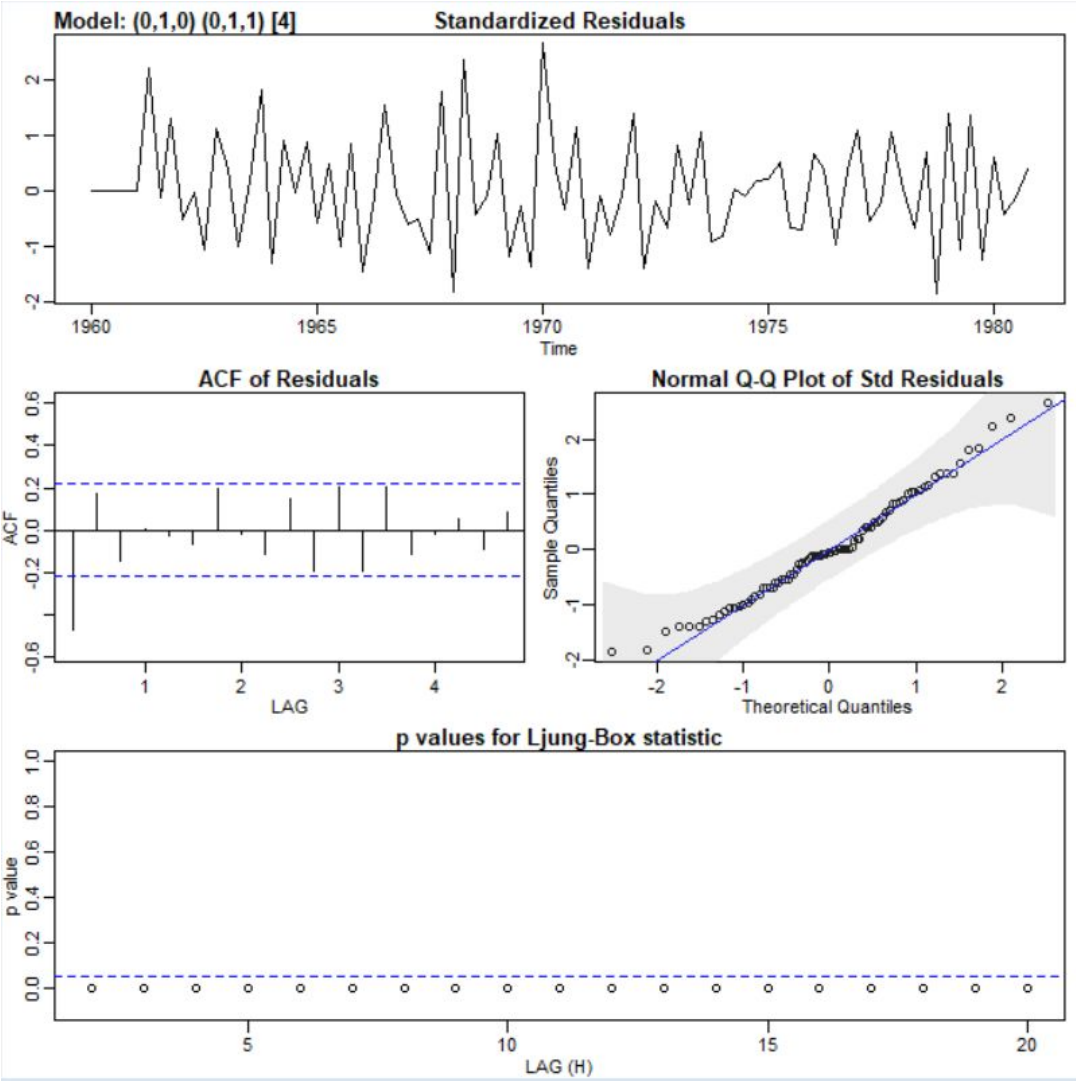
$0 \leq p,q,P,Q \leq 1$

ARIMA(0,1,0) × (0,1,0)4



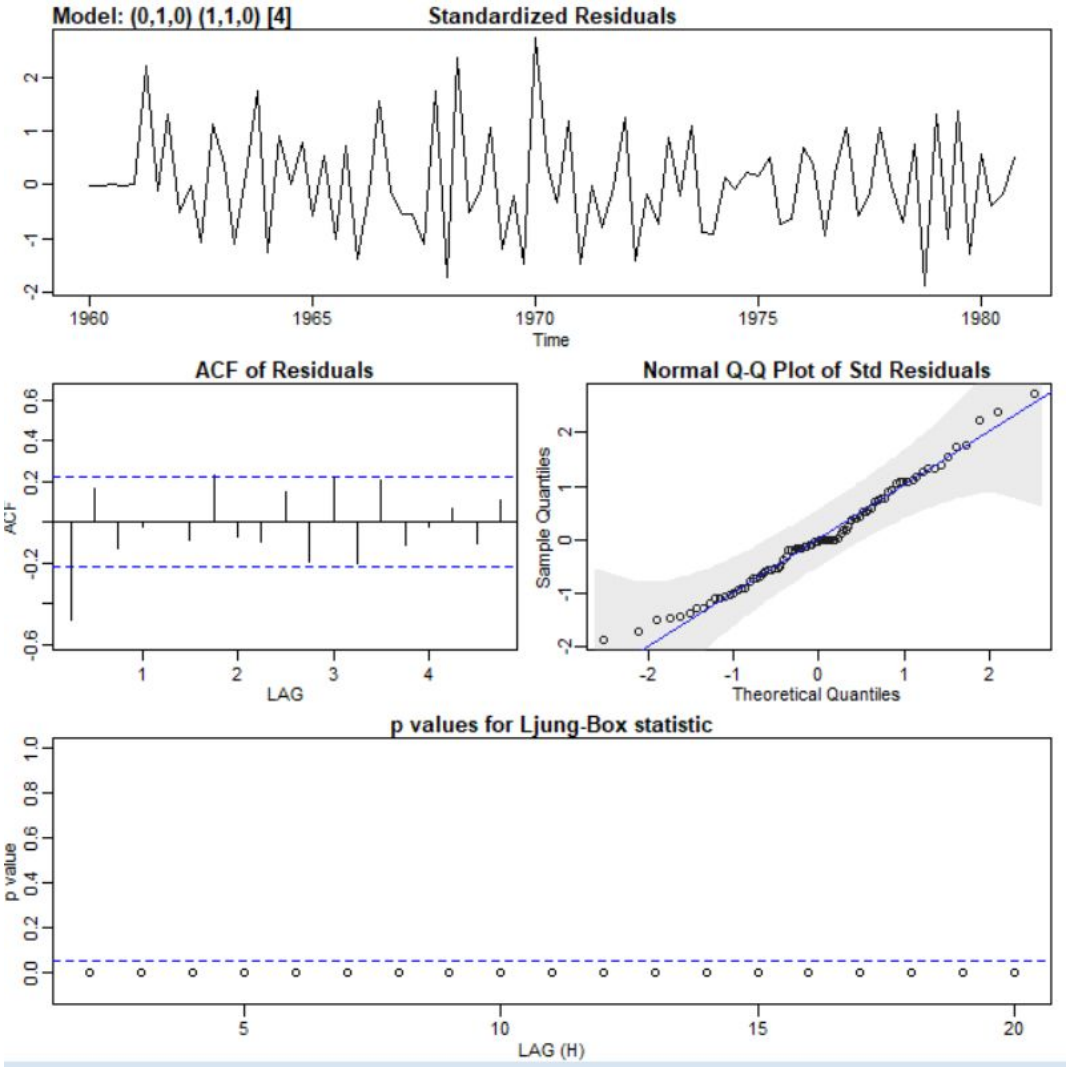
The p values for the Ljung - Box statistic are bad , and we see spikes in the ACF of the residuals . So , we have to reject this model.

