Name: Rochita Das

Registration No: A01-2112-0735-11

Roll No: 91/STS/141003

Time series and Regression analysis of Yahoo share price of Japan and USA

Objectives:

I. Firstly, **time series analysis**, separately on both yahoo share price data of USA and Japan.

II. Secondly, **regression analysis**, considering some relevant exogenous variables and also considering mutual interaction between the two share prices.

Key words: ARIMA, GARCH, VAR, Causality, 2SLS

Data collection:

Data are collected on the following variables:

1. Yahoo share price of USA - Y_1 (endogenous variable)
2. Yahoo share price of Japan - Y_2 (endogenous variable)
3. NASDAQ (an American stock exchange index) - X_1 (exogenous variable)
4. NIKKEI (Tokyo stock exchange index) - X_2 (exogenous variable)
5. Yen – Dollar exchange ratio - X_3 (exogenous variable)

For variables $\boldsymbol{Y_1}$,$\boldsymbol{Y_2}$,$\boldsymbol{X_1}$, $\boldsymbol{X_2}$ the market **closing values** are considered .

- Y_1, Y_2 are considered as **endogenous** variables as they affect each other.
- X_1, X_2, X_3 are considered as **exogenous** variable. For this reason their values are considered at one lag before than endogenous variables.

✓ Data on Y_1, Y_2 are collected from 20/01/2010 to 26/01/2016.
✓ Data on X_1, X_2, X_3 are collected from 19/01/2010 to 25/01/2016.
✓ For prediction purpose, data on each variable are collected for 10 days after the last time point.

➢ Data are collected from the following sites:

finance.yahoo.com

https://research.stlouisfed.org/fred2/series/EXJPUS

A snapshot of data set is,

Date	usa	japan	Date_1	nasdaq	nikkei	exc ratio	time
10-02-15	43.07	7.12	09-02-15	4726.01	17335.85	118.84	1274
11-02-15	42.96	7.13	10-02-15	4787.64	17678.74	118.68	1275
12-02-15	43.93	7.15	11-02-15	4801.18	17504.62	118.88	1276
13-02-15	44.42	7.27	12-02-15	4857.61	17648.5	119.13	1277
17-02-15	43.53	7.27	13-02-15	4893.84	17711.93	118.88	1278
18-02-15	43.65	7.37	17-02-15	4899.27	17652.68	119.36	1279
19-02-15	44.37	7.35	18-02-15	4906.36	17652.68	119.72	1280
20-02-15	44.11	7.27	19-02-15	4924.7	17979.72	120.06	1281
23-02-15	43.53	7.26	20-02-15	4955.97	17913.36	119.47	1282
24-02-15	43.38	7.55	23-02-15	4960.97	18004.77	119.76	1283
25-02-15	44.43	7.73	24-02-15	4968.12	17987.09	120.22	1284
26-02-15	44.45	8	25-02-15	4967.14	18199.17	120.93	1285
27-02-15	44.28	8.04	26-02-15	4987.89	18264.79	121.17	1286
02-03-15	44.11	8.01	27-02-15	4963.53	18332.3	121.2	1287
03-03-15	42.62	7.92	02-03-15	5008.1	18466.92	121.5	1288
04-03-15	43.99	7.86	03-03-15	4979.9	18603.48	121.28	1289
05-03-15	44.16	8.1	04-03-15	4967.14	18585.2	121.17	1290
06-03-15	43.44	7.7	05-03-15	4982.81	18785.79	121.3	1291

(I). Time Series Analysis

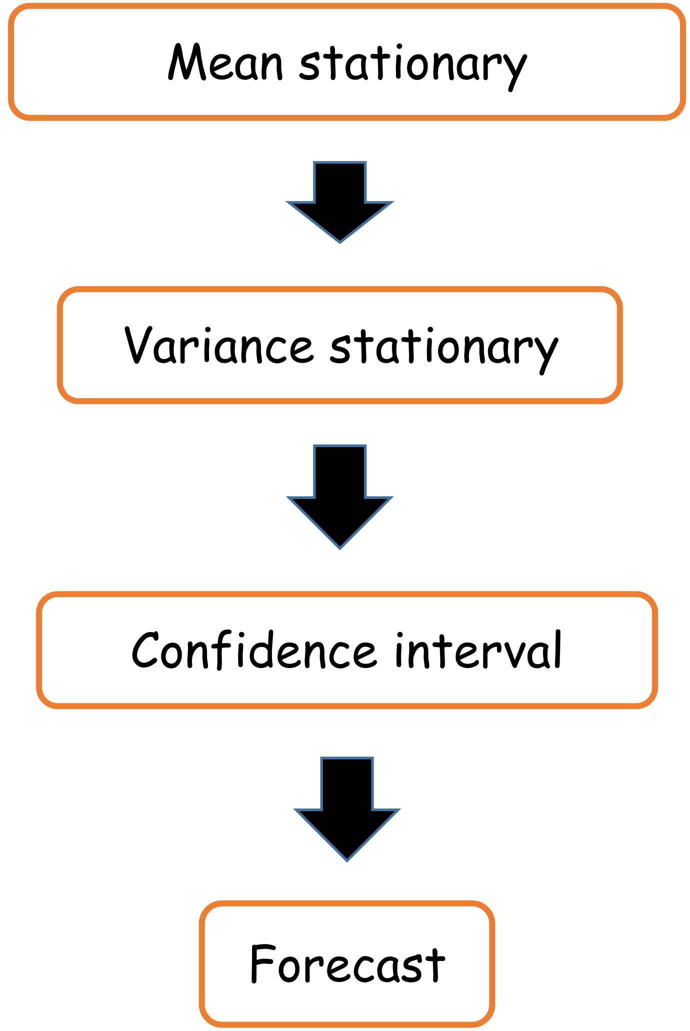
General approach:

Additive model on log transformed data

Removal of trend

Removal of seasonality

Stationary data


Mean stationary
Variance stationary
Confidence interval
Forecast

Different statistical tests, which are used in this analysis:

1. **Mann-Kendall Test** for Monotonic Trend

   H_0: No monotonic trend vs H_1: Monotonic trend is present

2. **Kruskal-Wallis test by ranks**

   (non-parametric One-way ANOVA on ranks)

   H_0 :medians of all groups are identical (seasonality is absent)
   vs H_1 : at least one population median of one group is different from the population median of at least one other group.

3. **Ljung-Box test**

   H_0: Data are independently distributed (autocorrelation in population is 0)
   H_1 : Data are not independently distributed; they exhibit serial correlation.

4. **Lagrange multiplier test**

H_0 : No heteroscedasticity is present
vs H_1 : Heteroscedasticity is present

5. **Anscombe test** on kurtosis

H_0 : Kurtosis = 3 vs H_1 : Kurtosis > 3

6. **Granger causality test**

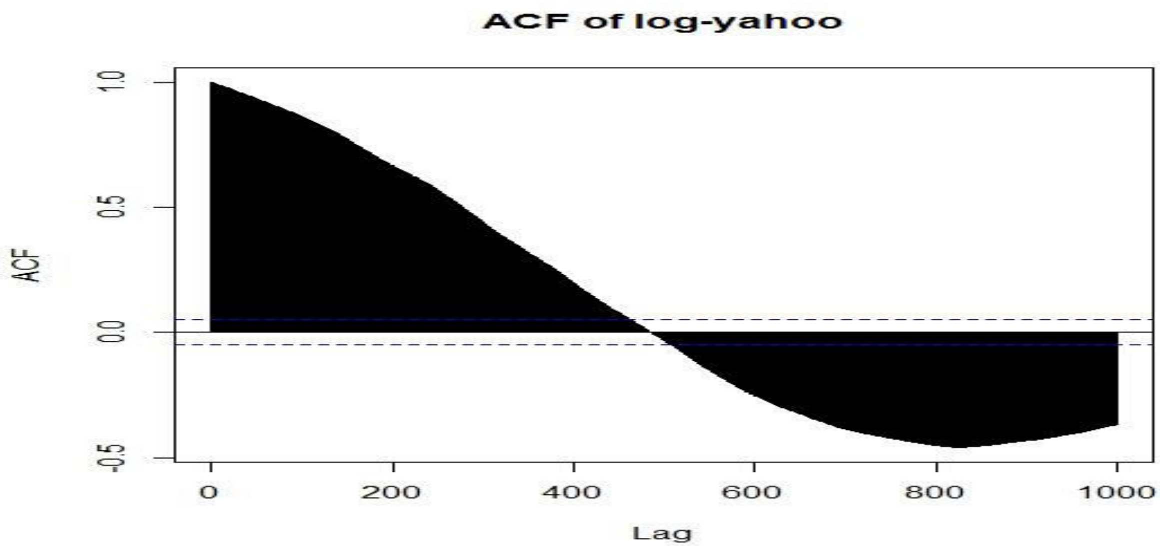
H_{0A} : one time series is ***not*** causal in forecasting another time series.
vs H_{1A} : one time series is causal in forecasting another time series.

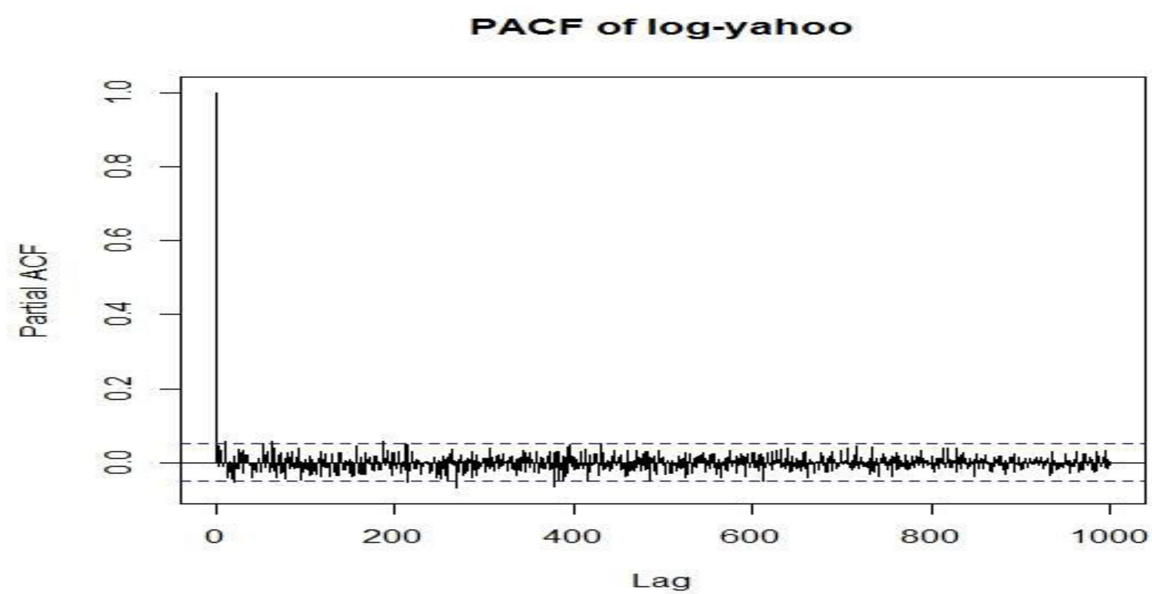
H_{0B} : two time series are ***not*** instantaneous causal to each other.
vs H_{1B} : two time series are instantaneous causal to each other.

(A). Time Series Analysis of Yahoo Share price of USA

- ✓ plot shows a very prominent **increasing trend**.

