Nirbhay Sharma (B19CSE114) ; Chirag Goyal (B19CSE105)

Programming assignment TSA

Question: Model the topic assigned using the data according to the topic mentioned. You should model the following-

1. ARMA/ ARIMA and its variation
2. ARCH/ GARCH and its variation.
3. Perform Forecasting and its evaluation. Make sure you test the relevant tests for each of these three.

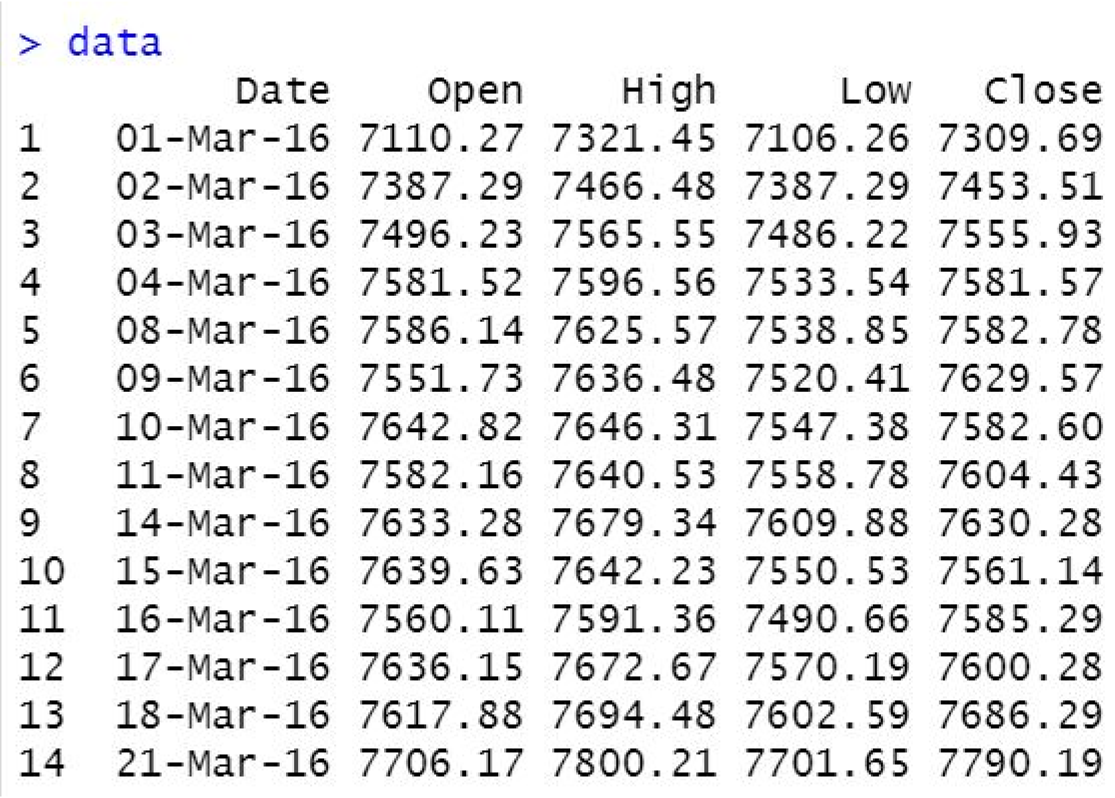
Approach and Implementation of the Problem

To implement the problem, we are going to utilize the Box-Jenkins methodology:

Data Collection & Data Format

Collect the time series data relevant to the problem at hand. This could include historical data on stock prices, sales data, weather data, or any other data that exhibits a time-dependent pattern.

Here, we are using the data that has four columns containing the daily data for opening, closing, high, and low values of the BSE 100 dataset from 2016 to 2023. For analysis purposes, we use Open columns only.