ARIMA(0,1,0) × (0,1,1)4



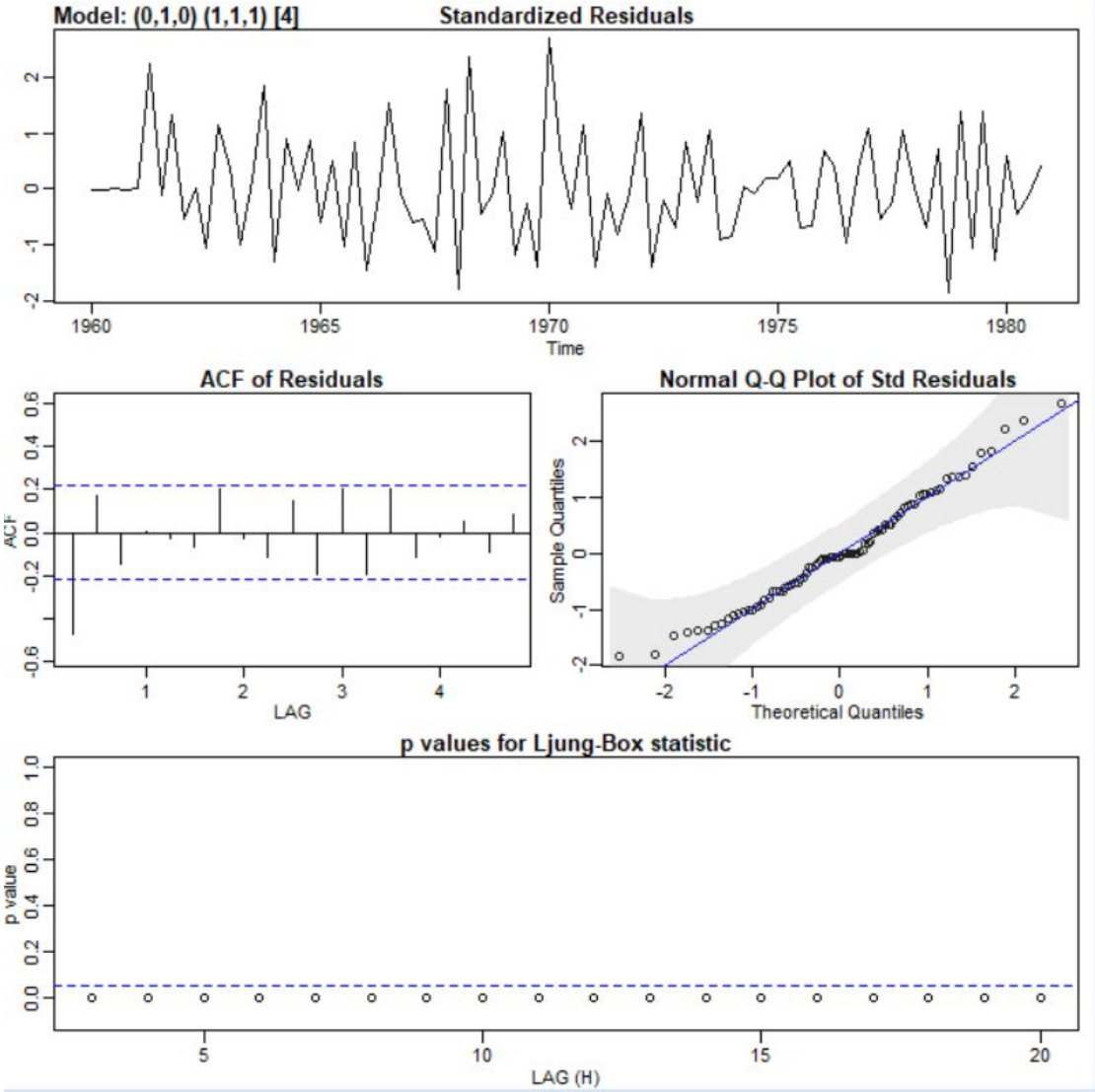
The p values for the Ljung - Box statistic are bad , and we see spikes in the ACF of the residuals . Also , p values of coefficients of the arima fit are greater than 0.05 . So , we have to reject this model.

ARIMA(0,1,0) × (1,1,0)4



The p values for the Ljung - Box statistic are bad , and we see spikes in the ACF of the residuals . So , we have to reject this model.