ACF, **PACF** plots of log transformed data are ,

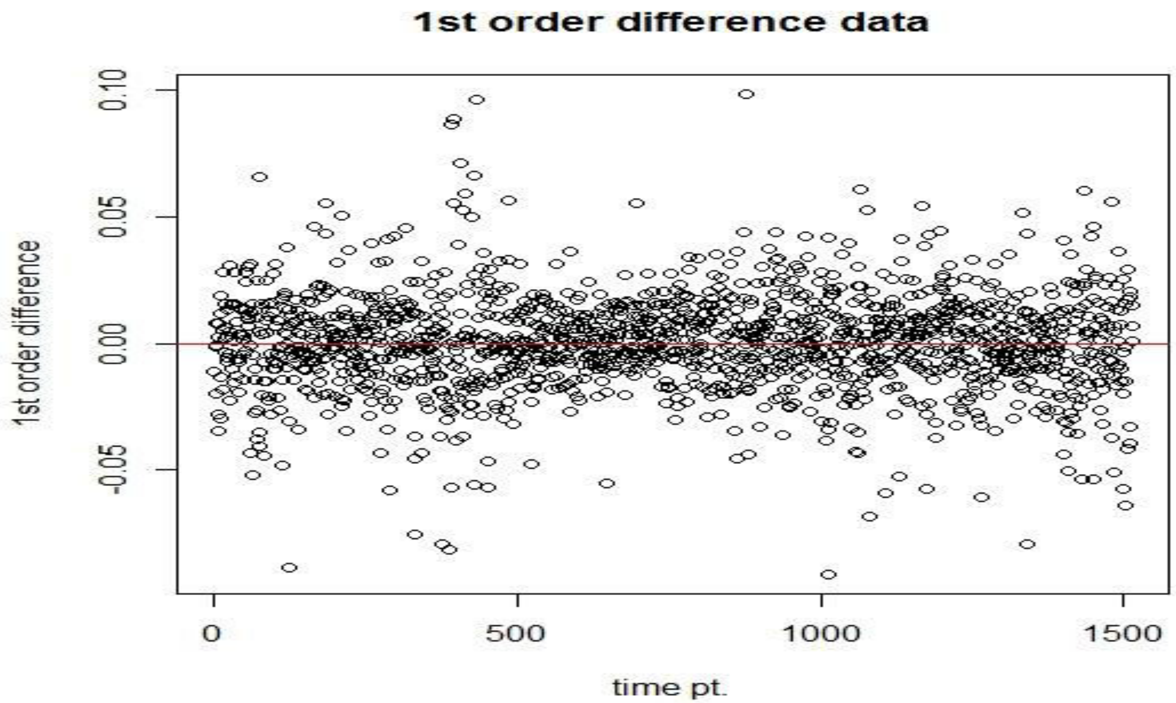




✓ The **Mann kendal test** : ***pvalue =< 2.22e-16***

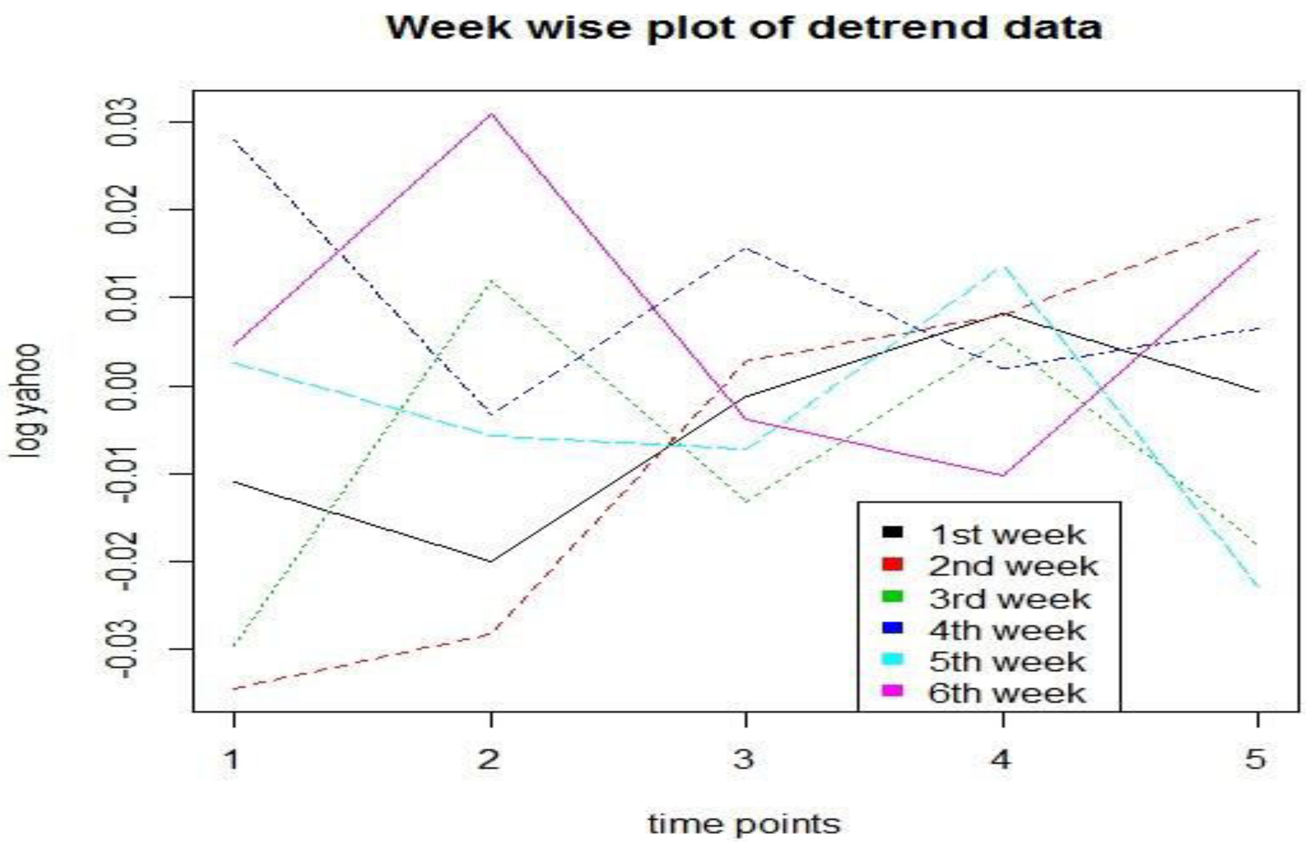
➡ presence of significant trend

1^{st} order difference plot is as follow,

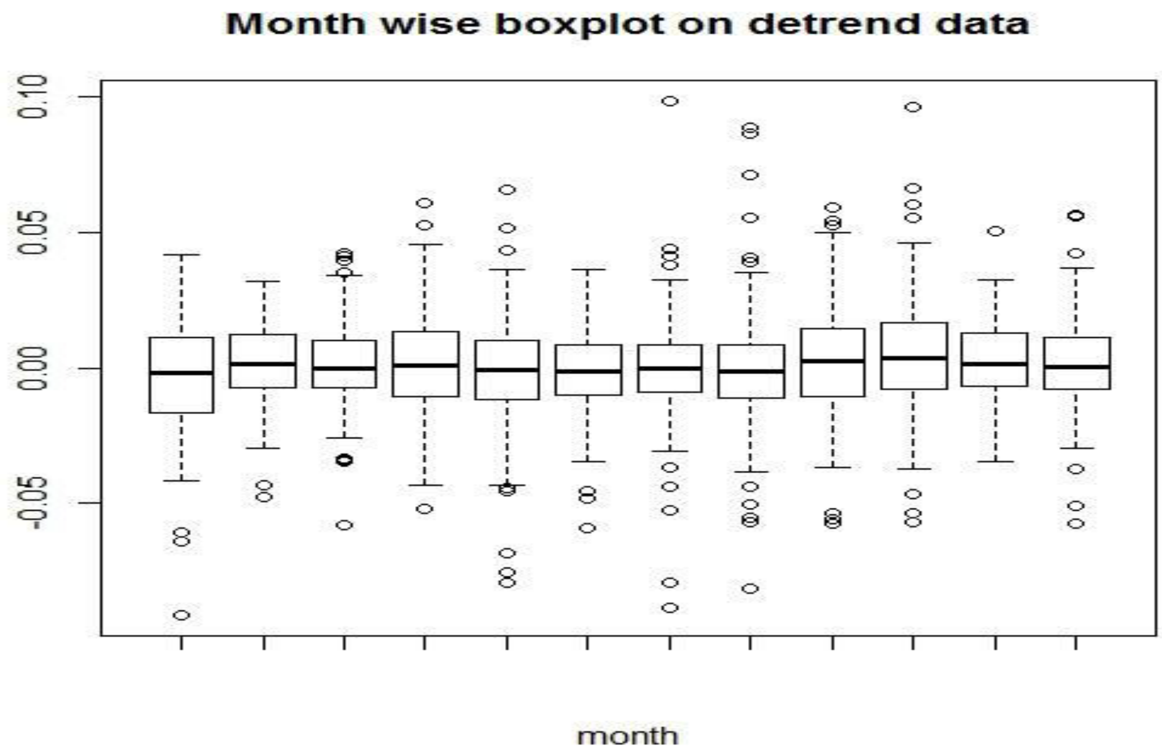


- ✓ The plot shows a random fluctuation about The line Y=0
- ✓ **Mann kendal** test on 1^{st} order difference gives ***p-value =0.83552*** confirming absence of significant trend

- Absence of **weekly seasonality**,



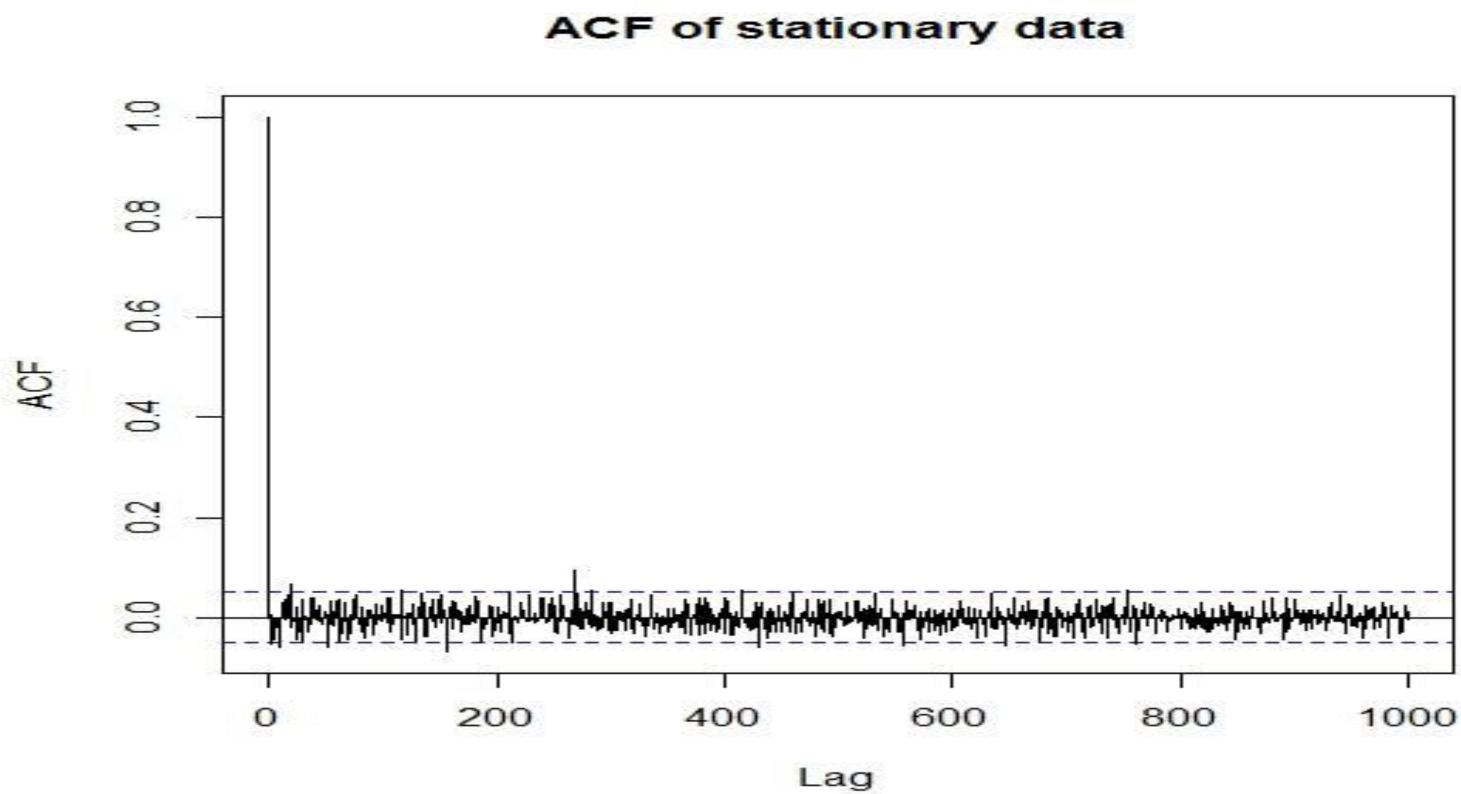
❖ Absence of **monthly seasonality**

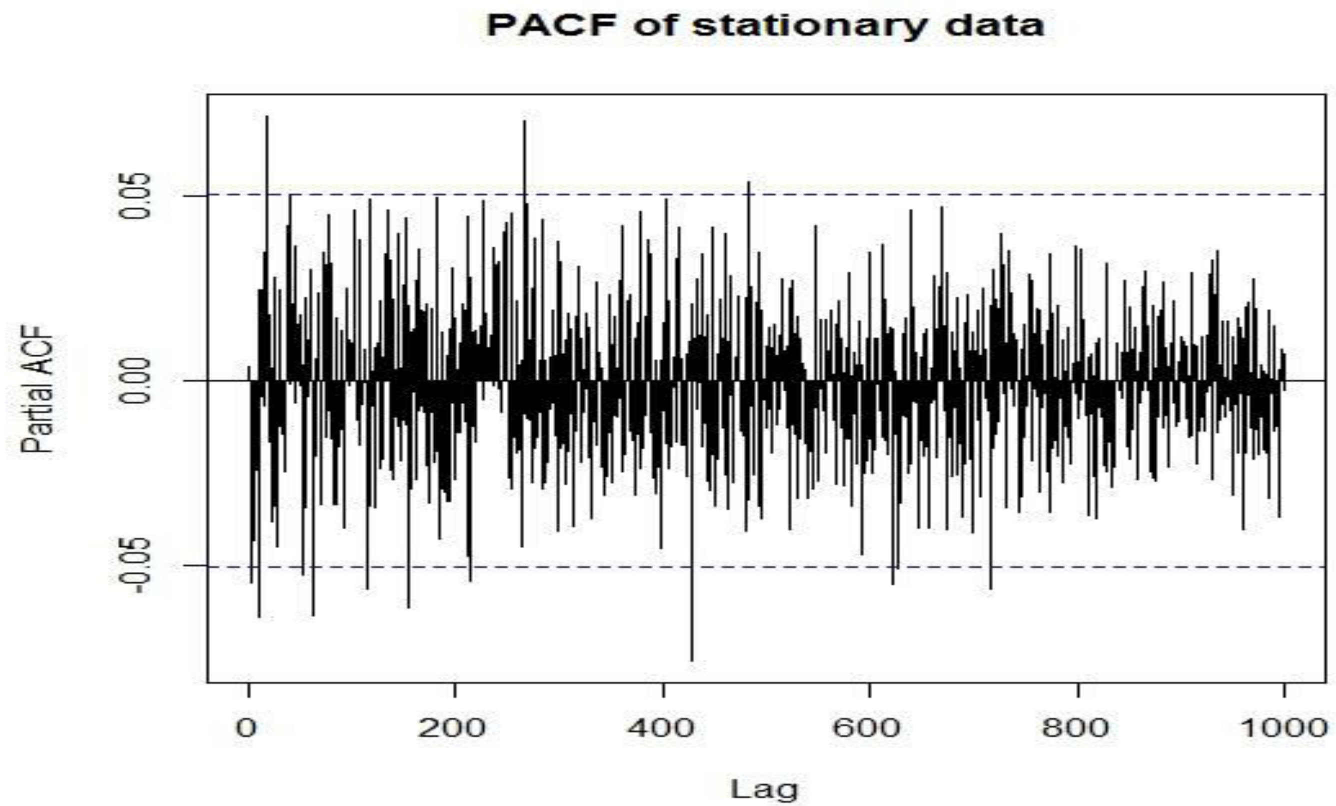


✓ **Kruskal-Wallis** test , ***p-value = 0.2003*** also supports it.

Hence the detrended data is considered as **stationary** data.

The ***ACF***, ***PACF*** plots of stationary data are as follows





- ACF tails off and PACF shows few significant spikes, hence AR(p) model with a large p value is appropriate.
- To keep the model parsimonious we opt for ARMA (p,q)model.

The model **ARIMA(1,1,3)** is chosen by **minimum AIC criterion.**

Yahoo share price of USA Y_{1t} at t^{th} time point,

$$\boldsymbol{Y_{1t} - Y_{1\overline{t-1}} = 0.6422 * (Y_{1\overline{t-1}} - Y_{1\overline{t-2}}) + \varepsilon_t - 0.6426 * \varepsilon_{t-1} - 0.0041 * \varepsilon_{t-2} - 0.0562 * \varepsilon_{t-3}}$$

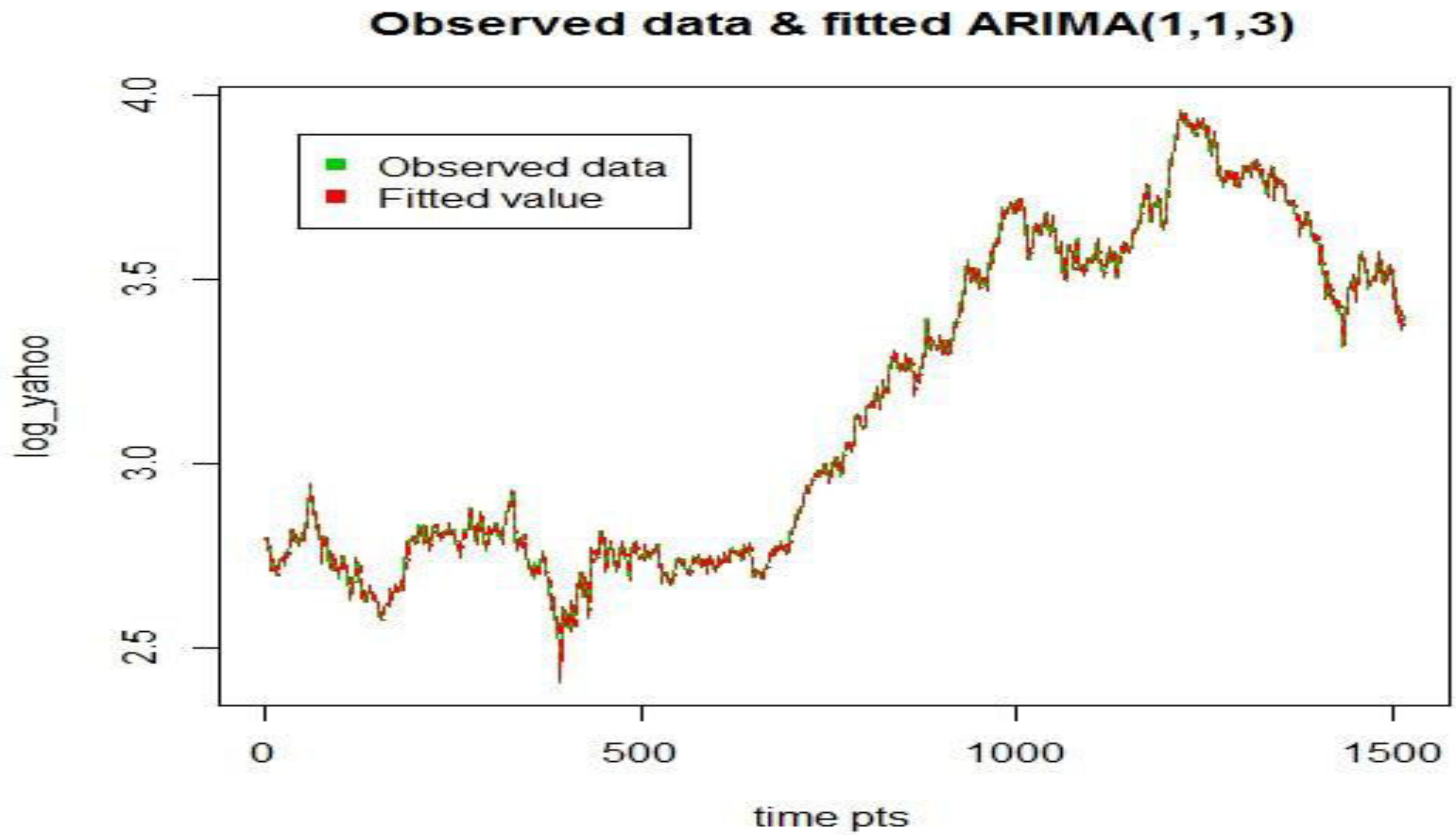
With aic = -7633.88

Ljung-Box test on residuals of ARIMA(1,1,3) is as follow,

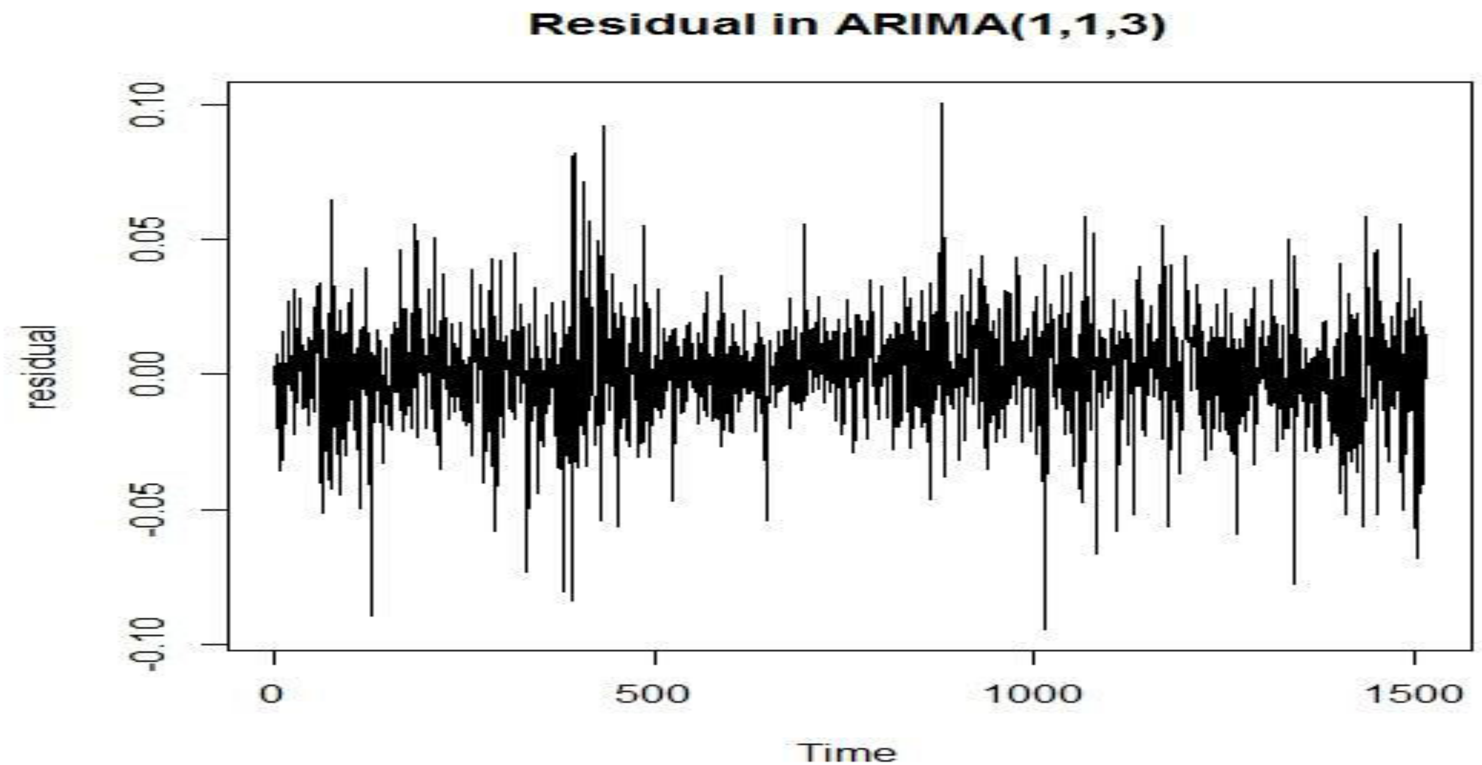
Lags	*Statistic*	*df*	*p-value*
5	0.3959217	5	0.9990010
10	5.4452888	10	0.8611389
15	11.1606360	15	0.7422577

The test says ARIMA residuals are purely random. Hence a good fit.