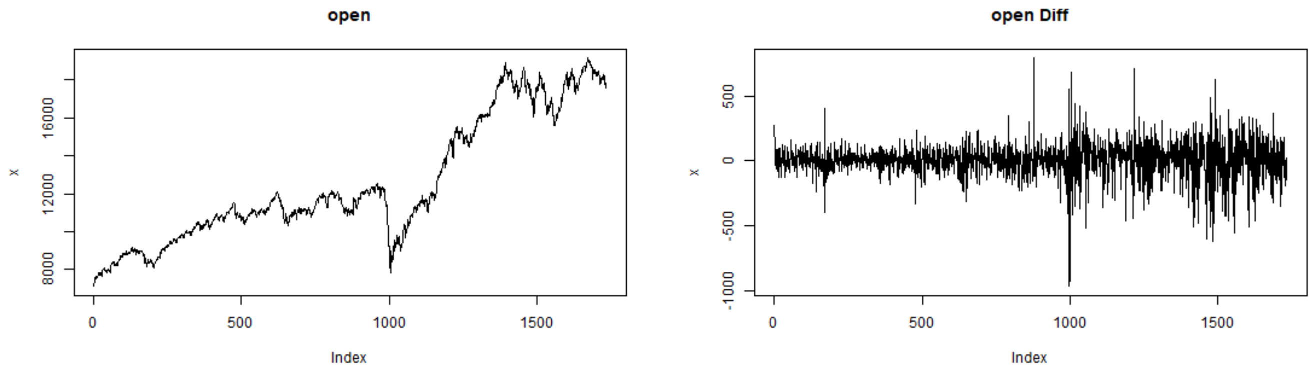
```
> data
         Date    Open    High     Low   Close
1   01-Mar-16 7110.27 7321.45 7106.26 7309.69
2   02-Mar-16 7387.29 7466.48 7387.29 7453.51
3   03-Mar-16 7496.23 7565.55 7486.22 7555.93
4   04-Mar-16 7581.52 7596.56 7533.54 7581.57
5   08-Mar-16 7586.14 7625.57 7538.85 7582.78
6   09-Mar-16 7551.73 7636.48 7520.41 7629.57
7   10-Mar-16 7642.82 7646.31 7547.38 7582.60
8   11-Mar-16 7582.16 7640.53 7558.78 7604.43
9   14-Mar-16 7633.28 7679.34 7609.88 7630.28
10  15-Mar-16 7639.63 7642.23 7550.53 7561.14
11  16-Mar-16 7560.11 7591.36 7490.66 7585.29
12  17-Mar-16 7636.15 7672.67 7570.19 7600.28
13  18-Mar-16 7617.88 7694.48 7602.59 7686.29
14  21-Mar-16 7706.17 7800.21 7701.65 7790.19
```

Analysing y and $\triangle y$

1. Analyzing the time series data for variable y and its corresponding change in y provides valuable insights into the trends and dynamics of the dataset over time.
2. Studying the patterns and behavior of both y and its change allows for a more comprehensive analysis of the dataset, revealing important information about the underlying factors driving the changes

observed in y. This analysis can aid in making informed decisions and predictions based on the dataset's characteristics.

Plots for the data patterns



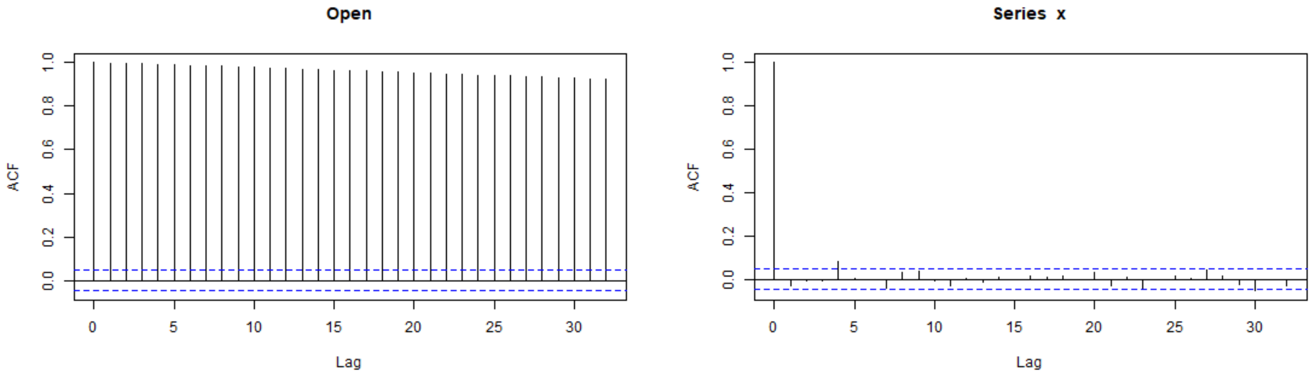
(a) y

(b) △y

Analysing ACF and PACF of y and $\triangle y$

1. Analyzing the Partial Autocorrelation Function (PACF) and Autocorrelation Function (ACF) of variable y and its change provides valuable insights into the dataset's time-dependent relationships and potential lag effects.
2. The PACF and ACF plots offer valuable visualizations that can help identify the significant lags in y and its change, which can be useful in determining the appropriate order for AR (AutoRegressive) and MA (Moving Average) components in time series models like ARIMA and GARCH.