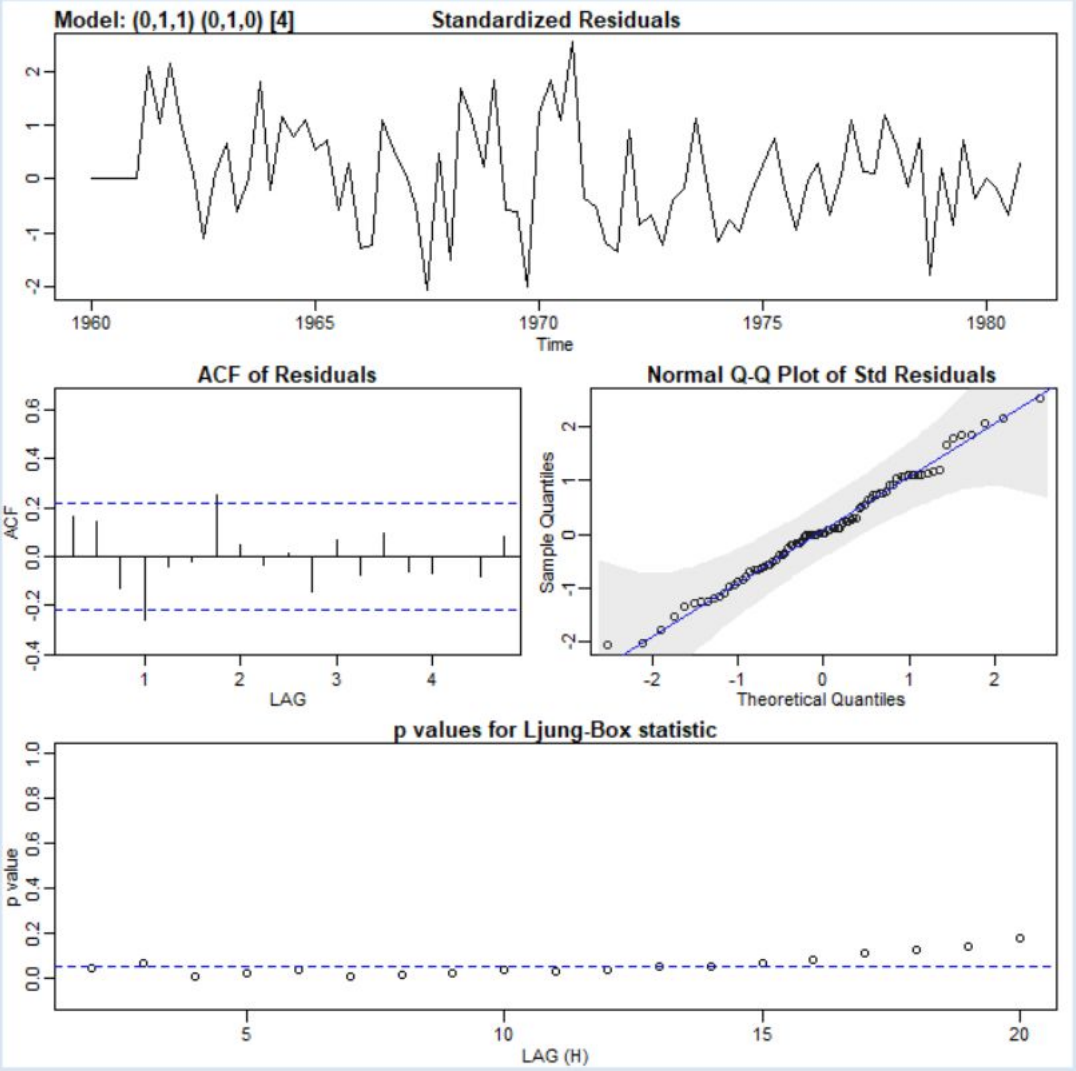
ARIMA(0,1,0) × (1,1,1)4



The p values for the Ljung - Box statistic are bad , and we see spikes in the ACF of the residuals . Also , p values of coefficients

of the arima fit are greater than 0.05 . So , we have to reject this model.

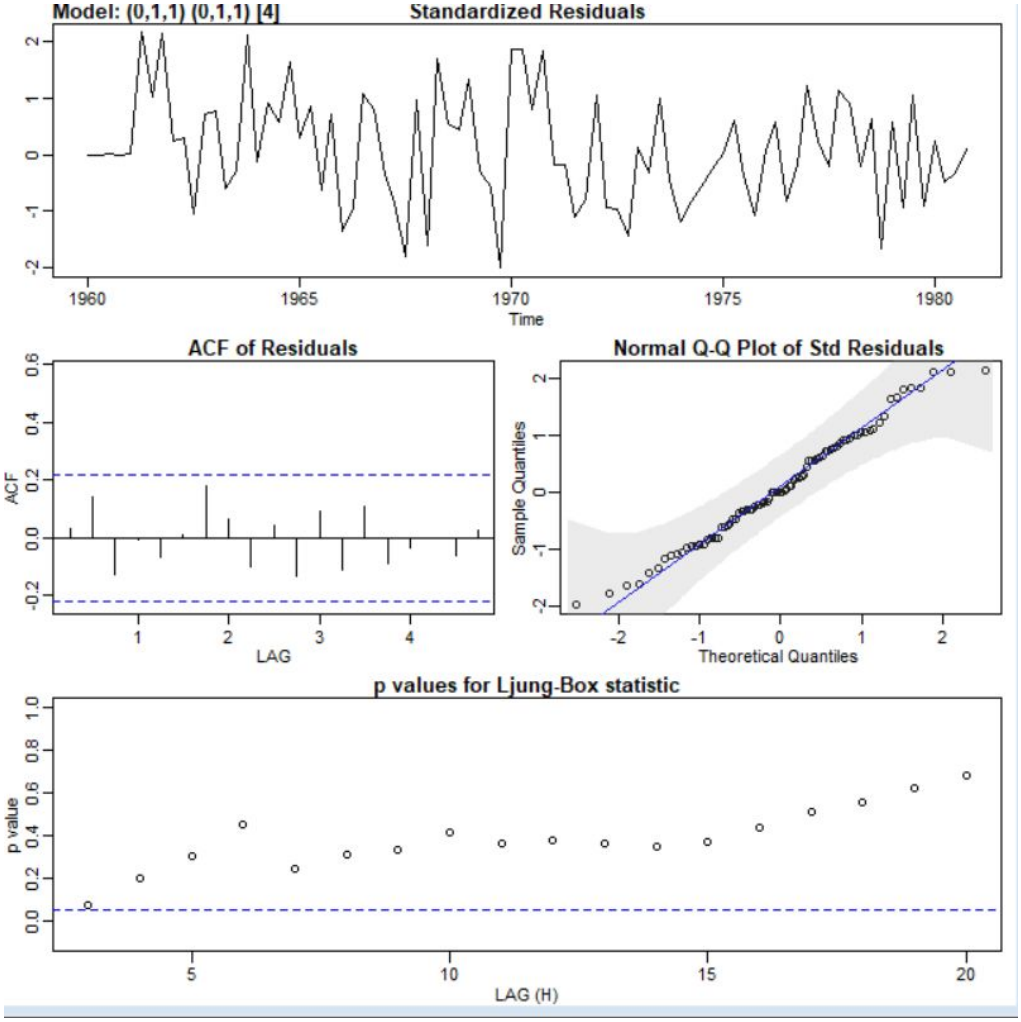
ARIMA(0,1,1) × (0,1,0)4



The p values for the Ljung - Box statistic are bad , and we see spikes in the ACF of

the residuals . So , we have to reject this model.