The following plot shows fitted value with original value .

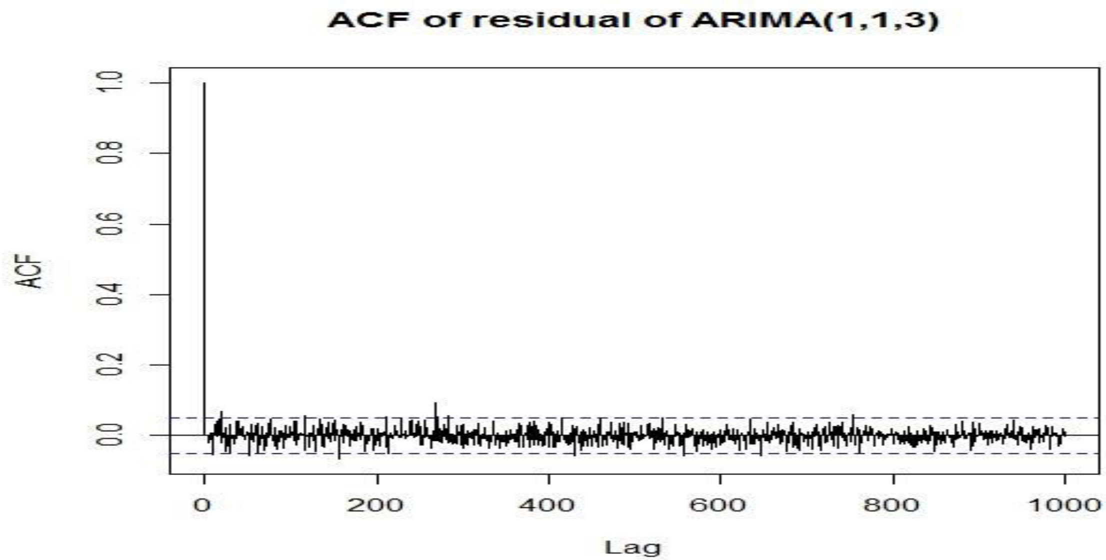


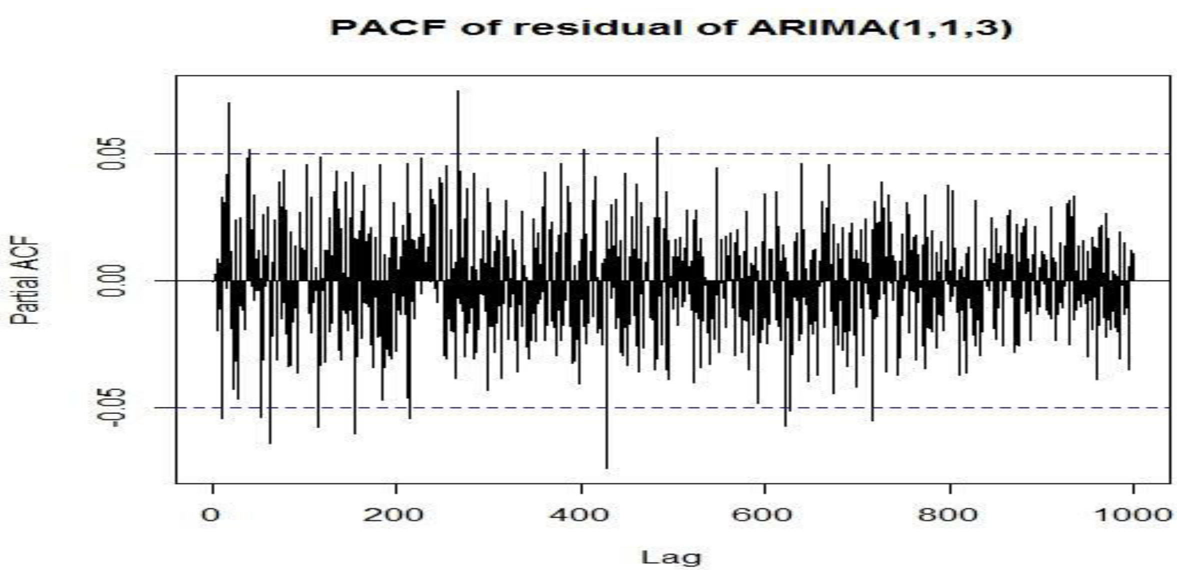
The Residual plot is as follows ,



- ✓ Large fluctuations followed by small fluctuations and again followed by large and this continues
- ✓ The **Lagrange multiplier test** for ARCH(1) gives ***p-value = 7.818e-05,*** which confirms at least ARCH(1) is required to model the volatility present in the ARIMA residual .

ACF,PACF plots of residual are as follow,





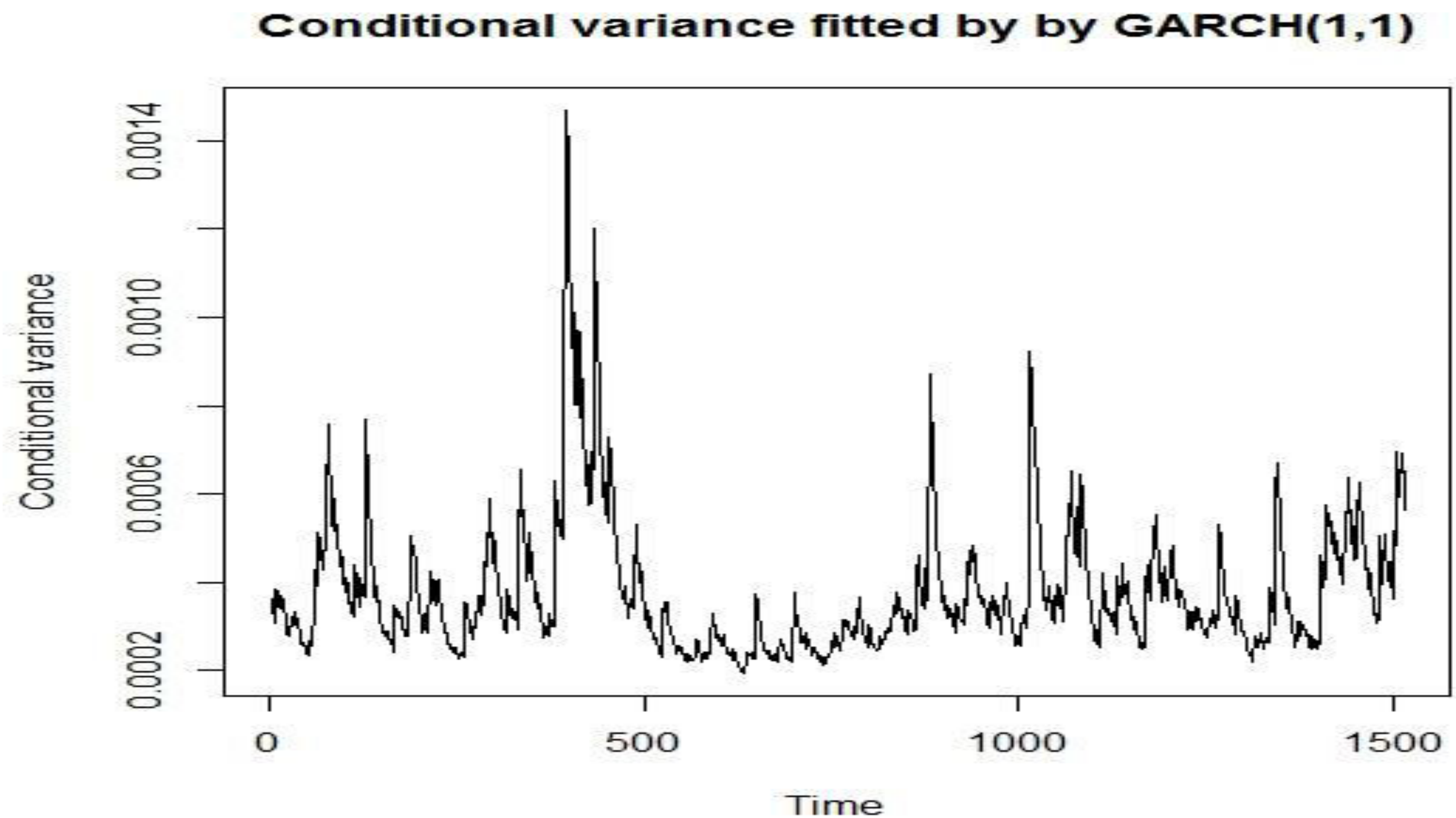
by minimum AIC criterion **GARCH(1,1)** is chosen and it is ,

ARIMA residual, $\boldsymbol{r_t = 7.213 * 10^{-4} + \varepsilon_t * \sqrt{\widehat{h_t}}}$

Where $\boldsymbol{\widehat{h_t} = 1.426 * 10^{-5} + 0.05264 * r_{t-1}^2 + 0.91 * h_{t-1}}$

- ✓ **Ljung-Box Test** on GARCH(1,1) residual gives ***p-value = 0.9164*** which says the model to be adqequate to capture volatility. The test result.

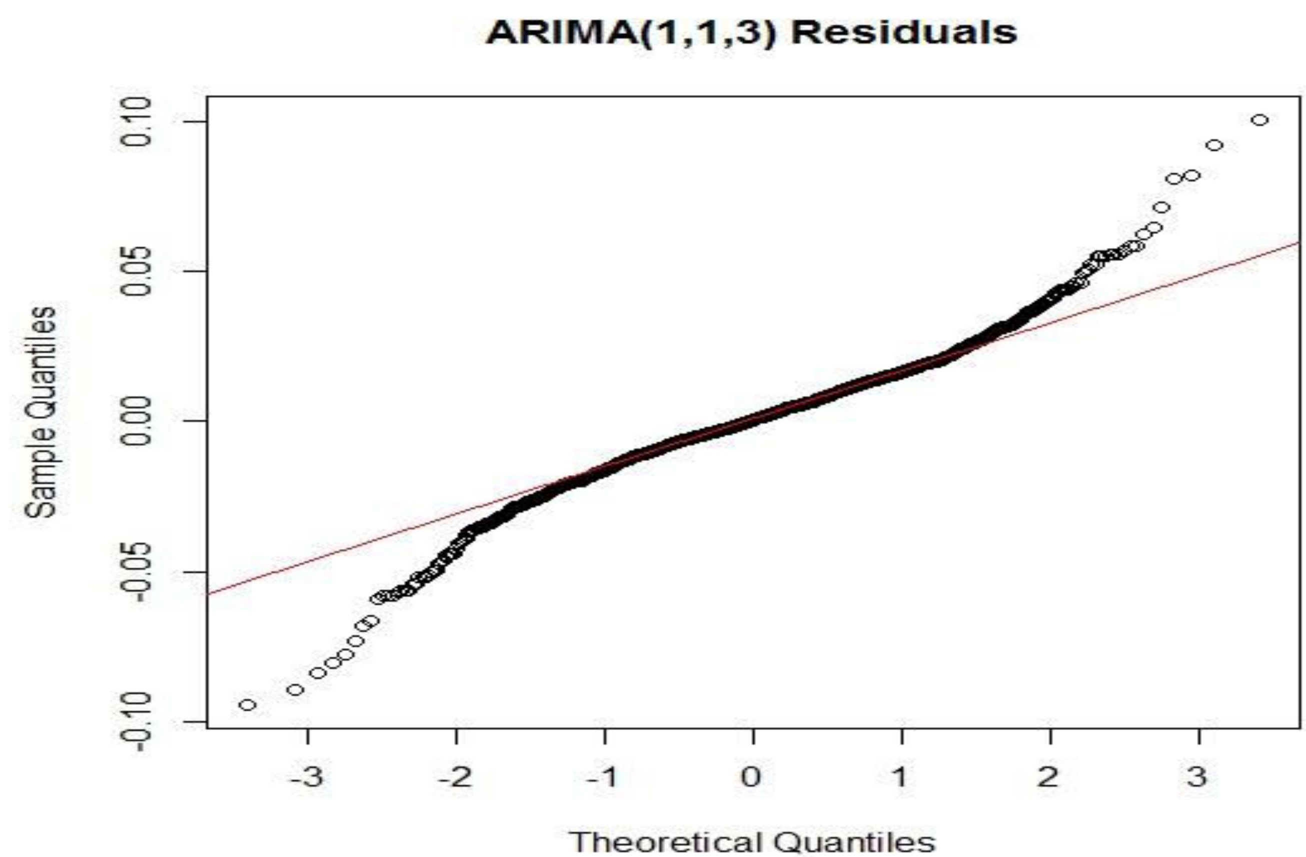
The plot of **conditional variance** i.e., $\widehat{h_t}$ is as follow,



Now to find the C.I. $dist^n$ of ARIMA residual is found first .

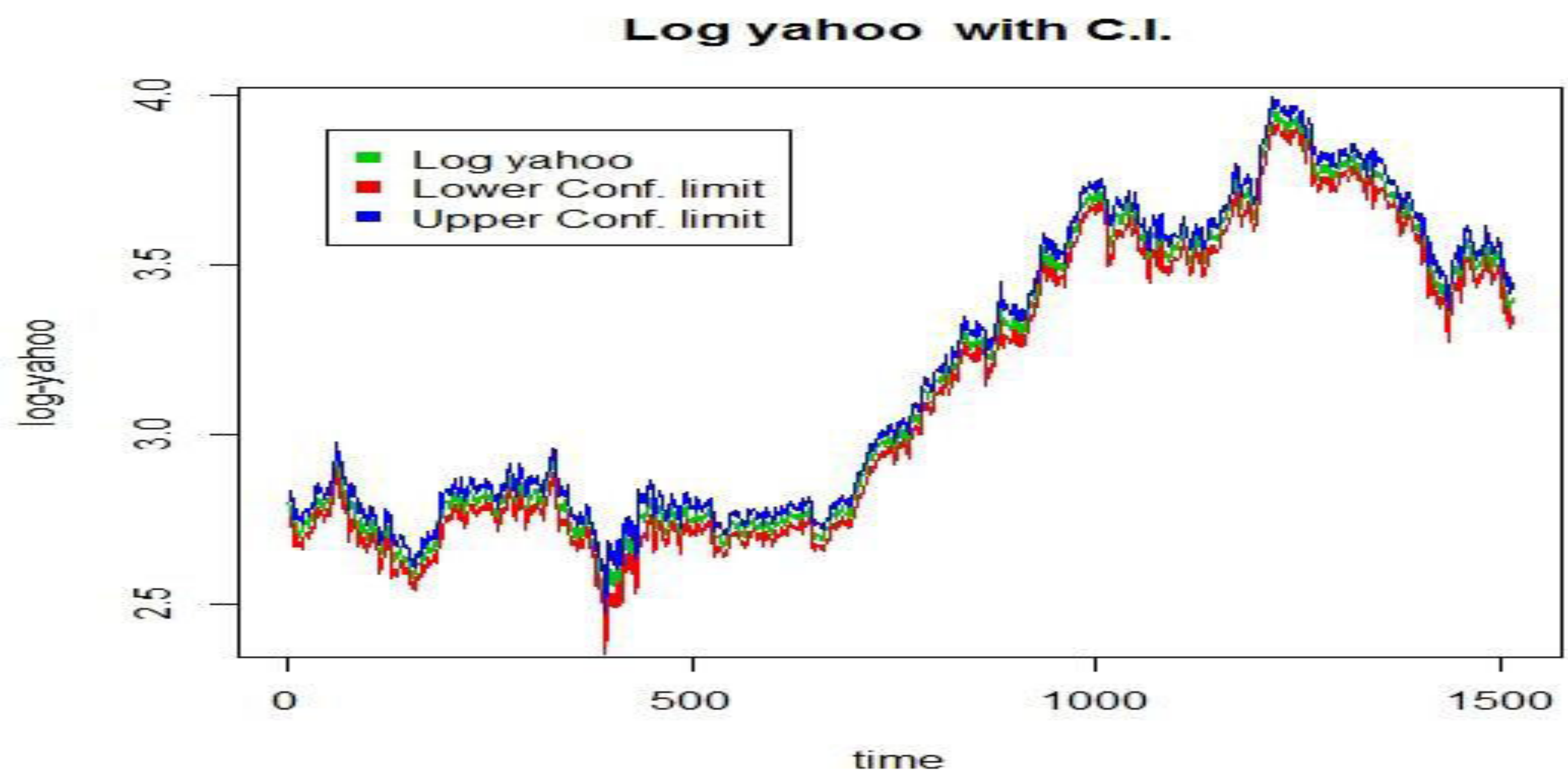
Kurtosis (r_t) =**5.867355** (>3)

- ✓ **Anscombe test** gives ***p-value < 2.2e-16*** which says kurtosis significantly differ from 3 .
- ✓ The **QQ-plot** also supports the fact obtained above



- Having **large frequency on tail** observations , we fit '**t(v)**' distribution.
- v is estimated from ARIMA residual as , v = 5 .