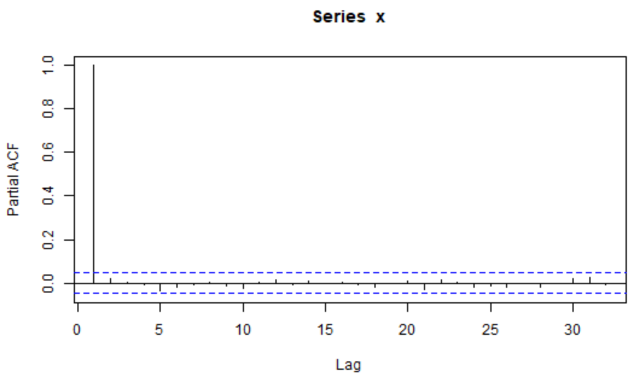
Plots for ACF for y and $\triangle y$



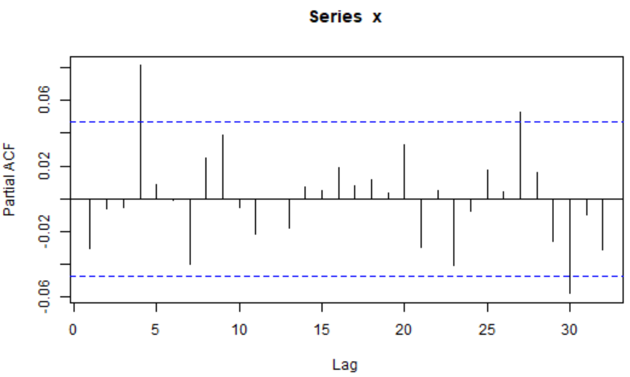
(a) y

(b) △y

Plots for PACF for y and $\triangle y$



(a) y



(b) △y

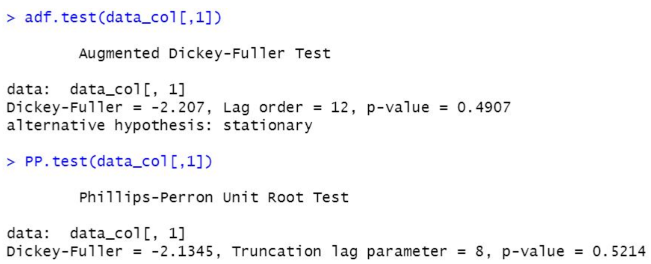
Exploratory Data Analysis (EDA)

Perform EDA on the time series data to gain insights into its properties, such as trend, seasonality, and stationarity. This can be done using visualizations, statistical tests, and other techniques.

Test of stationarity

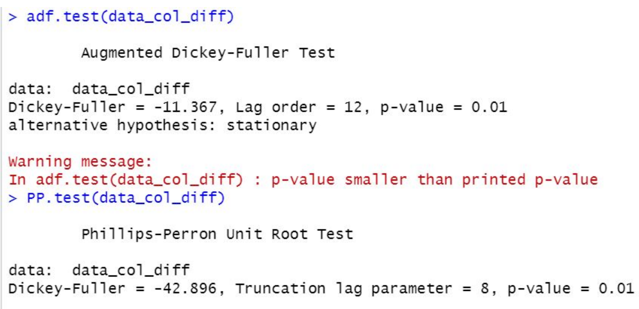
ADF & PP test of y and $\triangle y$

1.Testing for stationarity is a critical step in ARMA/ARIMA modeling, as it ensures that the time series data exhibits consistent statistical properties over time and is suitable for modeling. 2. Common tests for stationarity, such as the Augmented Dickey-Fuller (ADF) test or PP test, stand for "Phillips-Perron." The Phillips-Perron test is a statistical test used to test for the presence of a unit root in time series data, which indicates non-stationarity. It is a variation of the Dickey-Fuller test that corrects for autocorrelation and heteroscedasticity in the residuals, making it a useful tool for testing stationarity in time series data. The Phillips-Perron test is commonly used in econometrics and time series analysis to assess the stationarity of a dataset before applying various time series models, such as ARIMA or GARCH. It can help determine if the time series data is stationary or requires further preprocessing before fitting an ARMA/ARIMA model.



```
> adf.test(data_col[,1])

        Augmented Dickey-Fuller Test

data:  data_col[, 1]
Dickey-Fuller = -2.207, Lag order = 12, p-value = 0.4907
alternative hypothesis: stationary

> PP.test(data_col[,1])

        Phillips-Perron Unit Root Test

data:  data_col[, 1]
Dickey-Fuller = -2.1345, Truncation lag parameter = 8, p-value = 0.5214
```