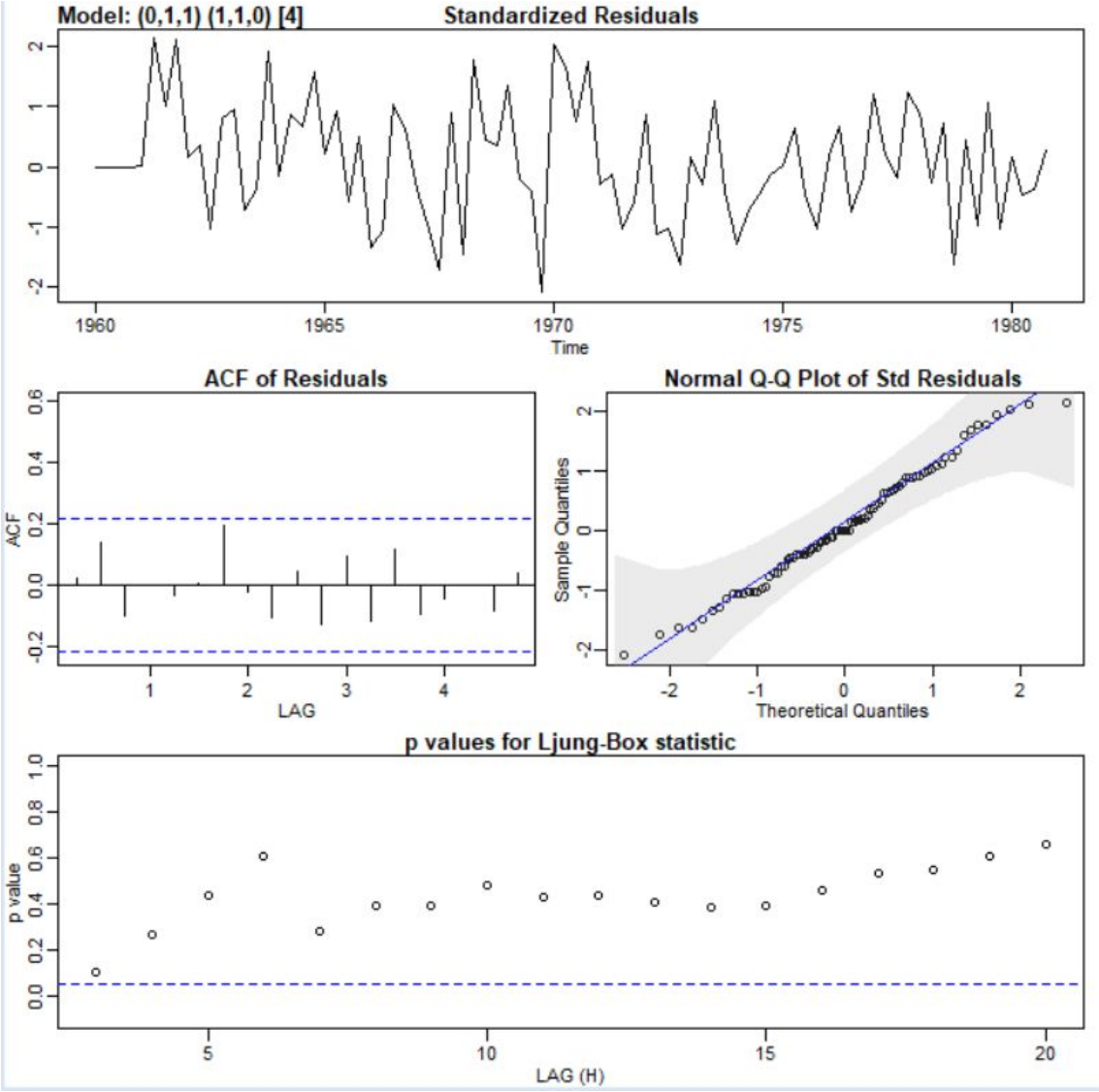
ARIMA(0,1,1) × (0,1,1)4 (Airline model)



We see that the fit is not bad due to non significant ACF of residuals and good p

values for Ljung Box statistic . We note that : AIC = -150.7528 , BIC = -143.4604

ARIMA(0,1,1)×(1,1,0)4



We see that the fit is not bad due to non significant ACF of residuals and good p

values for Ljung Box statistic . Also , the p values of the coefficients of the arima model are less than 0.05 . But , observe that for this model : AIC = -150.9134 , BIC = -143.621 . Both are smaller than that of the previous model . So we prefer this sarima fit over the previous one.

ARIMA(0,1,1) × (1,1,1)4

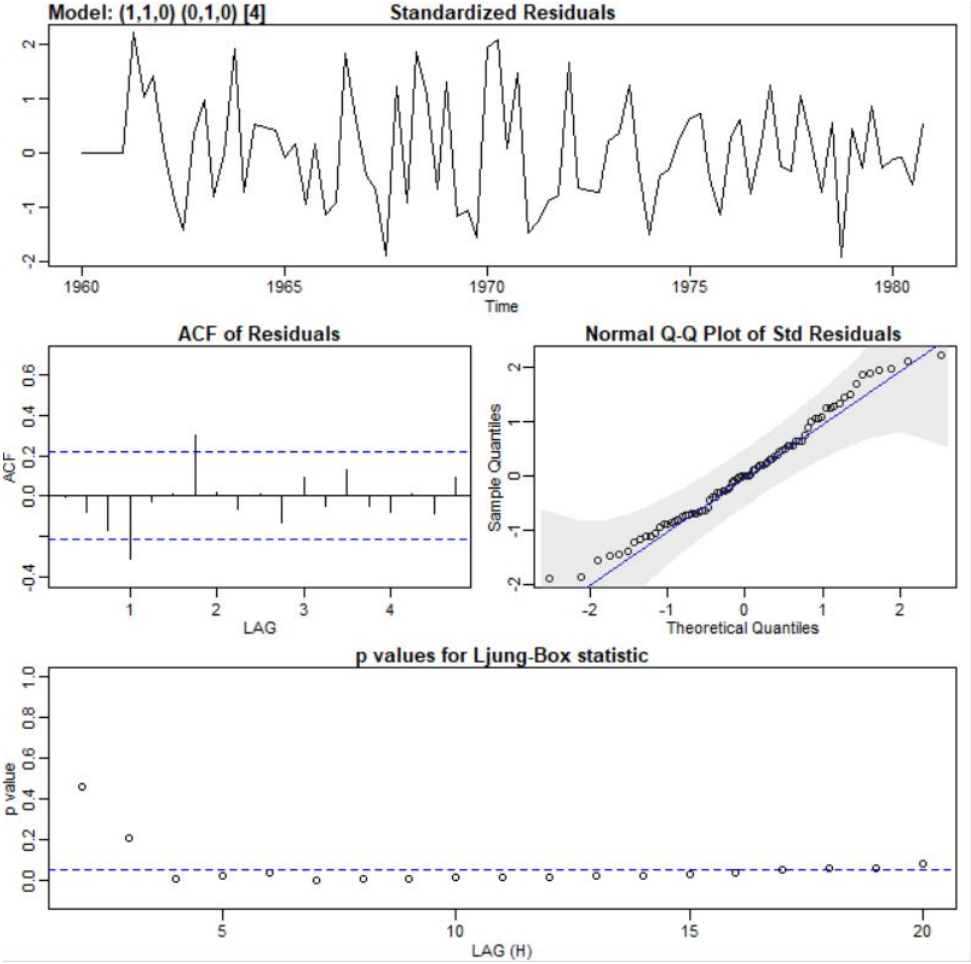
We look at the p values of the coefficients of the various terms below :

```
      Estimate     SE t.value p.value
ma1   -0.6749 0.0970 -6.9595  0.0000
sar1  -0.2004 0.3008 -0.6663  0.5072
sma1  -0.1390 0.2991 -0.4646  0.6435
> |
```

We can observe that the p values of seasonal AR term and seasonal MA term are greater than 0.05 . This tells us that this cannot be used as a model as the coefficients are not significant . The AIC =

-149.1317 , BIC = -139.4084 , both greater than the ARIMA(0,1,1)$\times$(1,1,0)4 that we have chosen till now .