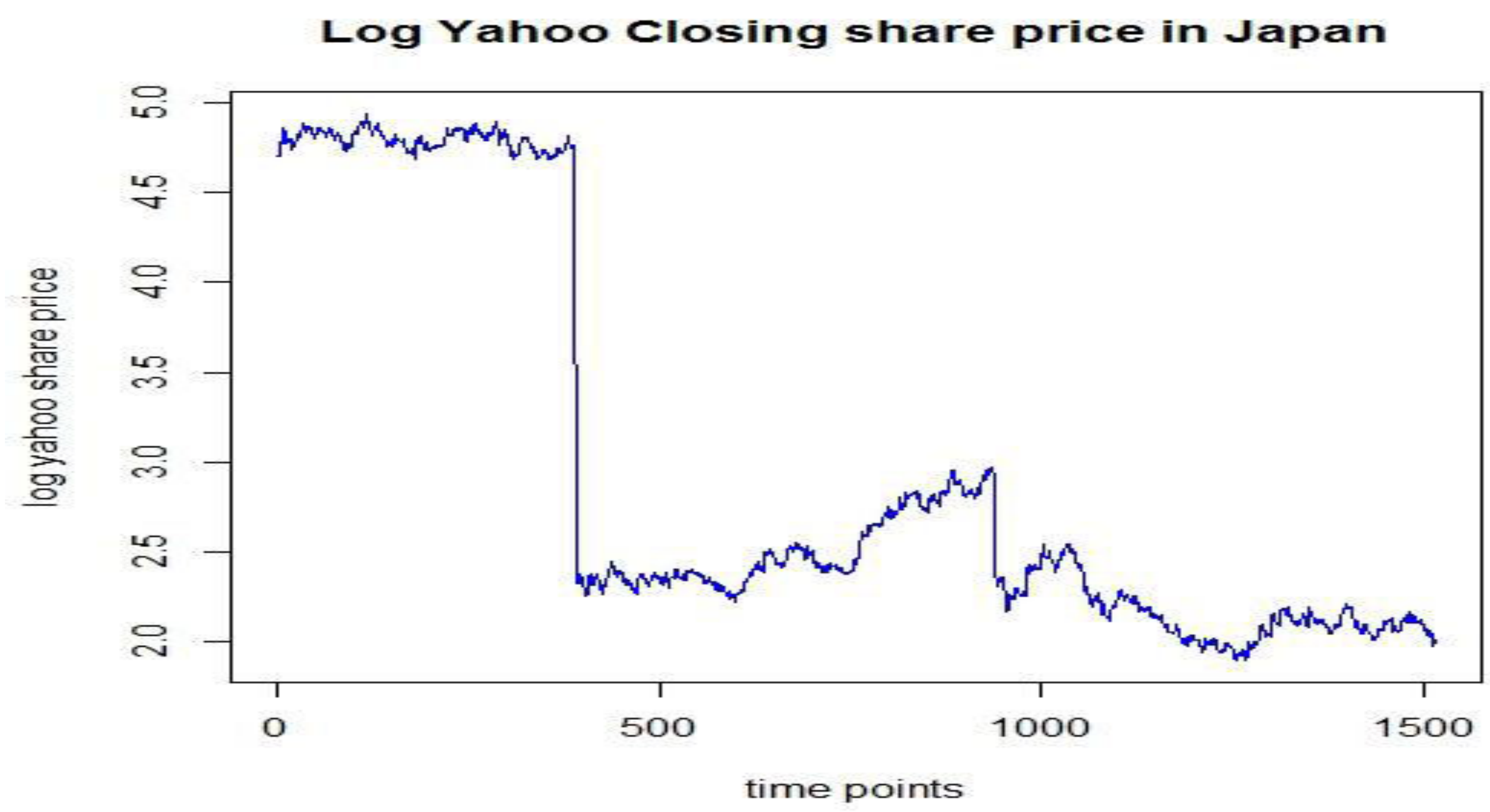
Now the original log transformed data with confidence interval is plotted as follow,



Forecast:

Time point	Observed value	Predicted value	95 % lower conf limit	95 % upper conf limit
27-01-16	29.69	29.96680	28.60014	31.39876
28-01-16	28.75	29.95979	28.60064	31.38352
29-01-16	29.51	29.94493	28.59342	31.36032
01-02-16	29.57	29.93538	28.59103	31.34293
02-02-16	29.06	29.92924	28.59169	31.32937
03-02-16	27.68	29.92532	28.59425	31.31835
04-02-16	29.15	29.92278	28.59792	31.30901
05-02-16	27.97	29.92116	28.60228	31.30086
08-02-16	27.05	29.92012	28.60698	31.29352
09-02-16	26.82	29.91946	28.61187	31.28680

(B). Time Series Analysis of Yahoo Share price of Japan



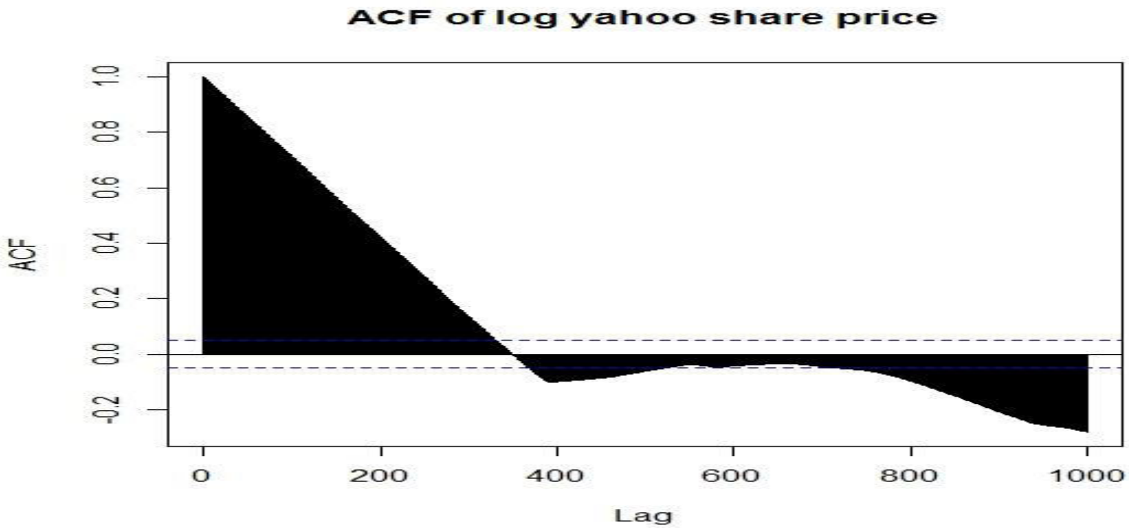
✓ The plot shows prominent **decreasing trend** with two sharp jumps.

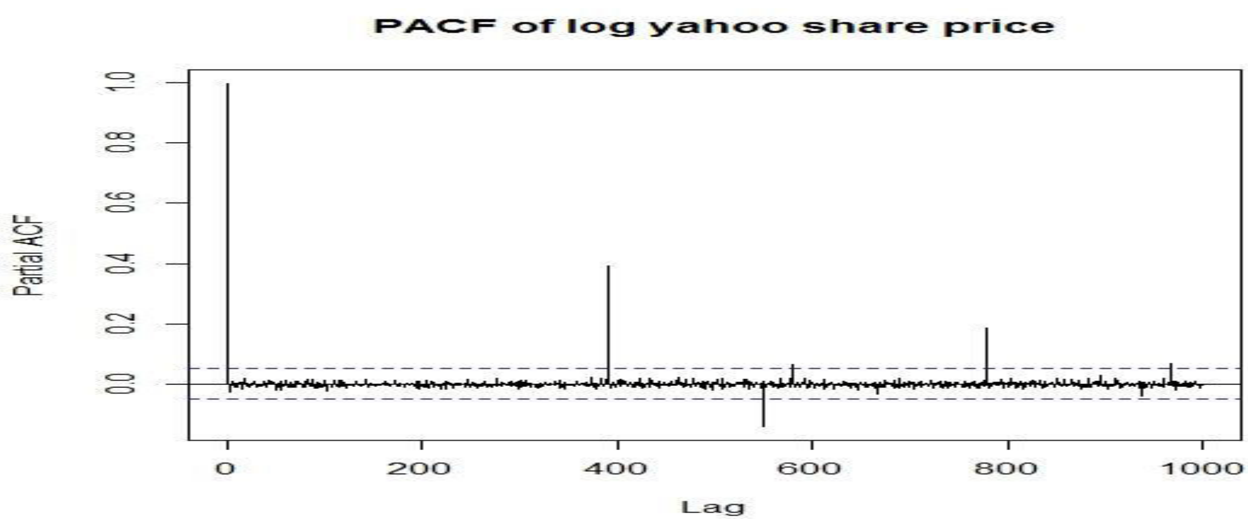
About the sharp jumps:

✓ The plot shows prominent decreasing trend with two sharp jumps at time points t=389 and t=937, i.e., at 03/08/2011 and 08/10/2013 respectively.

✓ The first sharp jump can be assumed to be caused by the **effect of earthquake and tsunami** that hit Japan in 2011.

ACF, **PACF** plots of log transformed data are,

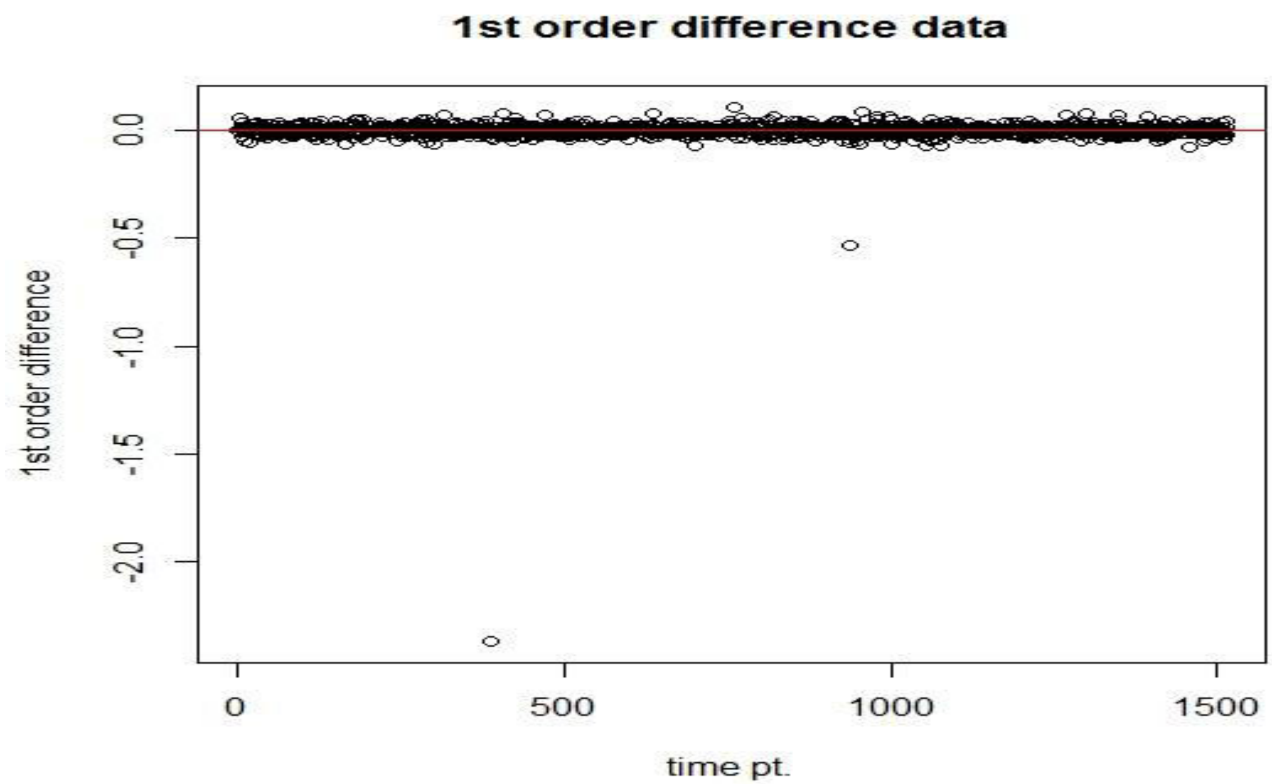




✓ **Mann kendal test** 2-sided ***p-value =< 2.22e-16***

➡ presence of significant trend.

The plot of $\mathbf{1^{st}}$ **order difference** is as follow,

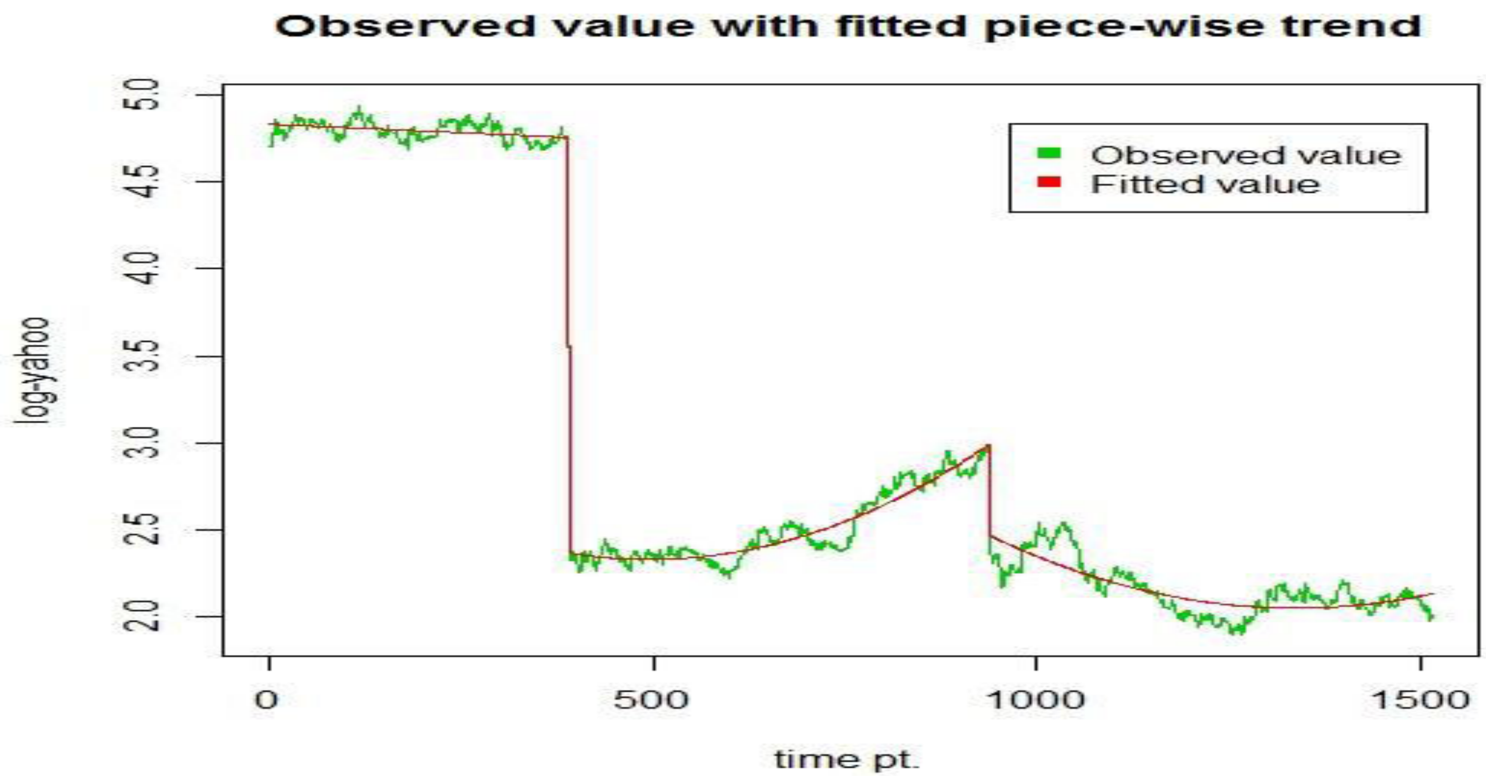


plot shows presence of **two very big outliers**, hence the method is discarded.