(a) y



```
> adf.test(data_col_diff)

        Augmented Dickey-Fuller Test

data:  data_col_diff
Dickey-Fuller = -11.367, Lag order = 12, p-value = 0.01
alternative hypothesis: stationary

Warning message:
In adf.test(data_col_diff) : p-value smaller than printed p-value
> PP.test(data_col_diff)

        Phillips-Perron Unit Root Test

data:  data_col_diff
Dickey-Fuller = -42.896, Truncation lag parameter = 8, p-value = 0.01
```

(b) △y

Box-Pierce test for y

1. The Box-Pierce test is a statistical test commonly used in time series analysis to assess the goodness-of-fit of a model by examining the residuals of the fitted model for variable y.
2. The Box-Pierce test statistic, calculated from the autocorrelation function (ACF) of the residuals, provides a quantitative measure of the residual autocorrelation, which can indicate if the model is

capturing the underlying patterns in the data or if further model refinement is needed.

χ-squared	df	P-value
1726.9	1	$< 2.2e^{-16}$
3446.2	2	$< 2.2e^{-16}$
5157.9	3	$< 2.2e^{-16}$
6862	4	$< 2.2e^{-16}$
16915	10	$< 2.2e^{-16}$
25065	15	$< 2.2e^{-16}$
28269	17	$< 2.2e^{-16}$

ARIMA Modelling for y

Choose the appropriate ARMA or ARIMA model based on the properties observed in the EDA. ARMA (AutoRegressive Moving Average) and ARIMA (AutoRegressive Integrated Moving Average) are popular time series models that can capture linear dependencies and trends in the data. Fit the selected model to the data and tune its hyperparameters, such as the order of the autoregressive (AR), moving average (MA), and integrated (I) components.

We have selected three different ARIMA models to do the modeling:

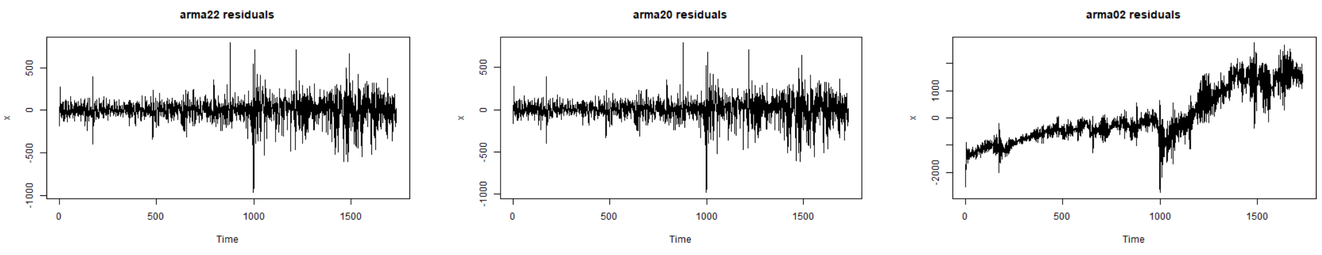
- ARIMA(2,0,2)
- ARIMA(2,0,0)
- ARIMA(0,0,2)

ARIMA(2,0,2), ARIMA(2,0,0), ARIMA(0,0,2)

Evaluate the fitted ARMA/ARIMA model using diagnostic tests such as residual analysis, normality tests, and autocorrelation tests. This ensures the model assumptions are met and the residuals are white noise

Plots for residuals

1. Residual plotting is important after modeling because it helps to assess the adequacy of the model's fit to the data. Residuals are the differences between the observed values and the predicted values from the model, and analyzing their behavior can provide insights into the model's performance.
2. Residual plots can help identify patterns or trends in the residuals, such as autocorrelation or heteroscedasticity, indicating potential issues with the model's assumptions or specifications. Residual plots can also reveal if the residuals are centered around zero, normally distributed, and exhibit constant variance, which are important assumptions of many time series models.
3. By examining the residuals, we can identify potential problems with the model's fit and make necessary adjustments, such as refining the model parameters, adding or removing components, or accounting for any non-random patterns in the residuals. Residual plots provide a visual and quantitative way to assess the model's accuracy and appropriateness for the data, ensuring reliable forecasts and accurate interpretations of the model's results.