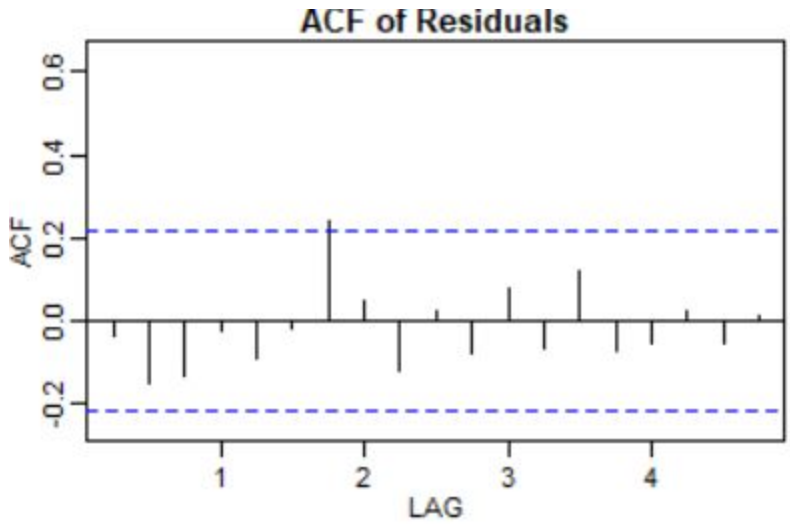
ARIMA(1,1,0)$\times$(0,1,0)4



We can observe significant spikes in the plot of the ACF of the residuals . Also , most of the the values of the Ljung box

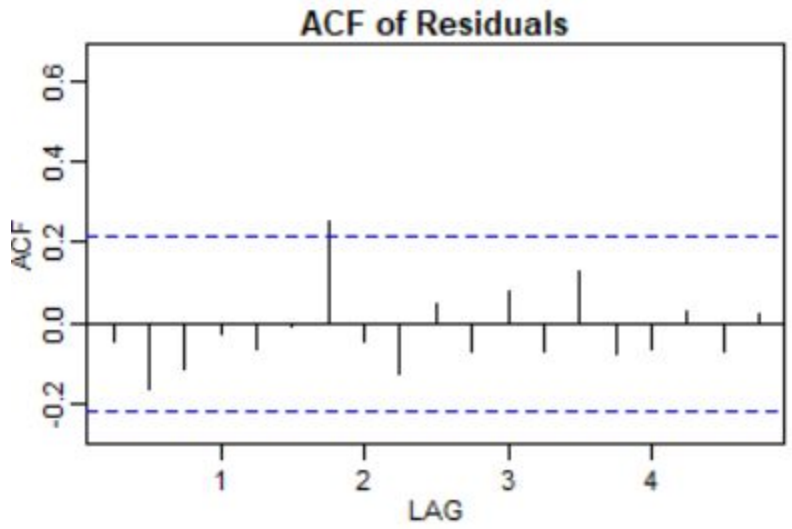
statistic are small . The AIC is -139.8248 and the BIC is -134.9632 both larger than the ARIMA(0,1,1)×(1,1,0)4 that we have chosen till now . So we reject this model .

ARIMA(1,1,0)×(0,1,1)4



We can observe a significant spike in the plot of the ACF of the residuals . The AIC = -146.0191 and BIC = -138.7266 , both greater than the ARIMA(0,1,1)×(1,1,0)4 that we have chosen till now . So we reject this model .

ARIMA(1,1,0)×(1,1,0)4



We can observe a significant spike in the plot of the ACF of the residuals . The AIC = -146.0319 and BIC = -138.7394 , both greater than the ARIMA(0,1,1)$\times$(1,1,0)4 that we have chosen till now . So we reject this model .

ARIMA(1,1,0)$\times$(1,1,1)4

```
     Estimate     SE t.value p.value
ar1   -0.5134 0.1009 -5.0893  0.0000
sar1  -0.1845 0.2892 -0.6378  0.5255
sma1  -0.1655 0.2827 -0.5855  0.5599
> |
```

We can observe that the p values of the seasonal AR and seasonal MA are greater

than 0.05 , so they are not statistically significant . The AIC = -144.3766 , BIC = -134.6534 , both greater than the ARIMA(0,1,1)×(1,1,0)4 that we have chosen till now .This leads us to reject this model .

ARIMA(1,1,1)×(0,1,0)4

```
     Estimate     SE t.value p.value
ar1   0.2384 0.1518  1.5702  0.1205
ma1  -0.8891 0.0958 -9.2798  0.0000
> |
```

We can observe that the p values of the seasonal AR and seasonal MA are greater than 0.05 , so they are not statistically significant . The AIC = -144.3766 , BIC = -134.6534 , both greater than the ARIMA(0,1,1)×(1,1,0)4 that we have chosen till now .This leads us to reject this model .

ARIMA(1,1,1) × (0,1,1)4

```
      Estimate     SE t.value p.value
ar1     0.0275 0.2066  0.1331  0.8944
ma1    -0.6990 0.1646 -4.2476  0.0001
sma1   -0.3072 0.1219 -2.5203  0.0138
> |
```

We can observe that the p value of the AR(1) term is greater than 0.05 , so this is not statistically significant . The AIC = -148.7706 and the BIC = -139.0473 , both greater than the ARIMA(0,1,1) × (1,1,0)4 that we have chosen till now .This leads us to reject this model .

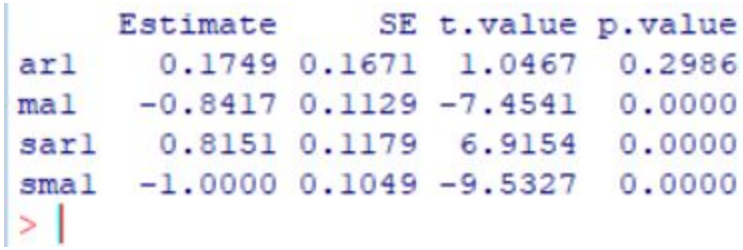
ARIMA(1,1,1) × (1,1,0)4

```
      Estimate     SE t.value p.value
ar1    -0.0141 0.2221 -0.0635  0.9495
ma1    -0.6700 0.1814 -3.6940  0.0004
sar1   -0.3265 0.1320 -2.4728  0.0156
```

We can observe that the p value of the AR(1) term is greater than 0.05 , so this is not statistically significant . The AIC = -148.9175 and the BIC = -139.1942 , both

greater than the ARIMA(0,1,1)×(1,1,0)4 that we have chosen till now .This leads us to reject this model .

ARIMA(1,1,1)×(1,1,1)4



```
      Estimate     SE t.value p.value
ar1     0.1749 0.1671  1.0467  0.2986
ma1    -0.8417 0.1129 -7.4541  0.0000
sar1    0.8151 0.1179  6.9154  0.0000
sma1   -1.0000 0.1049 -9.5327  0.0000
> 
```