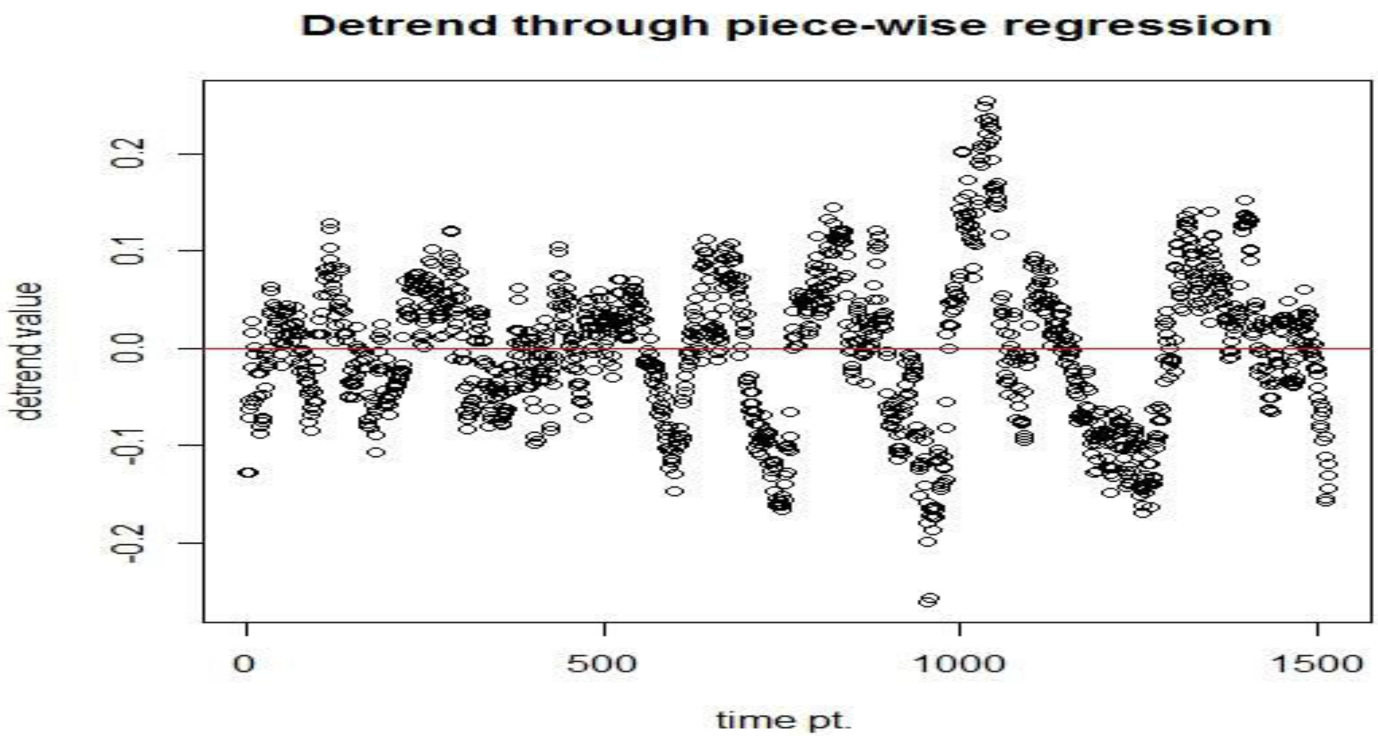
piece wise linear and quadratic **splines** are fitted. They are as follows,

Time point	***intercept***	***time***	***time^2***
1:389	4.83345	-0.0001942	----
390:937	3.154	$-3.337*10^{-3}$	$3.372*10^{-6}$
938:1515	6.744	$-7.028*10^{-3}$	$2.630*10^{-6}$

Plot of fitted trend with original log transformed data is,

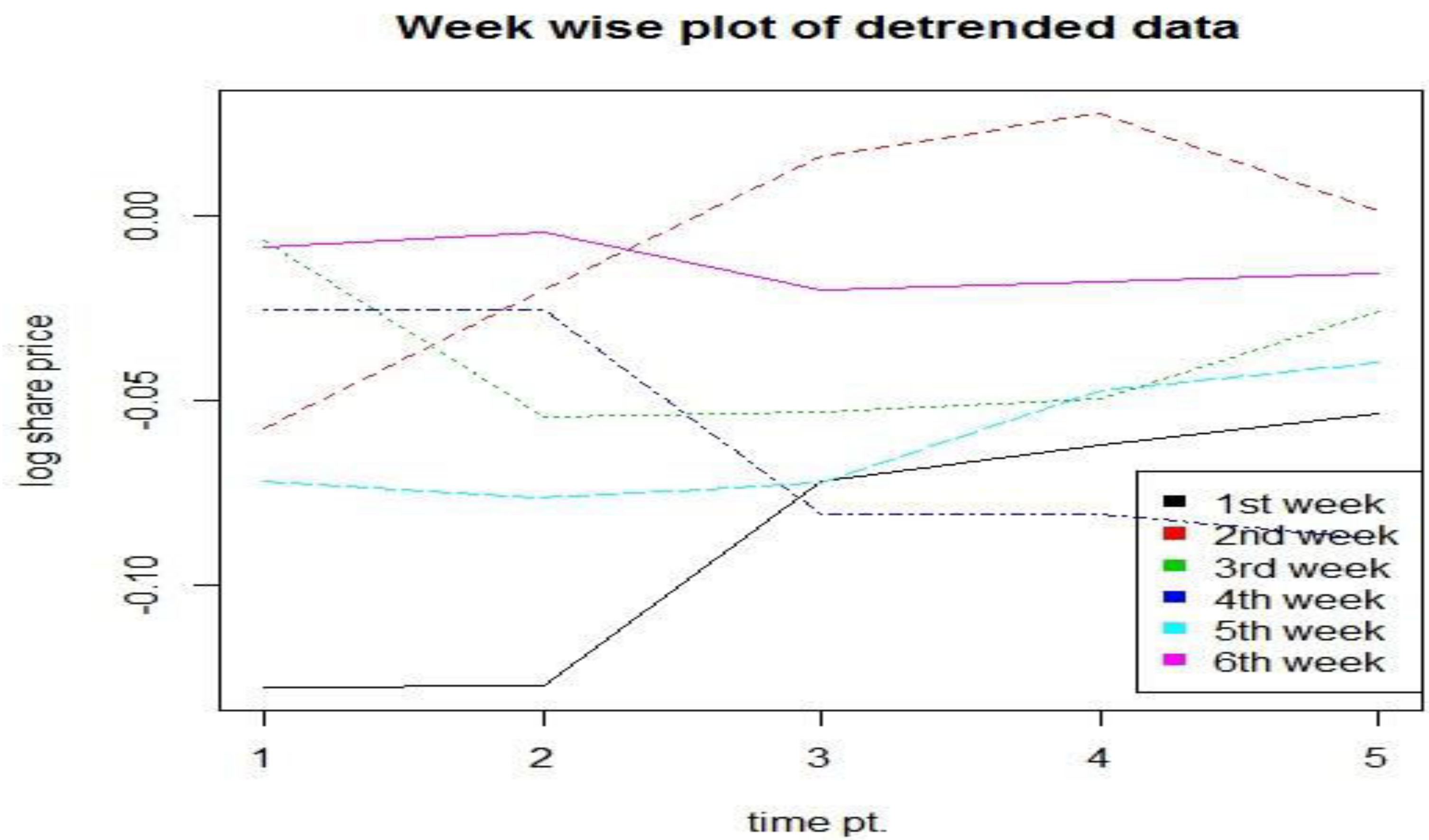


The plot of detrended data is shown below.

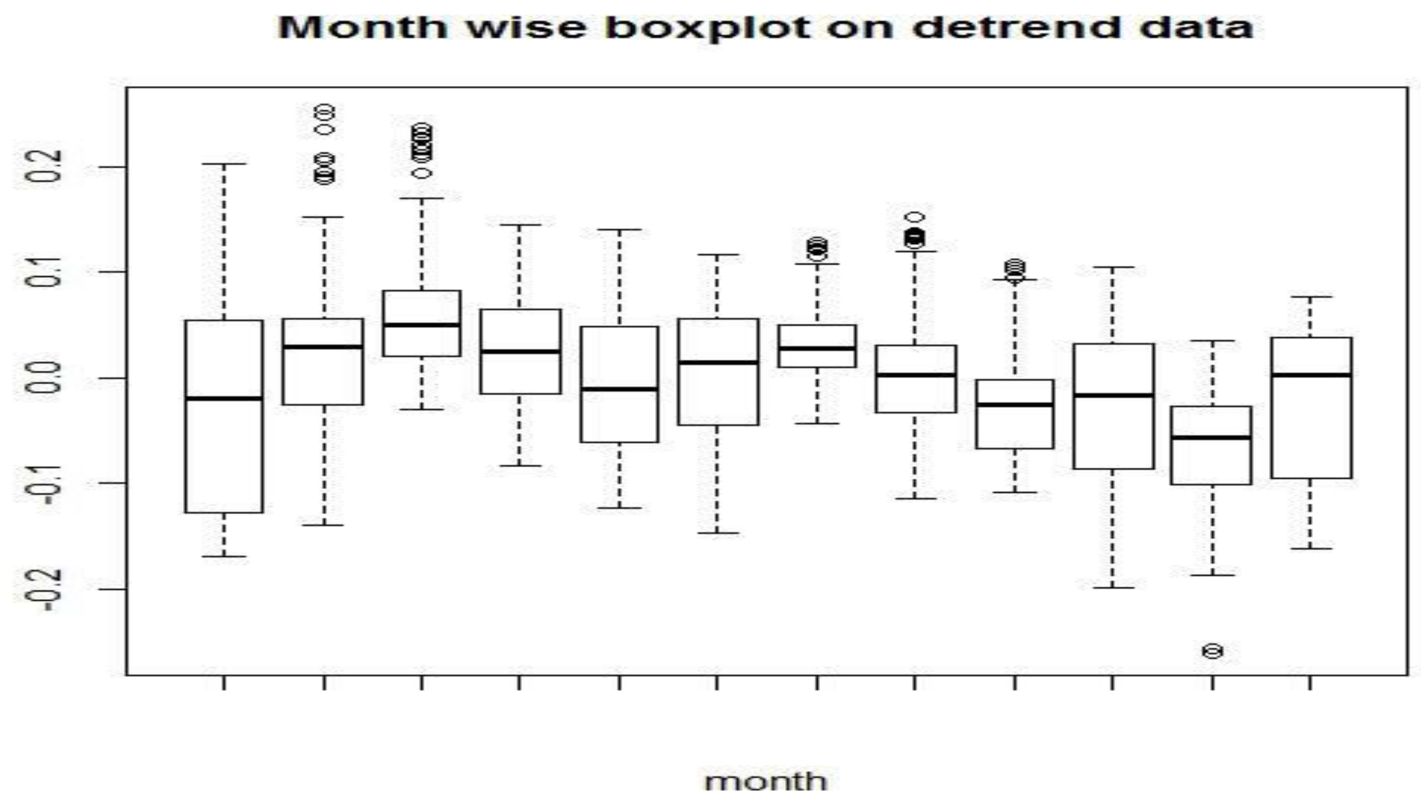


- ✓ The plot shows more or less a random fluctuation about The line Y=0,
- ✓ **Mann kendal test** on detrended data, 2-sided ***p-value =0.92061*** confirms absence of significant trend.

* Absence of weekly seasonality,



❖ Removal of monthly seasonality,



✓ Month wise boxplots reveals presence of monthly seasonality
✓ **Kruskal-Wallis test** : ***p-value < 2.2e-16*** i.e, significant monthly seasonality is present .

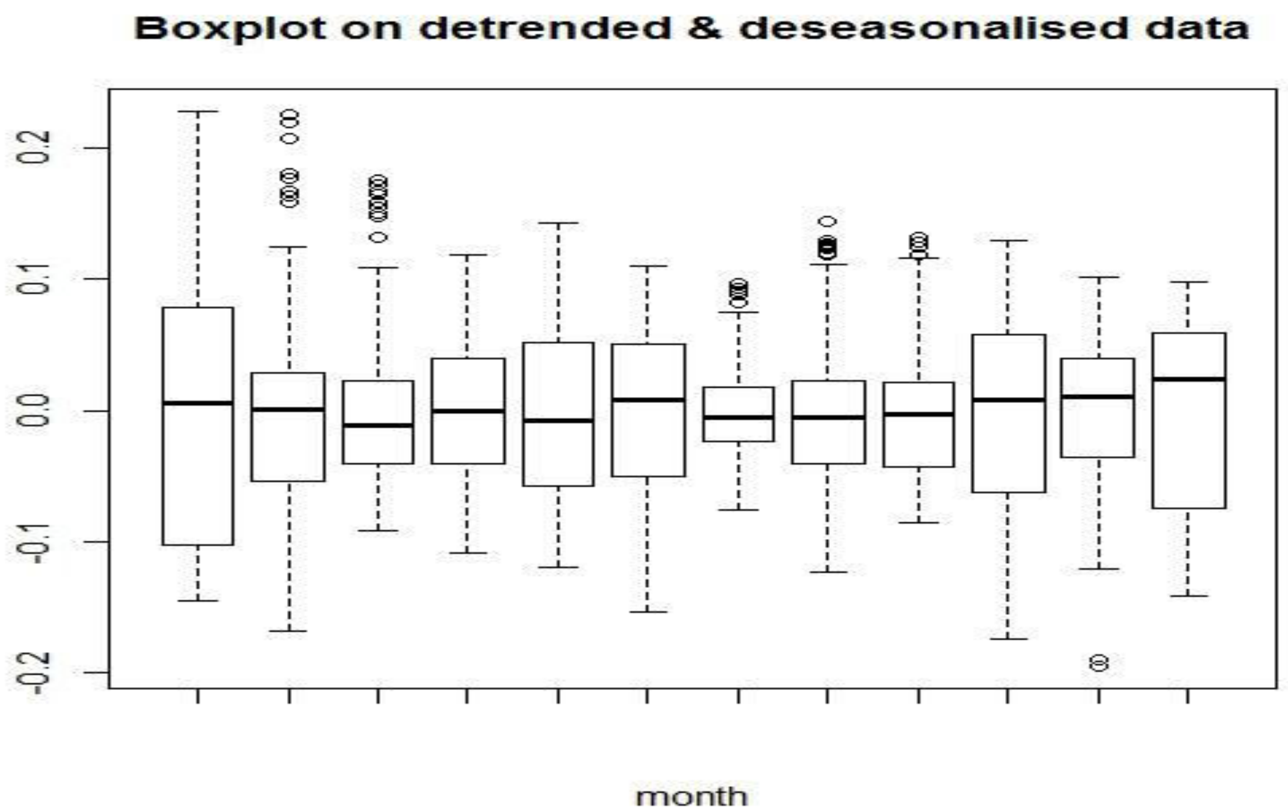
[***Note:*** this test is performed assuming month-wise variance not to differ significantly]

By **method of average** monthly seasonality is removed. The ***seasonal indices*** are,

Month	*Seasonal indices*	*Month*	*Seasonal indices*
January	-0.024448540	July	0.032761126
February	0.028520220	August	0.008119021
March	0.061377242	September	-0.023294866
April	0.025797036	October	-0.024831424
May	-0.002915900	November	-0.066095090
June	0.005856755	December	-0.020845580

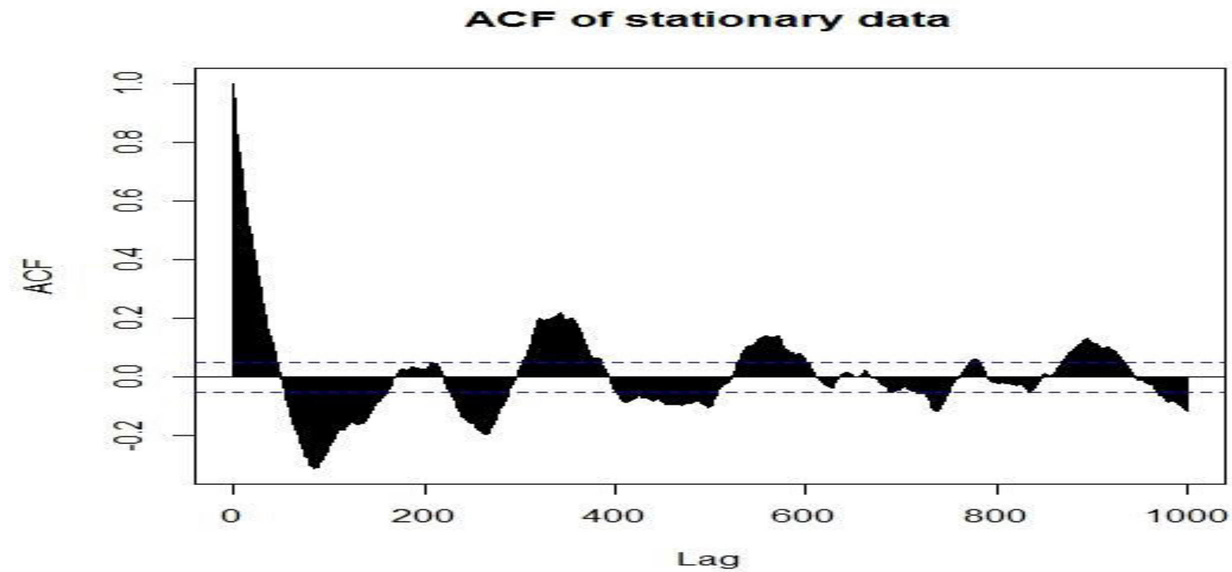
Detrended data is deseasonalised by subtracting corresponding seasonal indices.

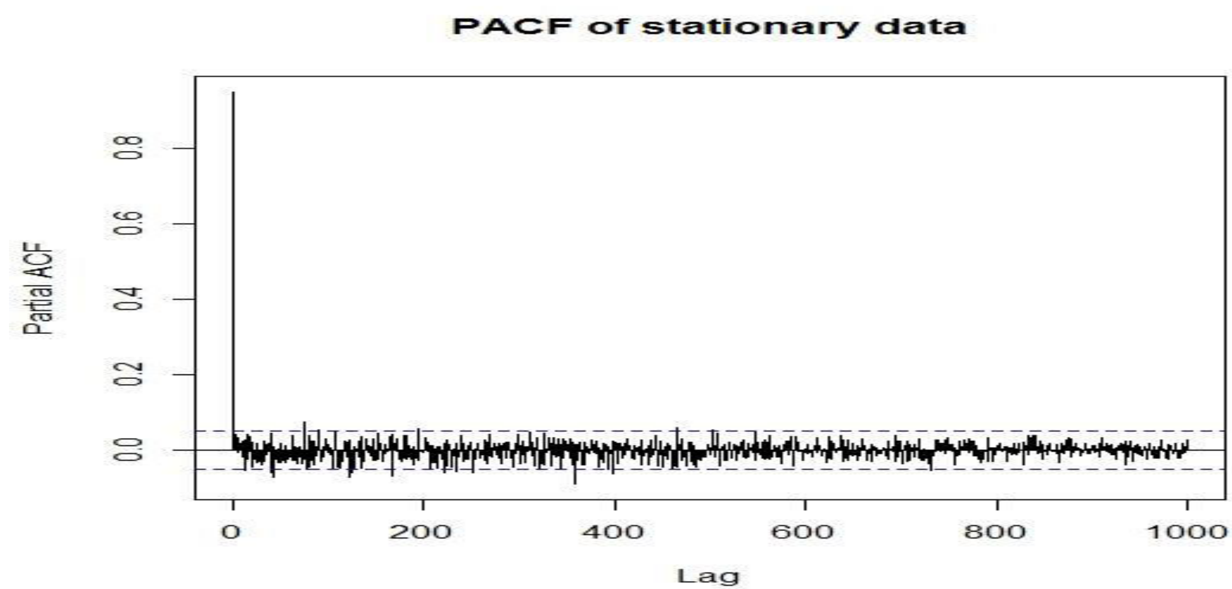
- ✓ **Kruskal-Wallis test** : ***p-value = 0.9879*** which says monthly seasonality is absent
- ✓ Following boxplot also supports it .



Hence we can consider this detrended and deseasonalised data as **stationary** data.

The **ACF**, **PACF** plots of stationary data are as follows,





- ✓ PACF tails off and ACF shows few significant spikes ,hence MA(q) model is appropriate.
- ✓ To keep the model parsimonious we opt for ARMA (p,q)model.