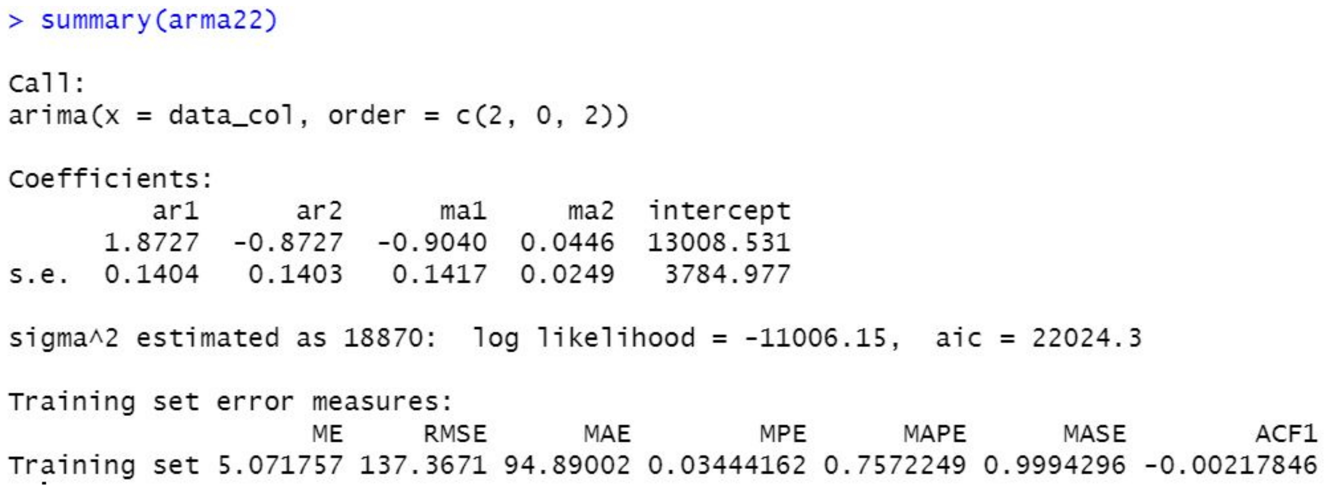


(a) ARIMA(2,0,2) (b) ARIMA(2,0,0) (c) ARIMA(0,0,2)

Model Selection

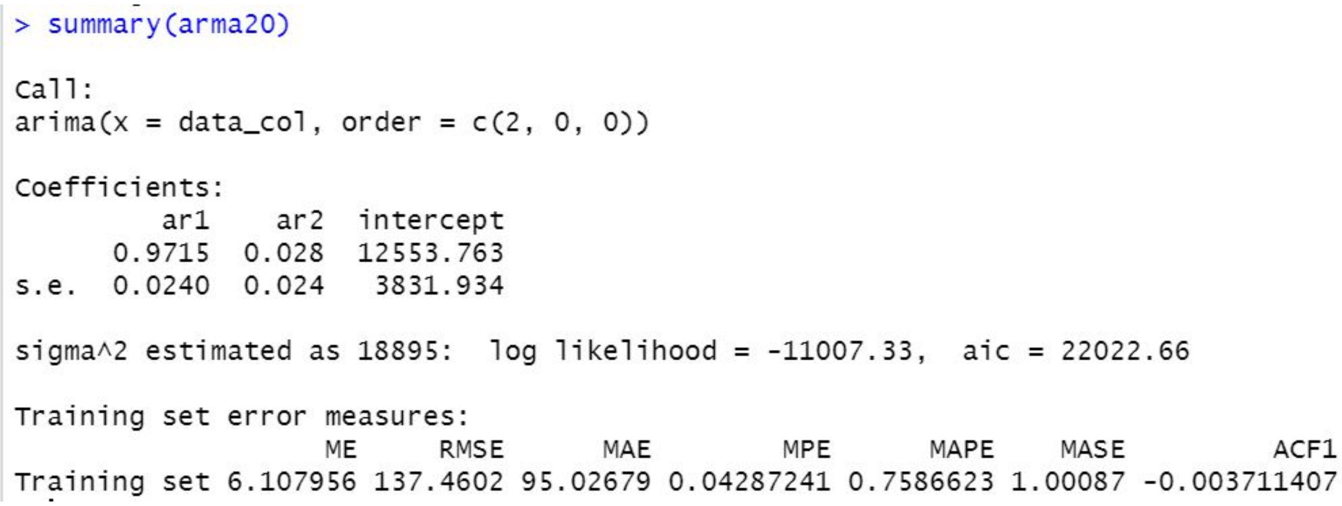
Compare the performance of different ARMA/ARIMA models using evaluation metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Akaike Information Criterion (AIC). Select the best-performing model based on these metrics. The summary of each model is represented below:

Summary for ARIMA(2,0,2)