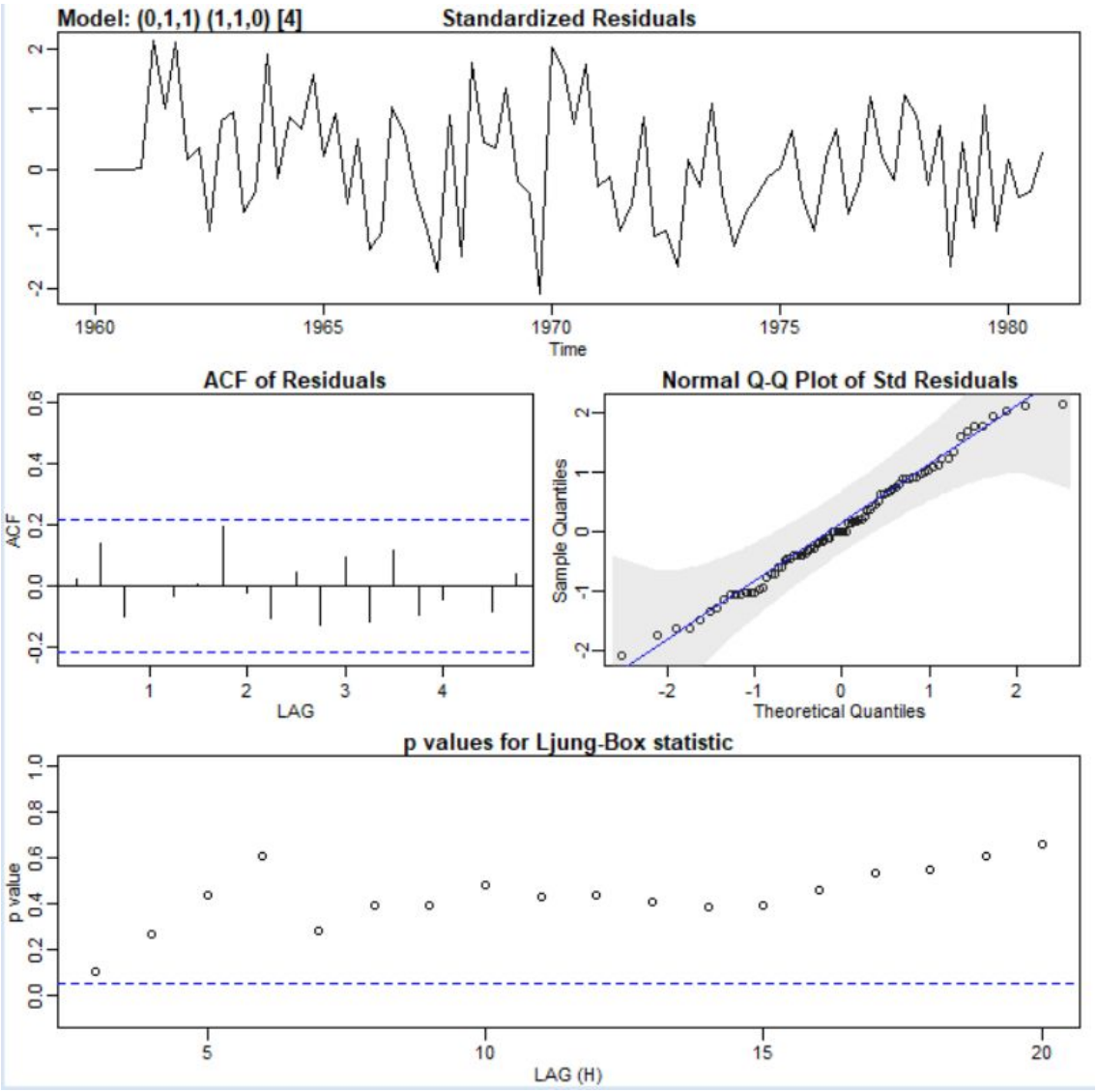
We can observe that the p value of the AR(1) term is greater than 0.05 , so this is not statistically significant . The AIC = -144.4483 and the BIC = -132.2942, both greater than the ARIMA(0,1,1)×(1,1,0)4 that we have chosen till now .This leads us to reject this model .

This analysis allows us to conclude that ARIMA(0,1,1)$\times(1,1,0)$ model is the best possible model to fit to this data .

	Estimate	SE	t.value	p.value
ma1	-0.6796	0.0969	-7.0104	0.0000
sar1	-0.3220	0.1124	-2.8641	0.0054

We can observe that the p values of both coefficients are very small ie; less than 0.05 . This tells us that the coefficients are significant . We have already seen that this model has the minimum AIC and BIC among all the candidates . Also , there are no spikes in the ACF of the residuals , and the Ljung Box test gives high p values . So , we will now fit this model to the data .We include the residual analysis once more :

Residual analysis of final model



Let X_t denote the earnings of the Johnson and Johnson company .

Take $Y_t = log(X_t)$.

The Multiplicative Seasonal ARIMA model ARIMA(p, d, q) × (P, D, Q)s is :

$$\Phi_P(B^s)\phi_p(B)\square_s^D \square^d Y_t = \Theta_Q(B^s)\theta_q(B)Z_t$$

Where:

$$\Phi_P(B^s) = 1 - \Phi_1 B^s - \ldots - \Phi_P B^{Ps}$$
$$\Theta_Q(B^s) = 1 + \Theta_1 B^s + \ldots + \Theta_Q B^{Qs}$$
$$\square_s^D = (1 - B^s)^D$$

$$\phi_p(B) = 1 - \phi_1 B - \ldots - \phi_p B^p$$
$$\theta_q(B) = 1 + \theta_1 B + \ldots + \theta_1 B^q$$
$$\square^d = (1 - B)^d$$

The first three polynomials correspond to the seasonal AR , seasonal MA and seasonal difference .

The next three polynomials correspond to the non-seasonal AR , MA and difference .

And Z_t refers to the errors .

ARIMA(0, 1, 1) × (1, 1, 0)4 model is :

$$\Phi_1(B^4)\phi_0(B)\square_4^1\square^1 Y_t = \Theta_0(B^4)\theta_1(B)Z_t$$

Simplifying this gives :

$$(1 - \Phi_1 B^4)(1 - B^4)(1 - B)Y_t = (1 + \theta_1 B)Z_t$$

Even further simplification yields :

$$Y_t = Y_{t-1} + (1+\Phi_1)Y_{t-4} - (1+\Phi_1)Y_{t-5} - \Phi_1 Y_{t-8} + \Phi_1 Y_{t-9} + Z_t + \theta_1 Z_{t-1}$$

Plugging in the values of θ_1 and Φ_1 from :

	Estimate	SE	t.value	p.value
ma1	-0.6796	0.0969	-7.0104	0.0000
sar1	-0.3220	0.1124	-2.8641	0.0054