By min^m **AIC criterion**, the model is **ARMA(2,2)** [on stationary data] as follow ,

detrended and deseasonalised Yahoo share price of Japan Z_{2t} at t^{th} time point,

$$\boxed{\boldsymbol{Z_{1t} = -0.0005 - 0.0356 * Z_{1\overline{t-1}} + 0.9437 * Z_{1\overline{t-2}} + \varepsilon_t + 0.9809 * \varepsilon_{t-1} - 0.0191 * \varepsilon_{t-2}}}$$

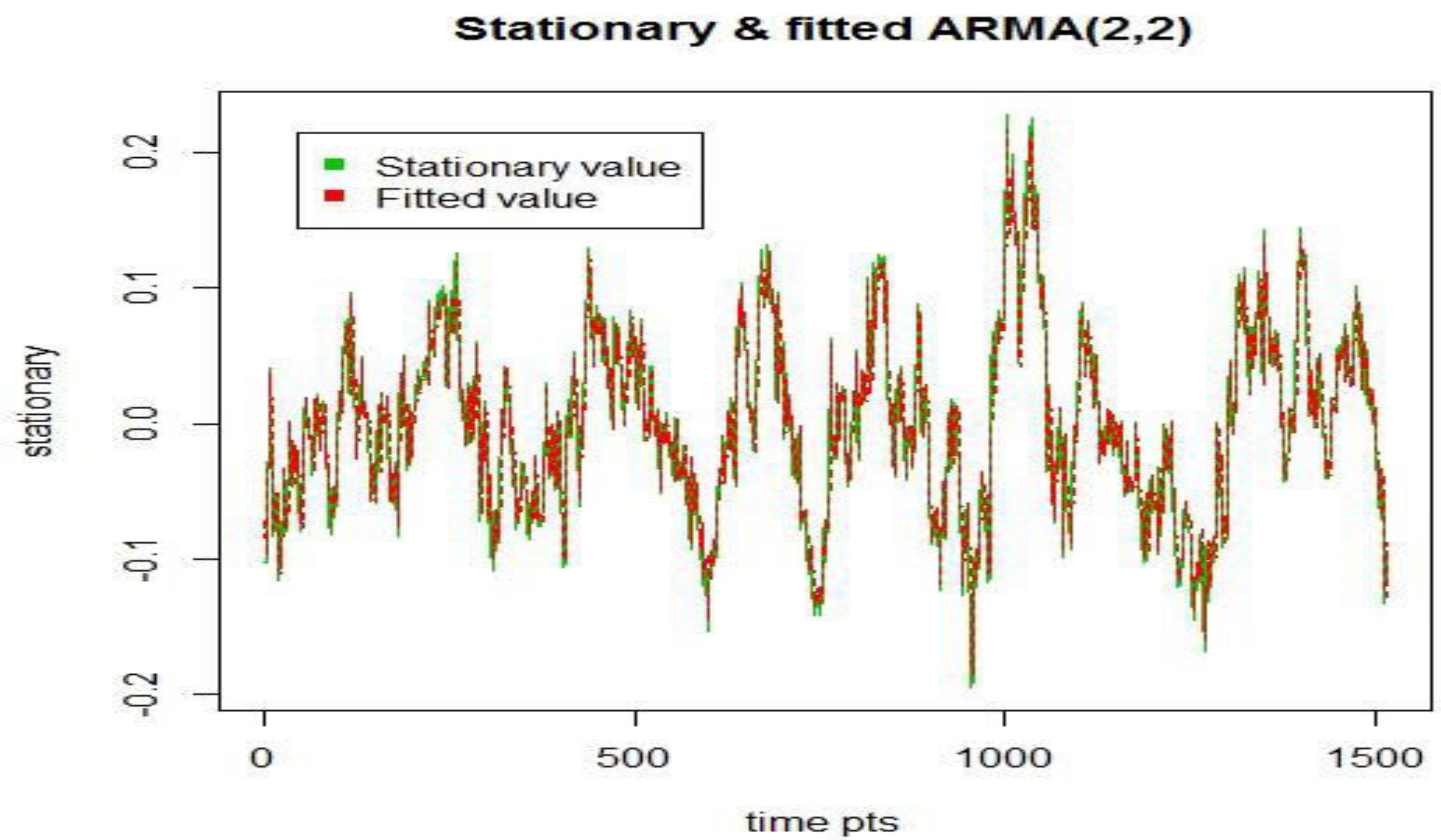
With aic = -7458.5

Ljung-Box test on ARMA residuals is as follow,

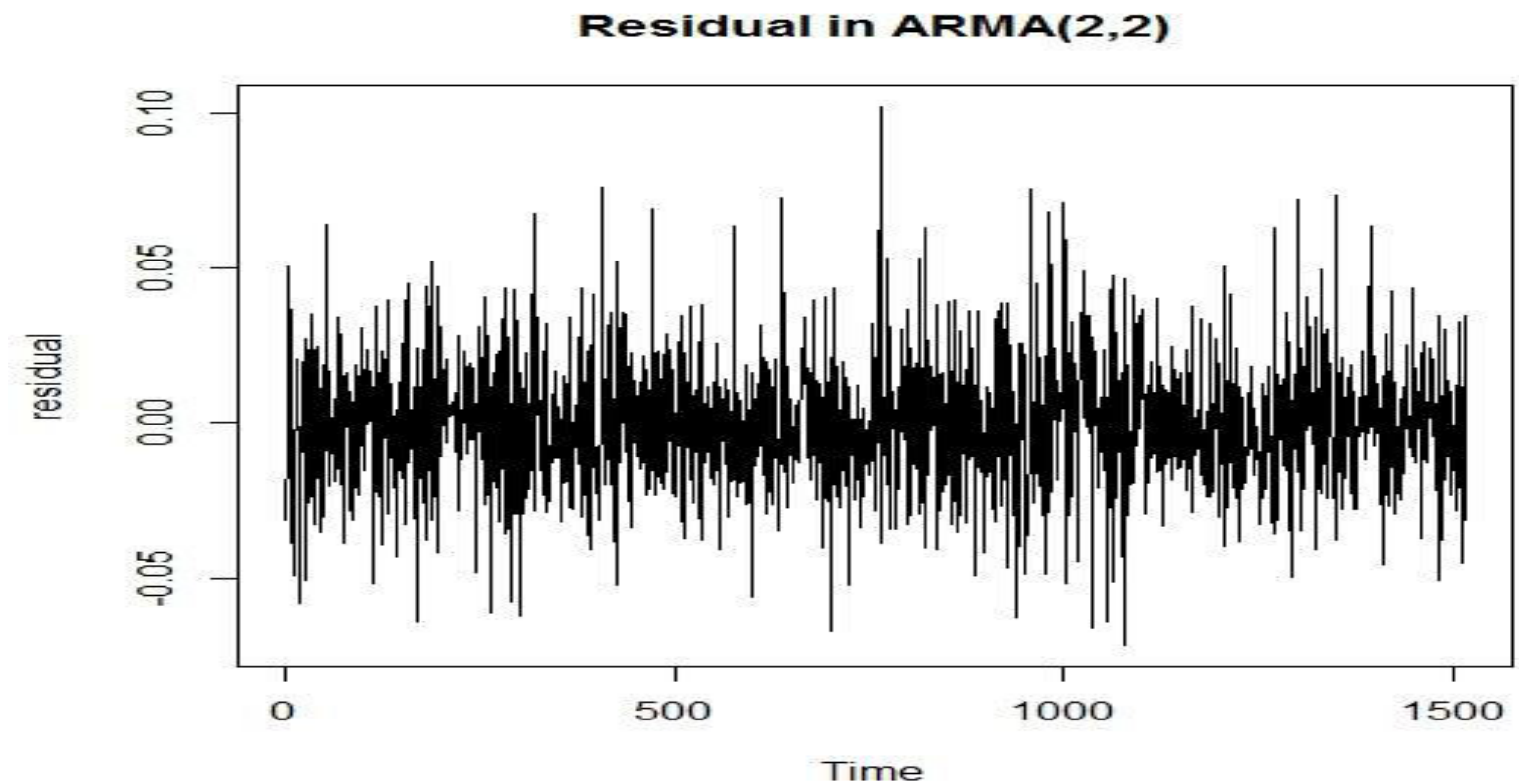
Lags	*Statistic*	*df*	*p-value*
5	5.552310	5	0.3246753
10	7.725947	10	0.6573427
15	15.566587	15	0.4175824

Test tells ARMA residuals are purely random. Hence a good fit.

The following plot shows fitted value with stationary value,

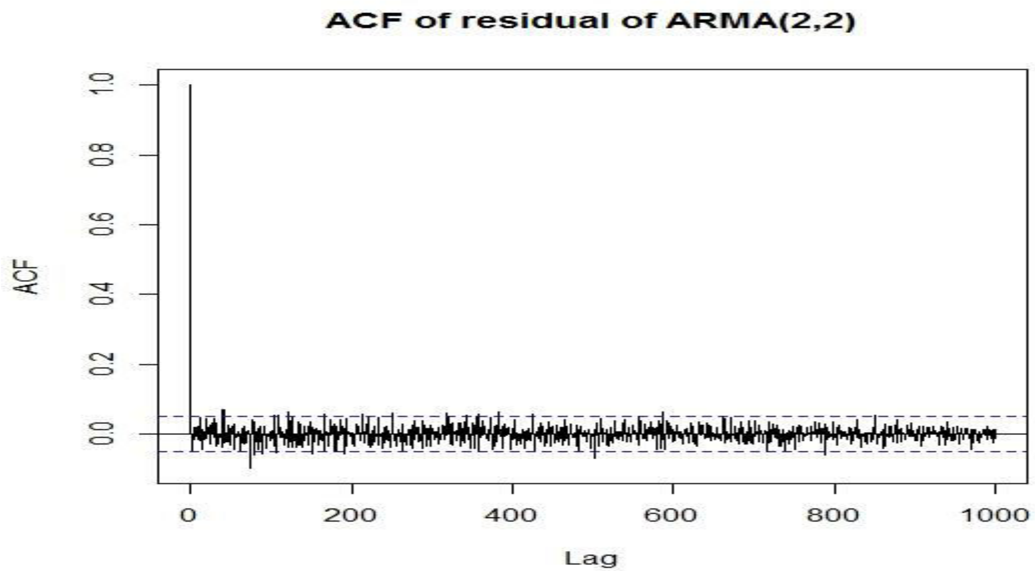


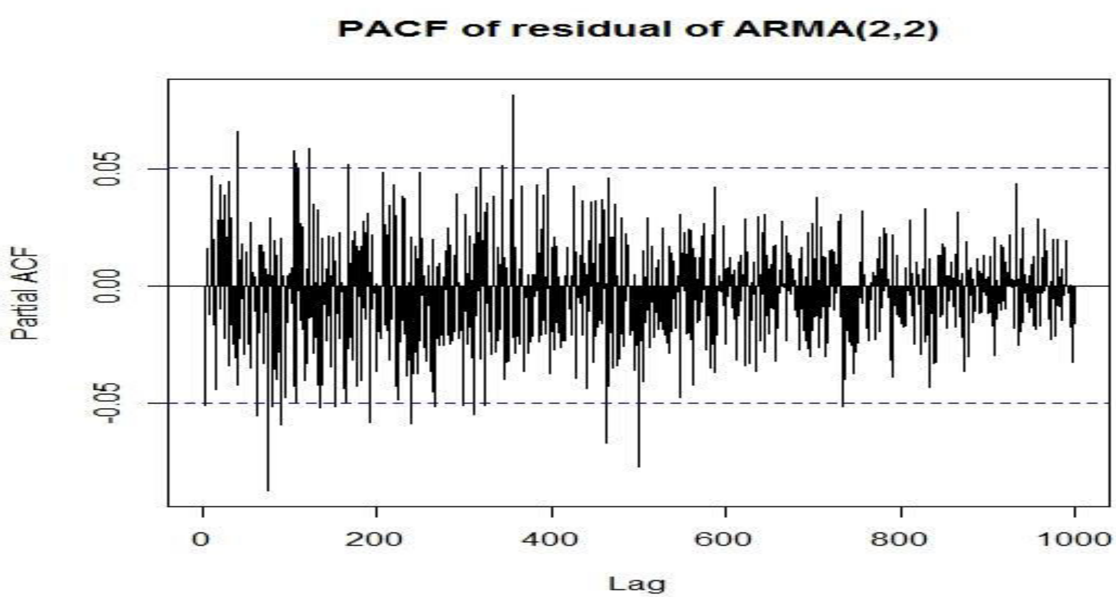
Residual plot is ,



✓ Large fluctuations followed by small fluctuations and again followed by large and this continues

ACF,PACF plot of residuals are,





- ✓ The **Lagrange multiplier test** for ARCH(1) gives **p-value = 0.04468**, which confirms atleast ARCH(1) is required to model the volatility present in the ARIMA residual .

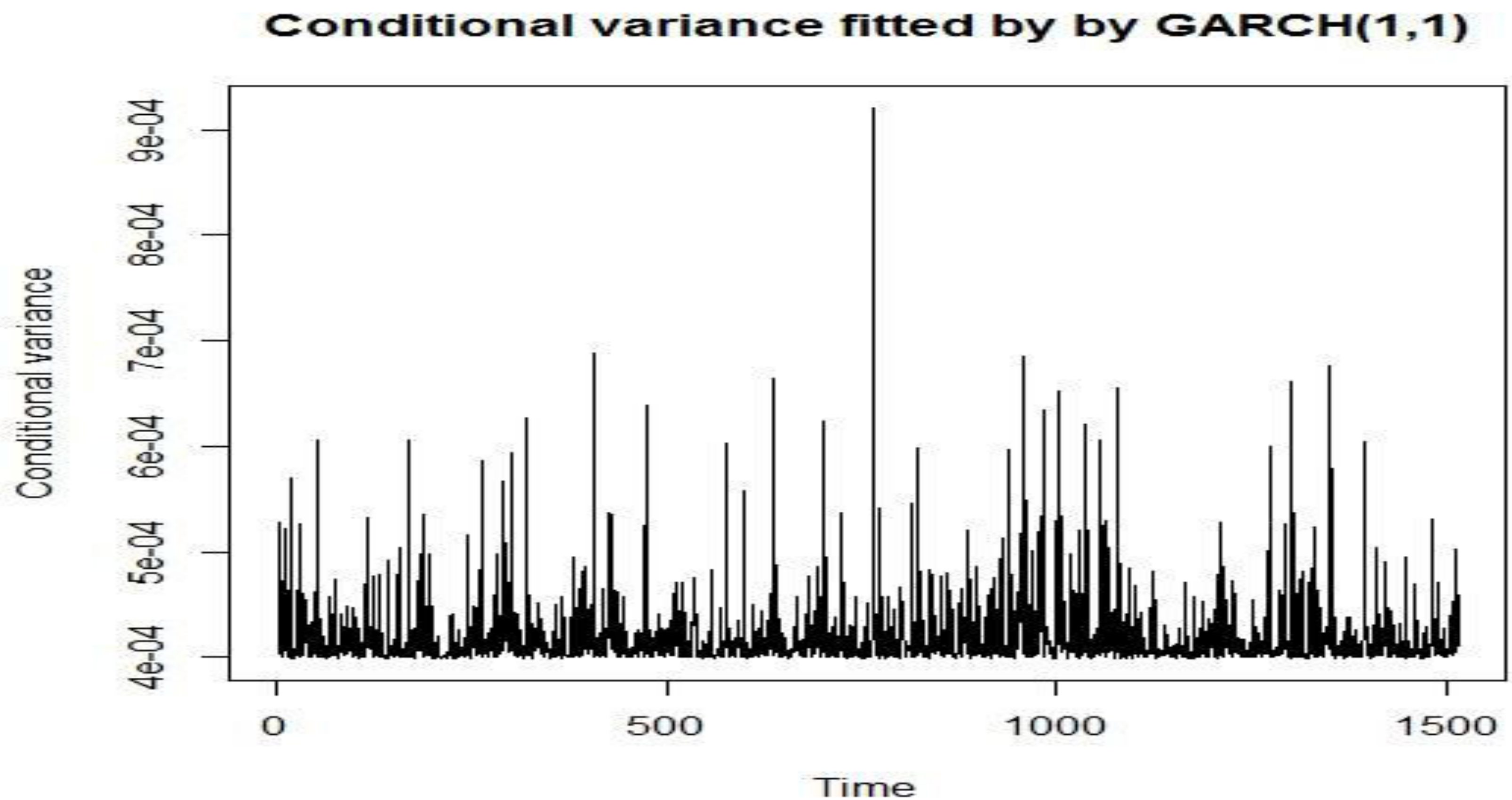
By $\boldsymbol{min^m}$ **AIC criterion** the model is GARCH(1,1)

ARMA residual, $\boldsymbol{r_t = -2.835 * 10^{-6} + \varepsilon_t * \sqrt{\widehat{h_t}}}$

Where $\boldsymbol{\widehat{h_t} = 1.755 * 10^{-5} + 0.01459 * r_{t-1}^2 + 0.9435 * h_{t-1}}$

- **Ljung-Box Test** on GARCH(1,1) residual: ***p-value = 0.7111*** which says the model to be adqequate to capture volatility.