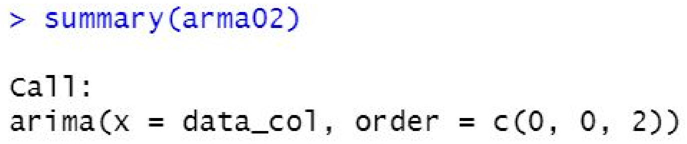
```
> summary(arma22)

Call:
arima(x = data_col, order = c(2, 0, 2))

Coefficients:
         ar1      ar2      ma1     ma2  intercept
      1.8727  -0.8727  -0.9040  0.0446  13008.531
s.e.  0.1404   0.1403   0.1417  0.0249   3784.977

sigma^2 estimated as 18870:  log likelihood = -11006.15,  aic = 22024.3

Training set error measures:
                   ME     RMSE      MAE        MPE      MAPE      MASE        ACF1
Training set 5.071757 137.3671 94.89002 0.03444162 0.7572249 0.9994296 -0.00217846
```

Summary for ARIMA(2,0,0)

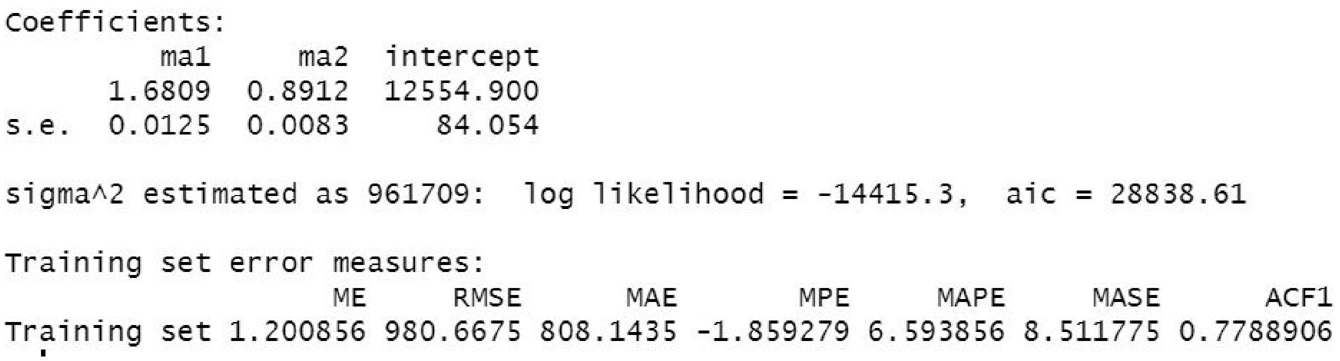


```
> summary(arma20)

Call:
arima(x = data_col, order = c(2, 0, 0))

Coefficients:
         ar1    ar2  intercept
      0.9715  0.028  12553.763
s.e.  0.0240  0.024   3831.934

sigma^2 estimated as 18895:  log likelihood = -11007.33,  aic = 22022.66

Training set error measures:
                   ME     RMSE      MAE        MPE      MAPE    MASE         ACF1
Training set 6.107956 137.4602 95.02679 0.04287241 0.7586623 1.00087 -0.003711407
```

Summary for ARIMA(0,0,2)



```
> summary(arma02)

Call:
arima(x = data_col, order = c(0, 0, 2))
```



```
Coefficients:
         ma1     ma2  intercept
      1.6809  0.8912  12554.900
s.e.  0.0125  0.0083     84.054

sigma^2 estimated as 961709:  log likelihood = -14415.3,  aic = 28838.61

Training set error measures:
                   ME     RMSE      MAE       MPE     MAPE     MASE      ACF1
Training set 1.200856 980.6675 808.1435 -1.859279 6.593856 8.511775 0.7788906
```

Interpretations from ARIMA models summary

When applying the ARIMA (2,0,2) model for time series data, the summary statistics typically include the coefficient values, standard errors, sigma squared, log-likelihood, AIC, and training set error measures.

Interpreting these summary statistics can provide insights into the adequacy of the ARIMA (2,0,2) model for the time series data, as well as the properties of the residuals and the model's goodness of fit.

- The coefficient values represent the autoregressive (AR) and moving average (MA) parameters that govern the behavior of the time series data. These values can be used to interpret the strength and direction of the relationships between past and future values of the time series data.
- The standard errors provide information about the variability of the coefficient estimates. Smaller standard errors indicate greater precision in the estimation of the coefficients.
- The sigma squared value represents the estimate of the variance of the residuals. This value can be used to assess the amount of unexplained variation in the time series data that the ARIMA (2,0,2) model does not account for.
- The log-likelihood value represents the maximum likelihood function value for the estimated model. This value can be used to compare the goodness of fit of different models, with higher values indicating better fits.
- The AIC value measures the relative quality of the model, taking into account the goodness of fit and the complexity of the model. Lower AIC values indicate better models.
- The training set error measures, such as mean absolute error (MAE), root mean squared error (RMSE), and mean absolute percentage error (MAPE), provide information about the accuracy of the model in predicting the time series data. These measures can be used to compare the performance of different models and identify potential improvement areas.