We get : θ_1 = -0.6796 , Φ_1 = -0.3220

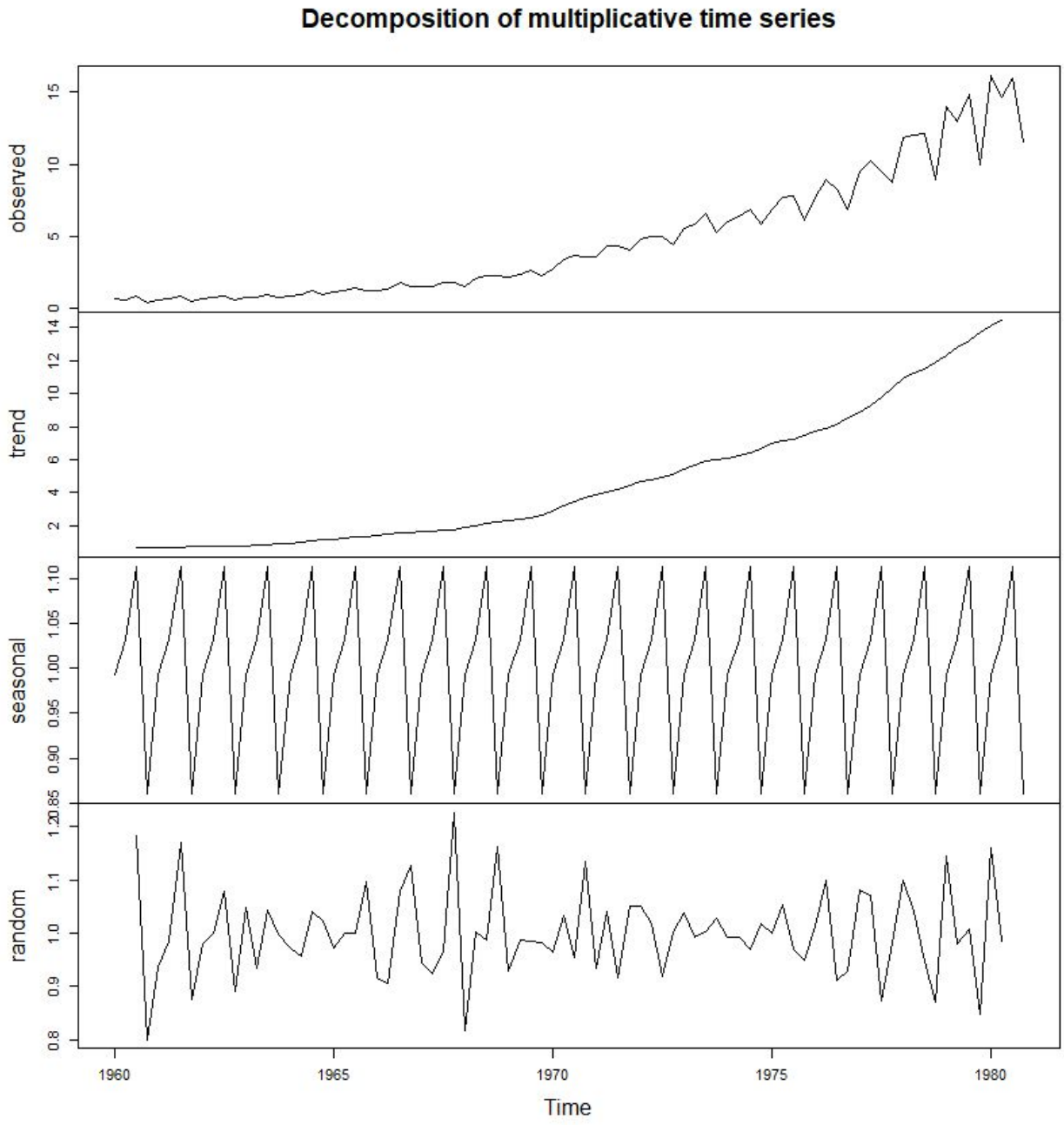
$$Y_t = Y_{t-1} + 0.678Y_{t-4} - 0.678Y_{t-5} + 0.3220Y_{t-8} - 0.3220Y_{t-9} + Z_t - 0.6796\,Z_{t-1}$$

$Y_t = log(X_t)$ and Z_t ~ N(0,0.0079)

We have thus successfully fitted an ARIMA model to the data .

We try to decompose the model and fit an ARIMA model to the residuals .The seasonal variation increases as we move across time. A multiplicative decomposition could be useful .

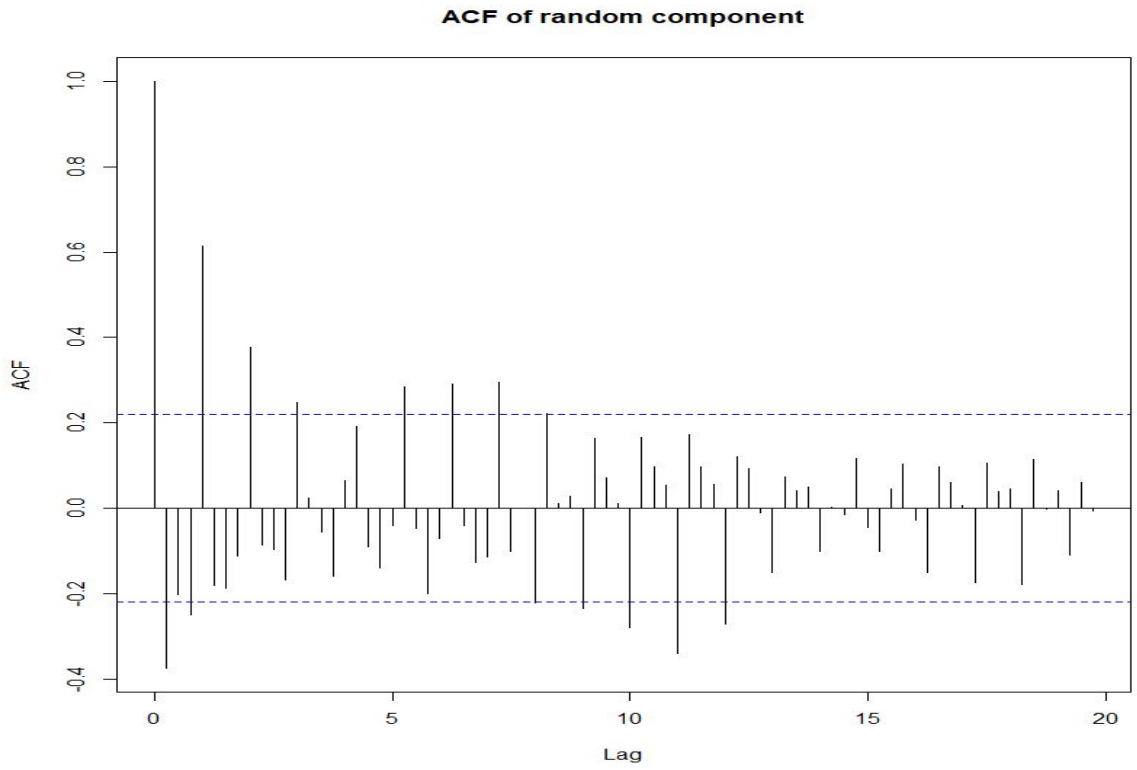
We use the inbuilt decompose() function of R to try a multiplicative decomposition of this model .