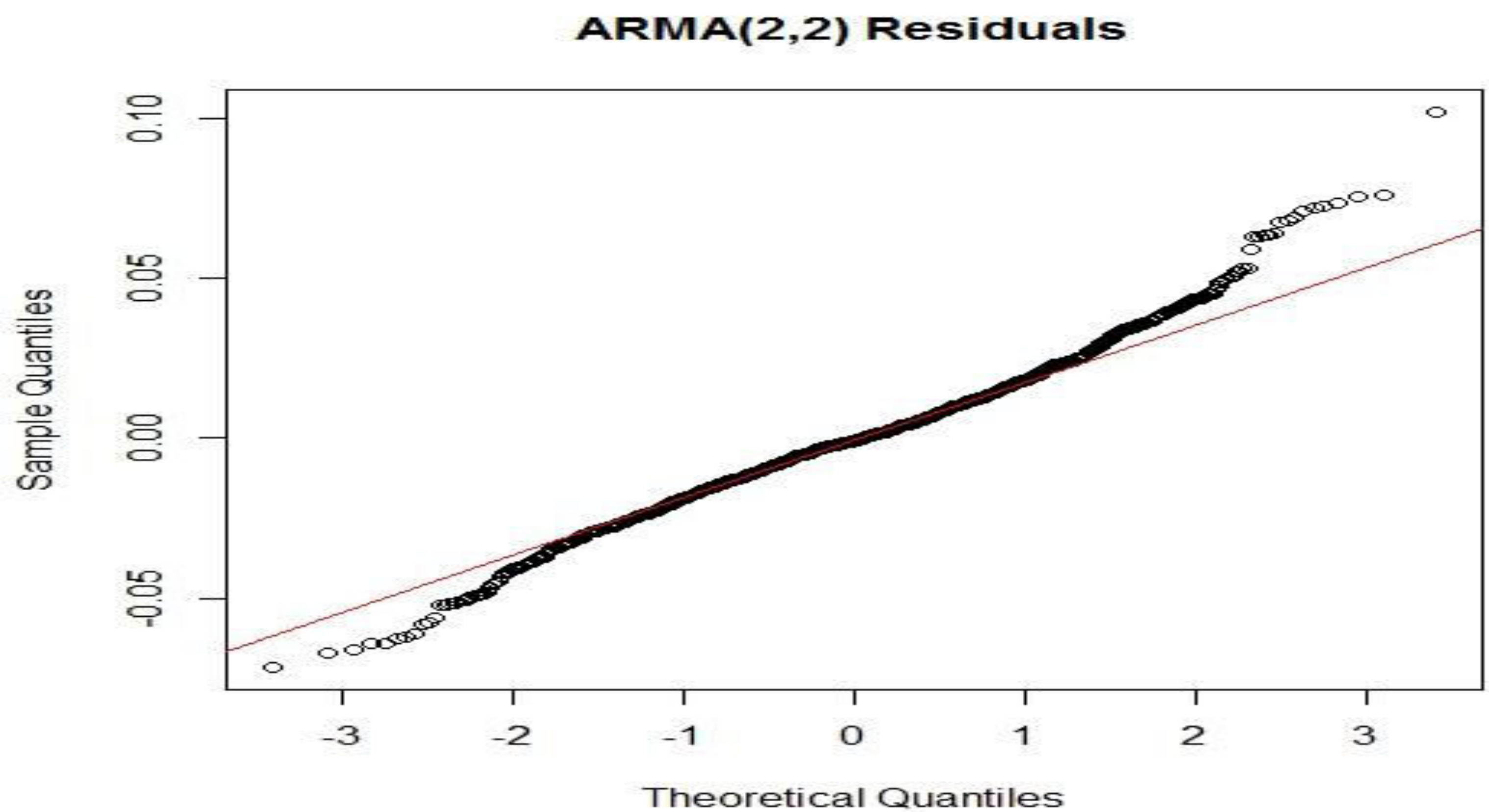
The plot of **conditional variance**, i.e., $\widehat{h_t}$ is as follow,



To find the C.I. $dist^n$ of ARIMA residual is found first

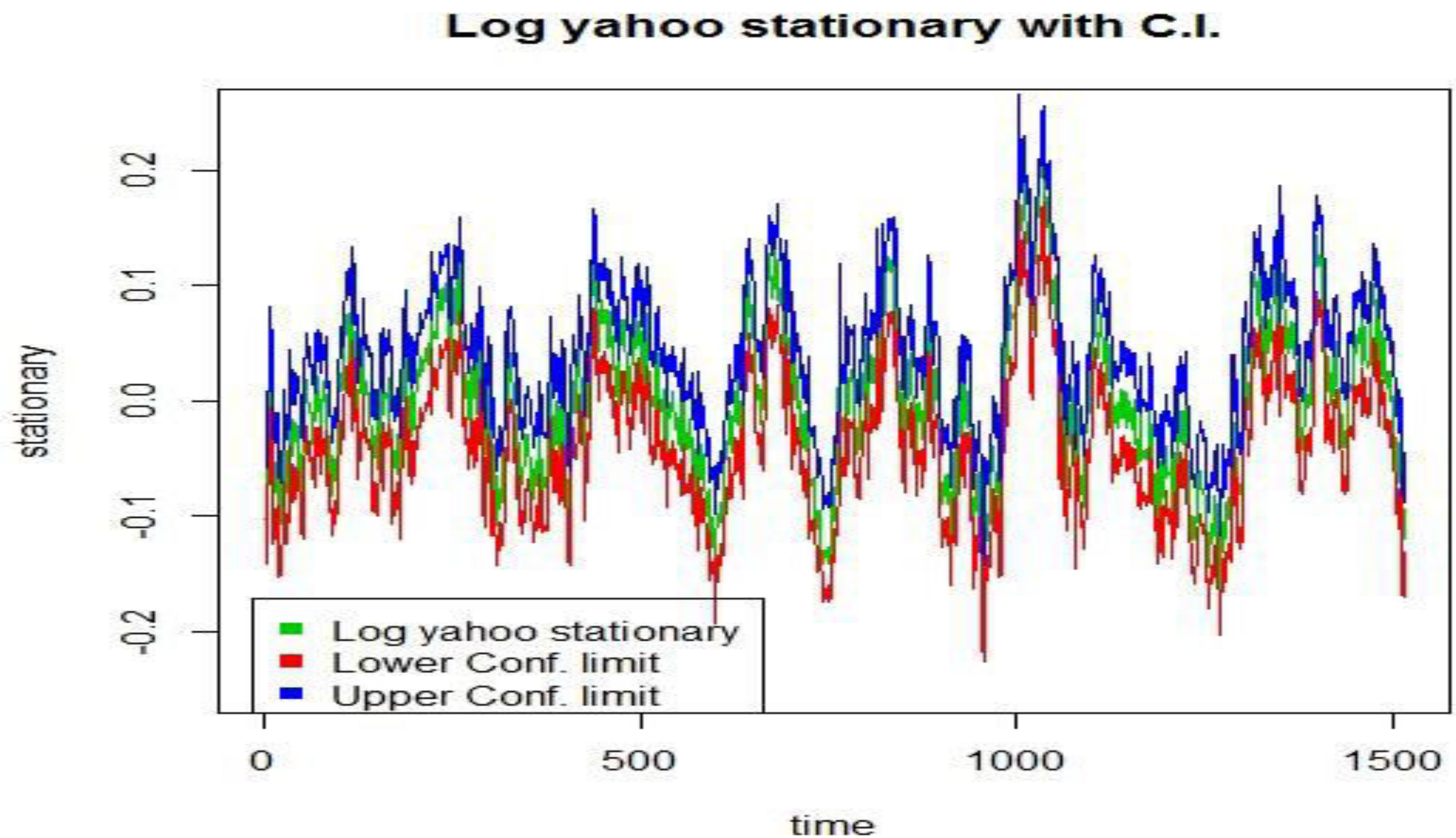
Kurtosis (r_t) =**4.35881** (>3)

- ✓ **Anscombe test** gives ***p-value = 4.768e-12*** which says kurtosis significantly differ from 3 .
- ✓ The **QQ-plot** also supports the fact obtained above

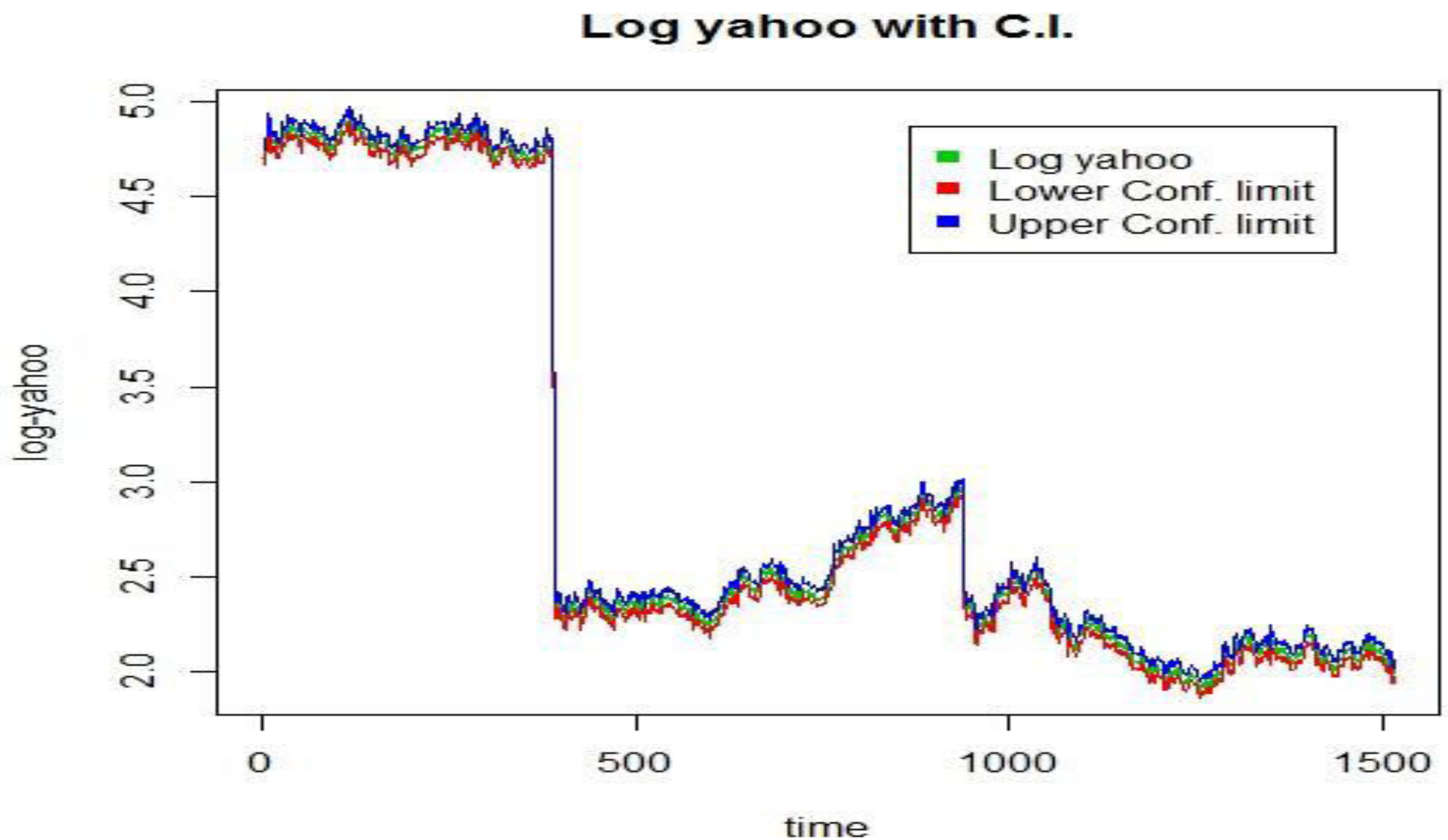


- For having **large frequency on tail** observations, we fit '**t(v)**' distribution.
- v is estimated from ARIMA residual as , v = 5 .

The original log transformed data with confidence interval is as follow,



Now adding the fitted trend and corresponding seasonal indices with conf. limits we get conf. limits for original log transformed data.

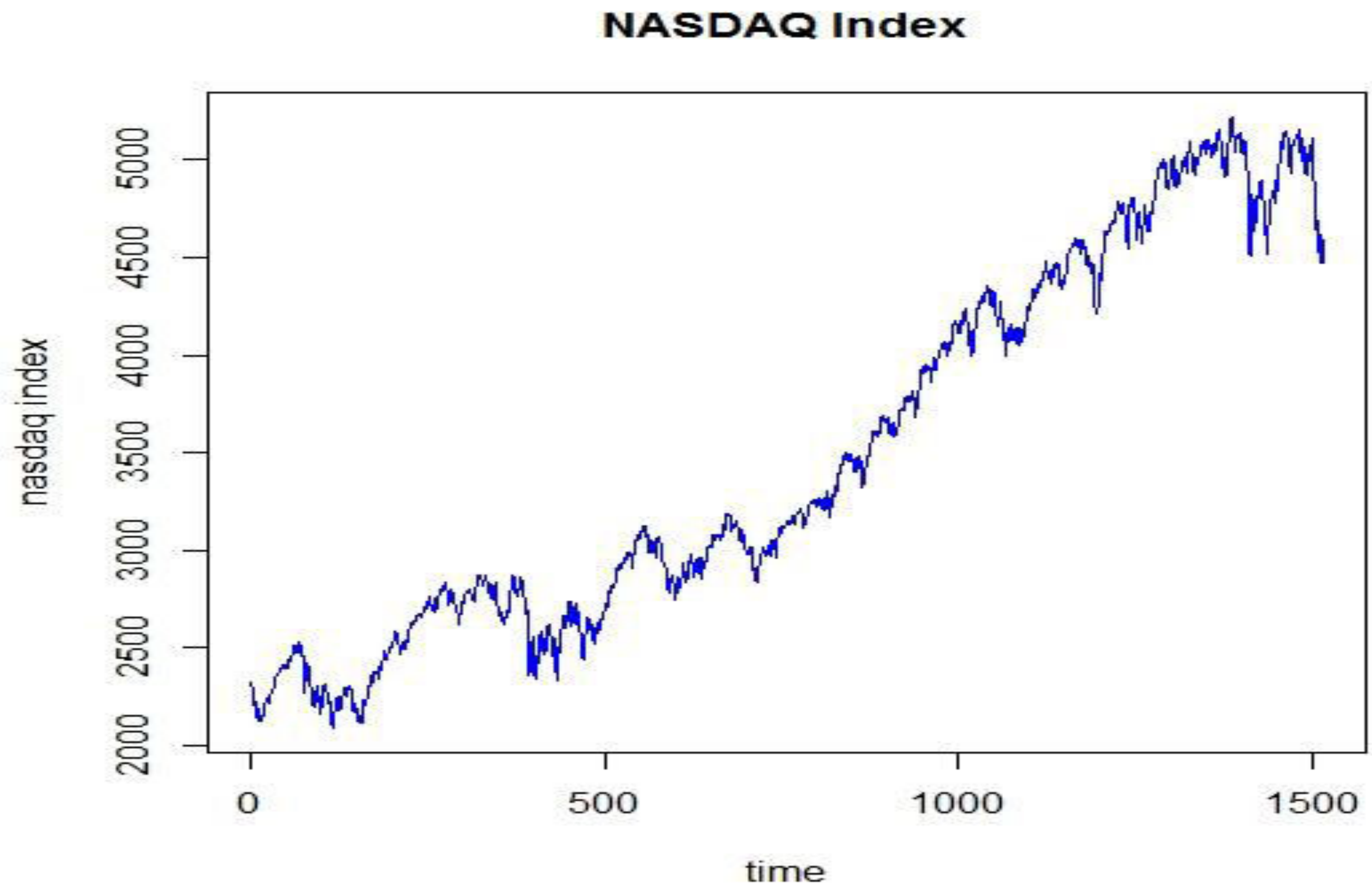


Forecast:

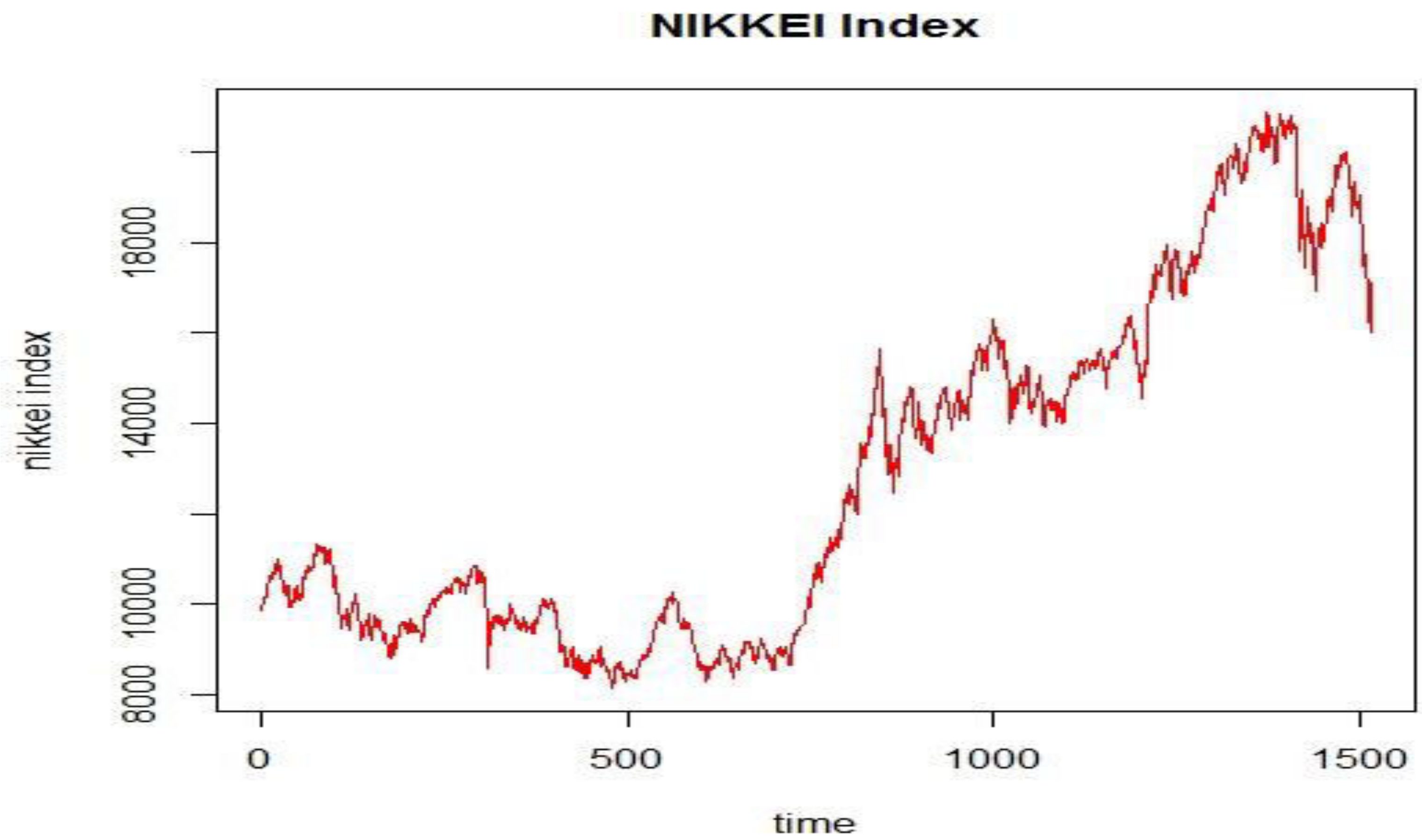
Time point	Observed value	Predicted value	95 % lower conf limit	95 % upper conf limit
27-01-16	7.44	7.462995	7.156204	7.782939
28-01-16	7.47	7.483270	7.175869	7.803841
29-01-16	7.58	7.544727	7.235017	7.867696
01-02-16	7.59	7.973986	7.646873	8.315092
02-02-16	7.74	8.035886	7.706445	8.379410
03-02-16	7.92	8.052431	7.722515	8.396442
04-02-16	7.74	8.111622	7.779476	8.457948
05-02-16	7.62	8.126072	7.793523	8.472811
08-02-16	7.78	8.182732	7.848046	8.531692
09-02-16	7.48	8.195284	7.860259	8.544589