Overall, the summary statistics for the ARIMA (2,0,2) model can be used to assess the adequacy of the model for the time series data, to identify potential areas for improvement, and to compare the performance of different models.

A similar interpretation can easily be made for ARIMA(2,0,0) and ARIMA(0, 0, 2) models

ARCH & GARCH Modelling for y

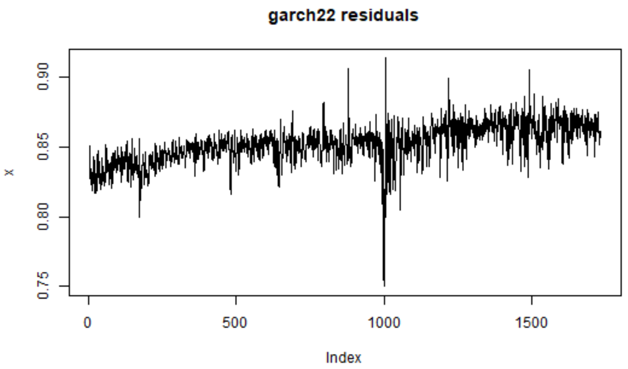
Similar to the ARMA/ARIMA modeling, we have followed the same steps to do the modeling of ARCH/GARCH:

We have used two different variations of the GARCH models:

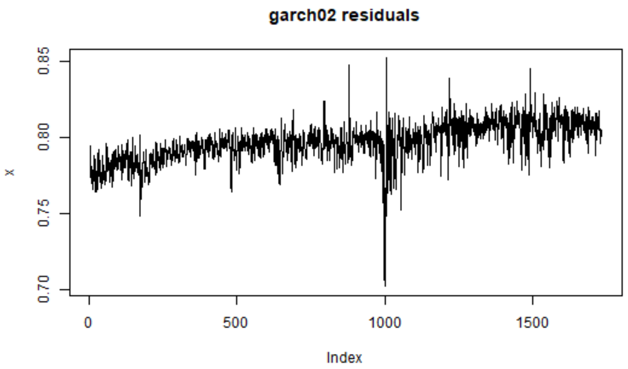
- GARCH(2,2)
- GARCH(0,2)

GARCH(2,2), GARCH(0,2)

Plots for residuals



(a) GARCH(2,2)



(b) GARCH(0,2)

- **The statistics of residuals are shown in quartile form:** When the statistics of residuals are shown in quartile form, it typically means that the residuals are being summarized using quartiles, which are a type of descriptive statistic that divides the data into four equal parts. Showing the statistics of residuals in quartile form can provide a useful summary of their distribution and can be used to identify potential outliers or skewness in the data.

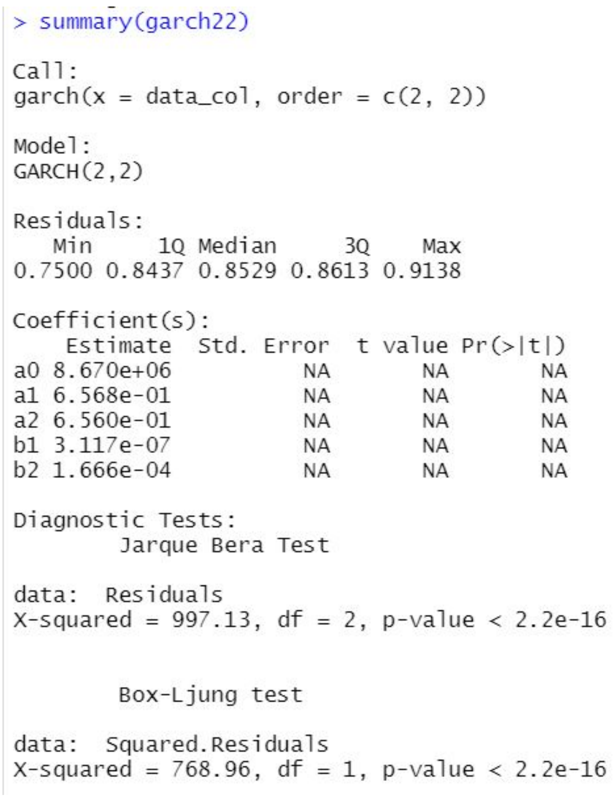
- **For each of the coefficients, the point estimates are shown:** In time series analysis, point estimates are often used to represent the values of the coefficients of a model. These coefficients represent the effect of various factors on the time series data, such as lagged values or exogenous variables. The point estimates of the coefficients provide valuable information about the relationship between the time series data and the factors that influence it. They can be used to make predictions or to understand the underlying patterns in the data, such as the degree of autocorrelation or seasonality.