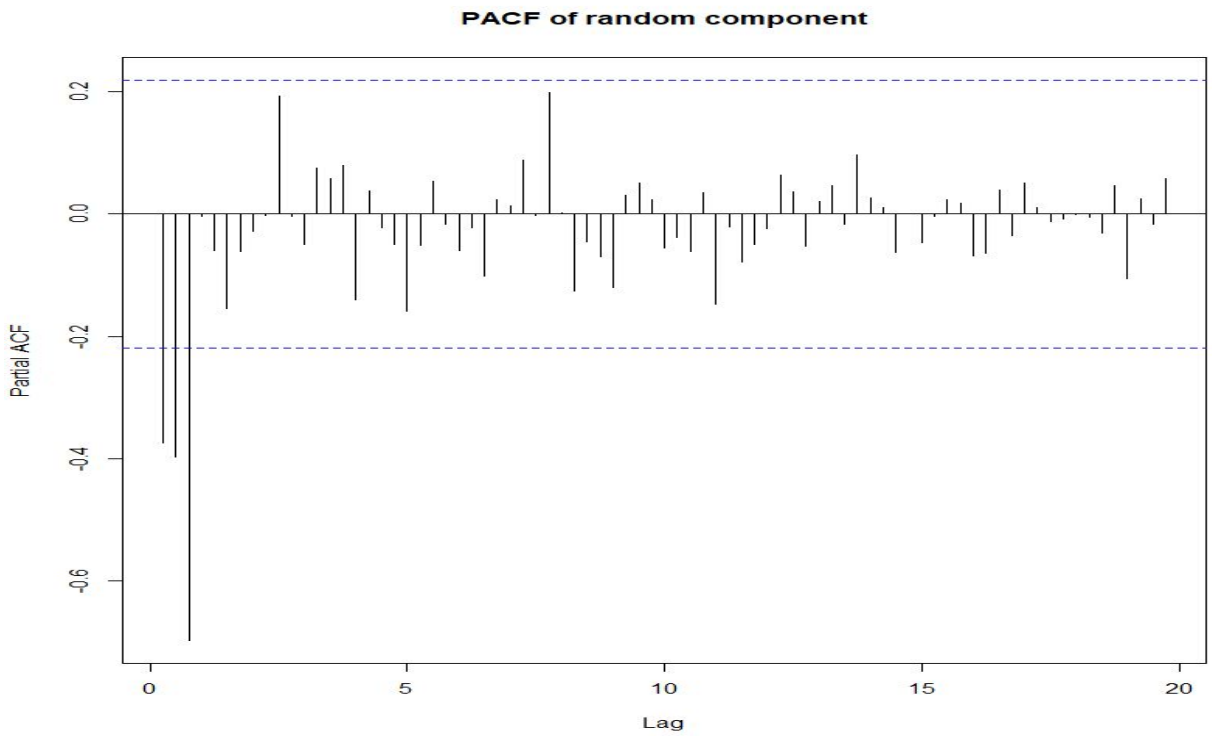


We observe that the random component does not have any visible trend or seasonality .

Let us try to fit arima model to this random component .

It can be considered a stationary time series . Let us look at the ACF and PACF plots :

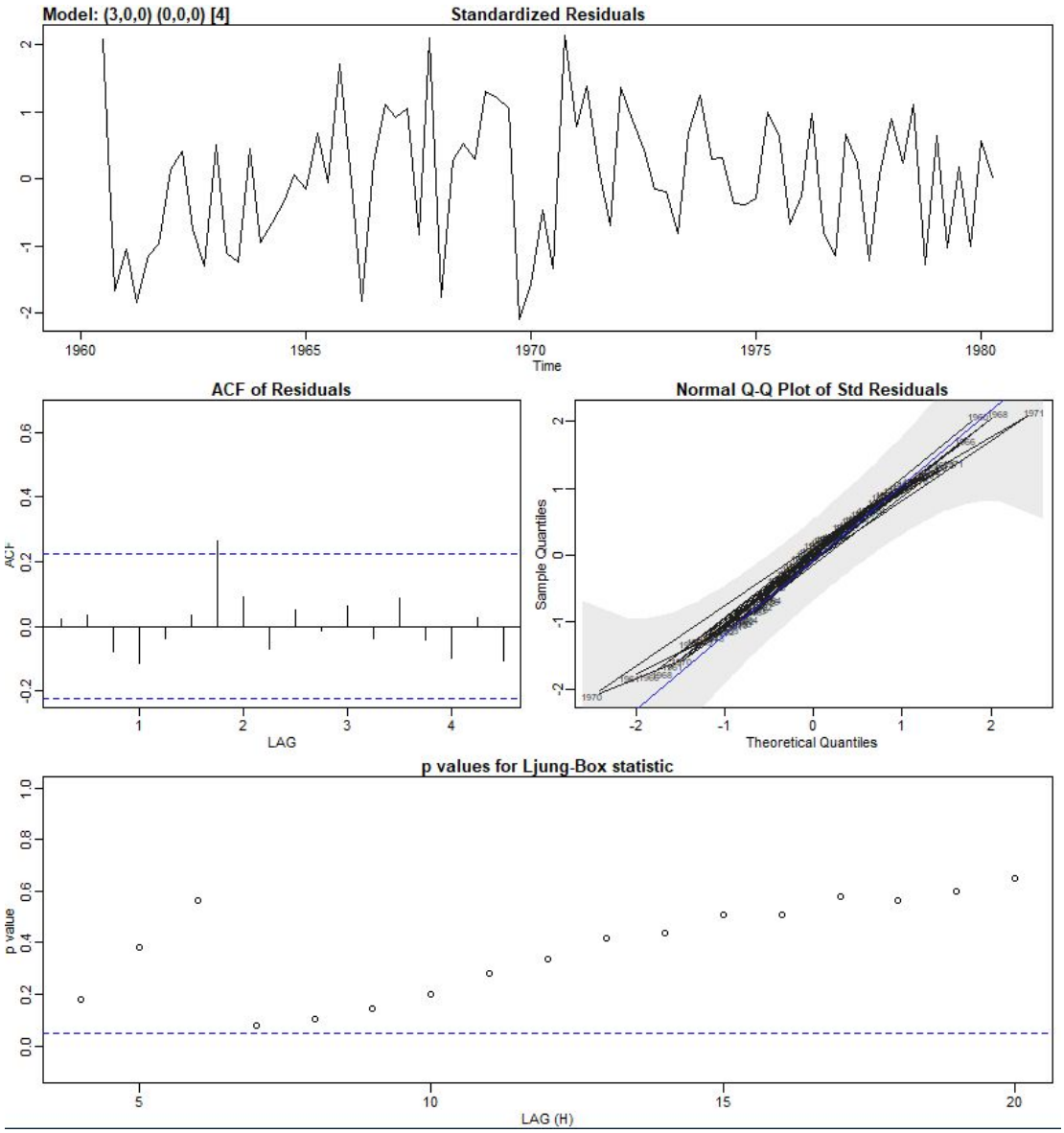




Now , the ACF looks quite complicated , there are many significant spikes in the plot.

But the PACF cuts off after lag 3 . This suggests that we should try to fit an AR(3) model .

The residual analysis after fitting the random component with ar(3) model is shown below :



We observe that there is only one spike in the acf of the residuals . The Ljung box statistic values are good , and the QQ plot is also not bad . This means that AR(3) is a good fit to the model .

	Estimate	SE	t.value	p.value
ar1	-0.8252	0.0725	-11.3777	0
ar2	-0.8163	0.0773	-10.5595	0
ar3	-0.7941	0.0738	-10.7630	0
xmean	0.9989	0.0015	660.1307	0

The p-values indicate the significance of these terms . This means that we can fit an AR(3) model to the residuals obtained from multiplicative decomposition . The AIC = -253.75 which is smaller than what we get after trying out some possible alternate models like arma(3,1) or arma(3,2) .

If r_t denotes the residuals , then we can write the AR(3) equation as follows :

$$(1+0.8252B+0.8163B^2+0.7941B^3)(r_t - 0.9989) = w_t$$

where w_t ~N(0,0.002073)

Simplifying this yields :

$$r_t = 3.4318 - 0.8252\, r_{t-1} - 0.8163\, r_{t-2} - 0.7941\, r_{t-3} + w_t$$

This is the arima model fitted to the random part of the decomposition performed by R .