(II). Regression Analysis

❖ **NASDAQ** (Nasdaq) is an American stock exchange, (X_1)

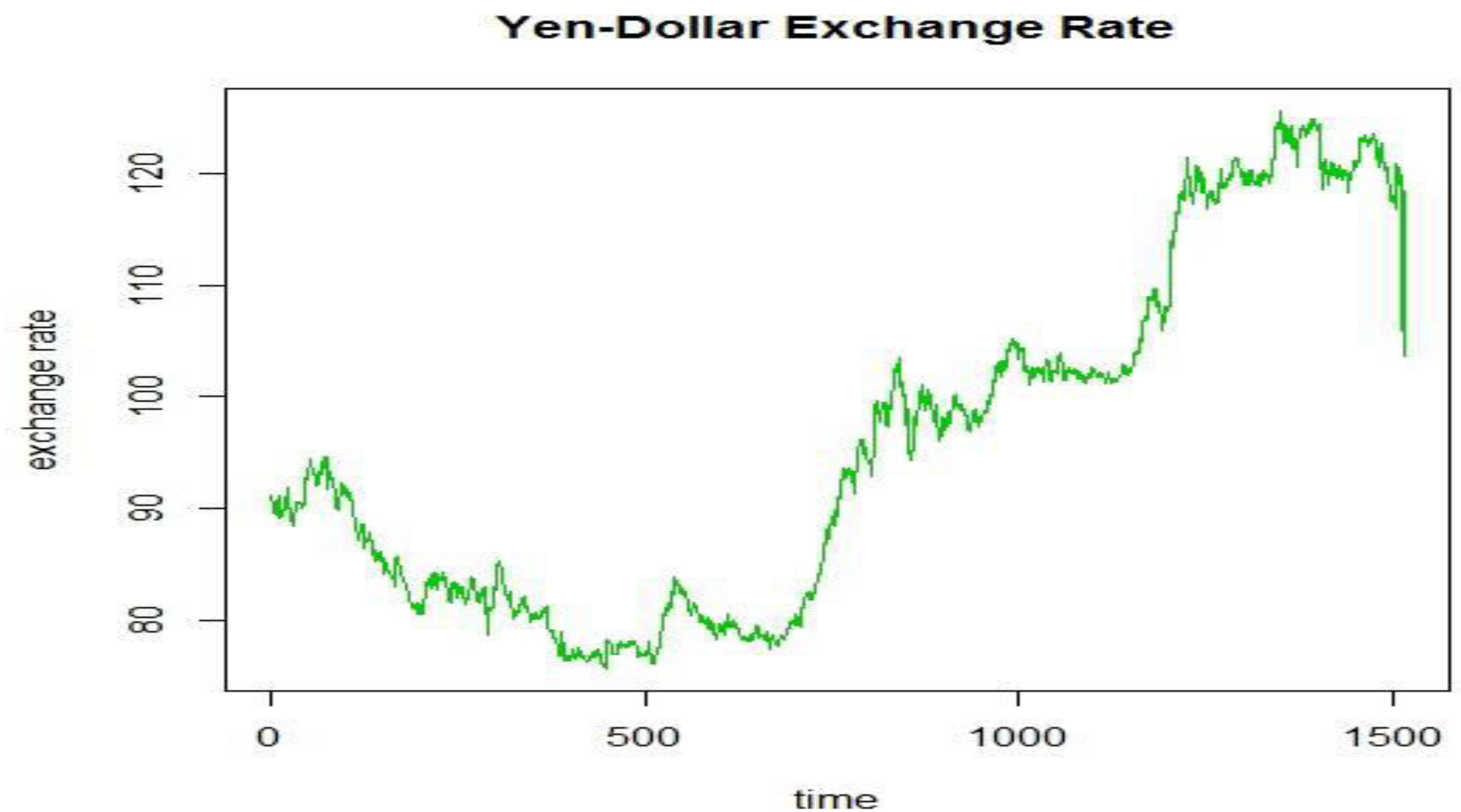


❖ **Nikkei** Stock *Average* is a stock market index for the Tokyo Stock Exchange, $(\boldsymbol{X_2})$



- **Yen-dollar exchange ratio** gives the amount of Yen which is equal to one dollar, **(X_3)**

- some observations were **missing** in this variable, they have been estimated through **MICE** algorithm.



VAR model:

. **Log transformation of all the variables** have been considered, as

- ✓ it stabilizes data
- ✓ it converts the multiplicative model into simpler additive model.

On basis of **minimum AIC**, (considering trend) the **fitted VAR(1)** model is,

$$\begin{bmatrix} Y_{1t} \\ Y_{2t} \end{bmatrix} = \begin{bmatrix} 0.28 \\ -0.24 \end{bmatrix} + \begin{bmatrix} 0.17 * 10^{-4} \\ -4.83 * 10^{-5} \end{bmatrix} * t + \begin{bmatrix} 1.01 & -0.7 * 10^{-3} \\ 4.6 * 10^{-3} & 0.98 \end{bmatrix} \begin{bmatrix} Y_{1\overline{t-1}} \\ Y_{2\overline{t-1}} \end{bmatrix}$$
$$+ \begin{bmatrix} -0.04 & 0.35 * 10^{-2} & -0.14 * 10^{-2} \\ 2.52 * 10^{-2} & -6.49 * 10^{-2} & 0.15 \end{bmatrix} \begin{bmatrix} X_{1t} \\ X_{2t} \\ X_{3t} \end{bmatrix} + \begin{bmatrix} e_{1t} \\ e_{2t} \end{bmatrix}$$

With Correlation matrix of residuals:

$$\begin{array}{cc} & \begin{array}{cc} Y_1 & Y_2 \end{array} \\ \begin{array}{c} Y_1 \\ Y_2 \end{array} & \begin{bmatrix} 1.0000 & 0.1531 \\ 0.1531 & 1.0000 \end{bmatrix} \end{array}$$

Granger causality test is performed on fitted var(1)

- ✓ H0: Y_2(yahoo japan) do not Granger-cause Y_1(yahoo usa): ***p-value = 0.5524***
- ✓ H0: Y_1(yahoo usa) do not Granger-cause Y_2(yahoo japan): ***p-value = 0.7458***
- ✓ H0: No instantaneous causality between: Y_1 and Y_2: ***p-value = 4.076e-09***

Conclusion :

- none of Y_1 and Y_2 are granger cause to each other.
- but there is a significant instantaneous causality between Y_1 and Y_2.

Hence we can opt for Simultaneous Equations Model **(SEM)**.

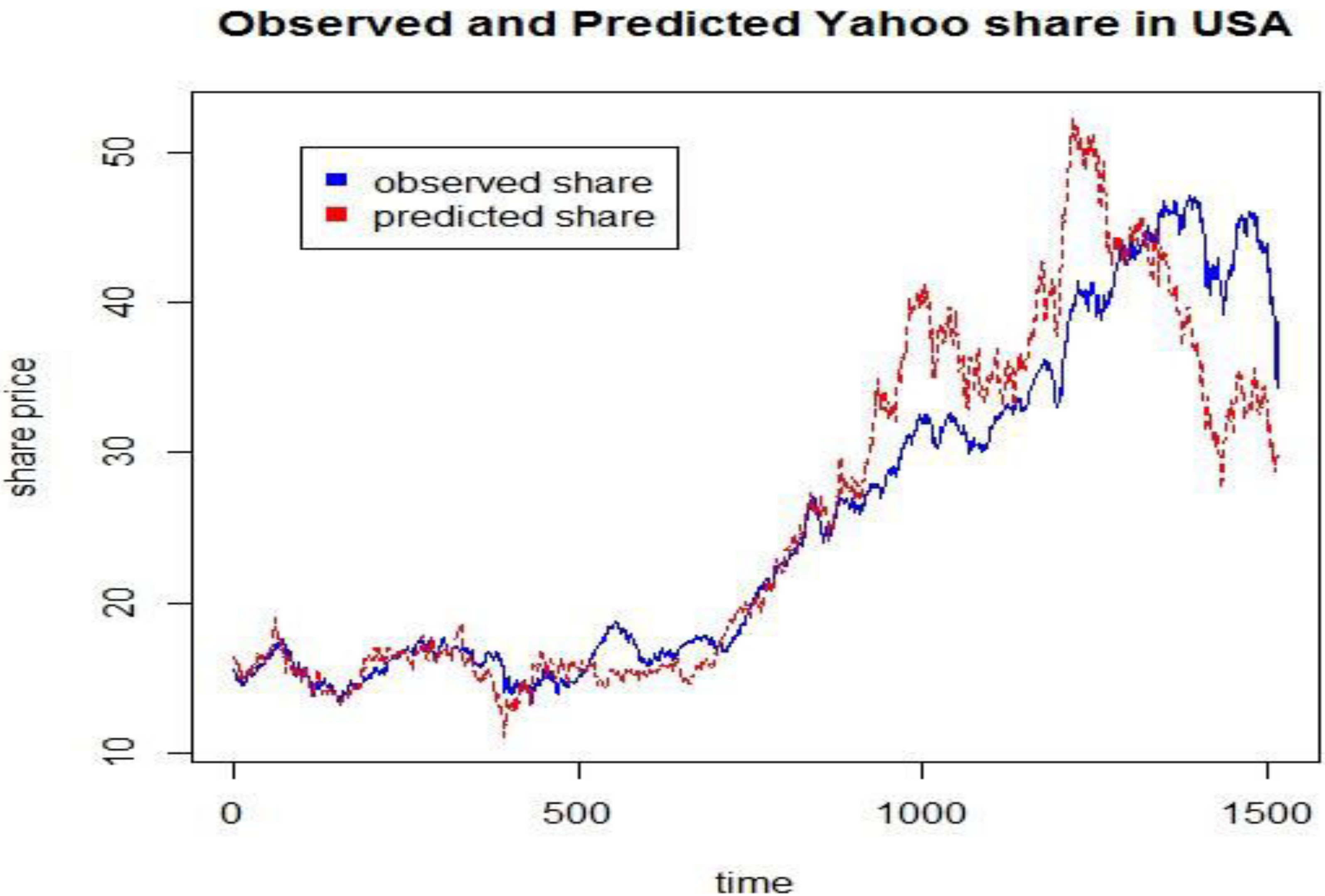
Simultaneous Equations Model (SEM):

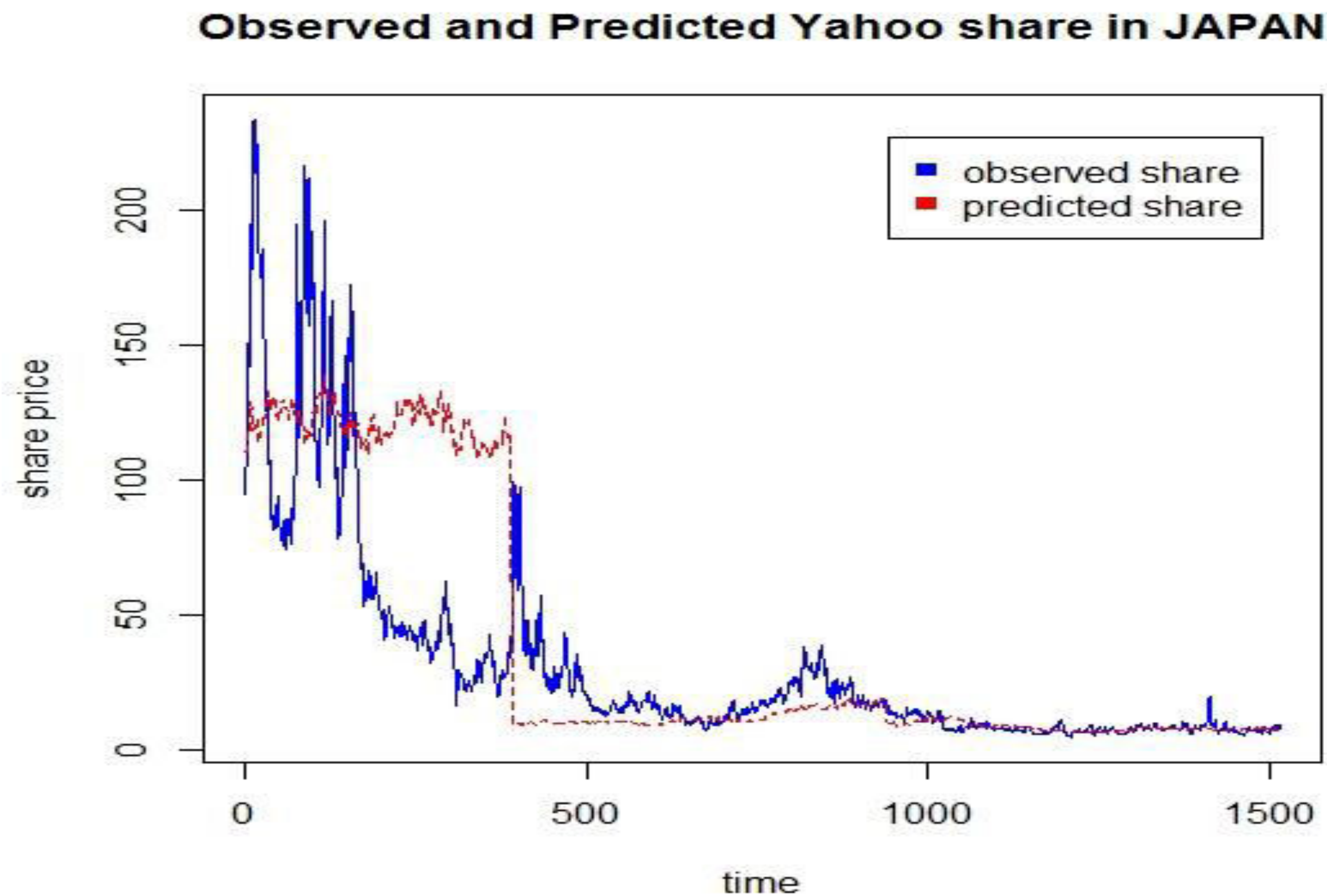
- **reduced form** on **log transformed data** by applying **2SLS** method is as follow,

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} -9.669 \\ 20.989 \end{bmatrix} + \begin{bmatrix} 0.862 & 0.290 & 0.677 \\ -6.612 & 4.066 & -0.574 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}$$

$\boldsymbol{X_2}$, $\boldsymbol{X_1}$ are **instrumented** respectively to estimate $\boldsymbol{Y_1}$, $\boldsymbol{Y_2}$

By reverse exponential transformation the plots of predicted values of Y_1 and Y_2 through 2SLS are as follow,


Observed and Predicted Yahoo share in USA
observed share
predicted share
share price
10
20
30
40
50
0
500
1000
1500
time


Observed and Predicted Yahoo share in JAPAN
observed share
predicted share
share price
0
50
100
150
200
0
500
1000
1500
time

Forecast:

Time point	Predicted Y_1	Observed Y_1	Predicted Y_2	Observed Y_2
27-01-16	38.354600	29.69	6.921515	7.44
28-01-16	37.609248	28.75	7.671636	7.47
29-01-16	37.86620	29.51	7.36738	7.58
01-02-16	39.639950	29.57	7.341739	7.59
02-02-16	39.836134	29.06	7.658916	7.74
03-02-16	38.92950	27.68	9.04666	7.92
04-02-16	38.251513	29.15	9.093485	7.74
05-02-16	37.948725	27.97	8.798173	7.62
08-02-16	36.84478	27.05	10.56277	7.78
09-02-16	36.12601	26.82	12.63097	7.48

❑ ***Remark:***

The main reason of predicted values to differ is insufficient covariates. Share price actually depends on so many factors and their mutual interactions. Consideration of more covariates should improve the predictions.

References:

1. Introduction to Modern Time Series Analysis - Kirchgassner, Wolters, Hassler.
2. Applied Econometric Time Series - Walter Enders.
3. Econometric methods – Johnston, Dinardo.
4. wikipedia

Acknowledgements:

*I am highly indebted to Prof. **Sugata Sen Roy**, Prof. **Manisha Pal** and Prof. **Uttam Bandyaopadhyay** for their guidance, support and constant supervision in completion of the project.*

*I want to offer my sincere gratitude to our department, **Department of Statistics, University of Calcutta** for giving me the opportunity to do this project.*

*I would also like to thank my **parents** and **friends** who helped me a lot in finalizing this project within the limited time frame.*

THANK YOU