- **Jarque Bera Test for Diagnostic Testing:** The Jarque-Bera test is a statistical test used for diagnostic testing in time series analysis to assess whether the residuals of a model have a normal distribution. The test is based on the skewness and kurtosis of the residuals, which are measures of the shape of the distribution.

  The Jarque-Bera test can be used to assess the adequacy of a model's fit to the data by checking whether the residuals are normally distributed or not. Suppose the p-value of the test is less than a chosen significance level, typically 0.05. In that case, it suggests that the residuals are not normally distributed, and the model may need to be revised or adjusted.

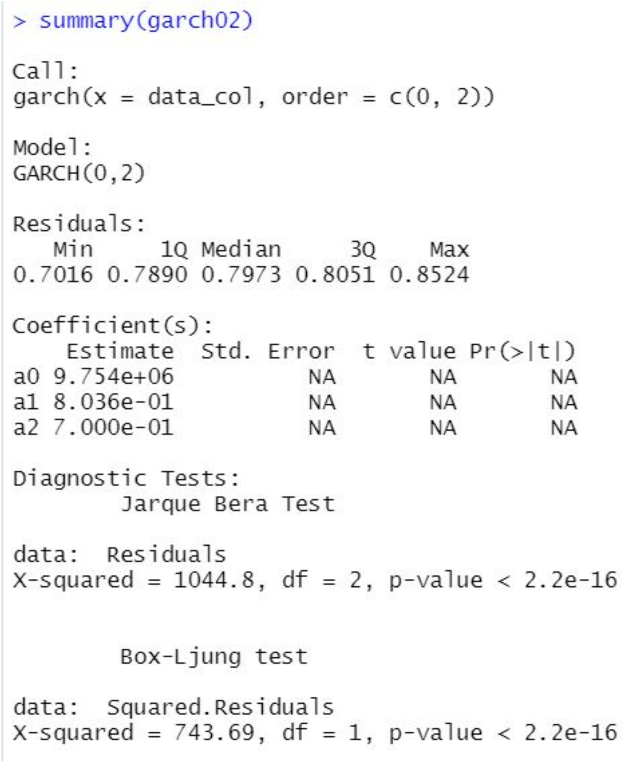
  The Jarque-Bera test is commonly used in time series analysis to ensure that the assumptions of normality are met for the model's residuals, which is important for making accurate forecasts and reliable statistical inferences.

Summary for GARCH(2,2) & GARCH(0,2)



```
> summary(garch22)

Call:
garch(x = data_col, order = c(2, 2))

Model:
GARCH(2,2)

Residuals:
   Min     1Q Median     3Q    Max 
0.7500 0.8437 0.8529 0.8613 0.9138 

Coefficient(s):
    Estimate  Std. Error  t value Pr(>|t|)
a0 8.670e+06          NA       NA       NA
a1 6.568e-01          NA       NA       NA
a2 6.560e-01          NA       NA       NA
b1 3.117e-07          NA       NA       NA
b2 1.666e-04          NA       NA       NA

Diagnostic Tests:
        Jarque Bera Test

data:  Residuals
X-squared = 997.13, df = 2, p-value < 2.2e-16


        Box-Ljung test

data:  Squared.Residuals
X-squared = 768.96, df = 1, p-value < 2.2e-16
```

(a) GARCH(2,2)



```
> summary(garch02)

Call:
garch(x = data_col, order = c(0, 2))

Model:
GARCH(0,2)

Residuals:
   Min     1Q Median     3Q    Max 
0.7016 0.7890 0.7973 0.8051 0.8524 

Coefficient(s):
    Estimate  Std. Error  t value Pr(>|t|)
a0 9.754e+06          NA       NA       NA
a1 8.036e-01          NA       NA       NA
a2 7.000e-01          NA       NA       NA

Diagnostic Tests:
        Jarque Bera Test

data:  Residuals
X-squared = 1044.8, df = 2, p-value < 2.2e-16


        Box-Ljung test

data:  Squared.Residuals
X-squared = 743.69, df = 1, p-value < 2.2e-16
```

(b) GARCH(0,2)

Interpretations from GARCH models summary

When applying the GARCH (2,2) model for time series data, the summary statistics typically include the coefficient values, residual χ^2 value, p-value, and degrees of freedom (df) for both the conditional variance equation and the conditional mean equation. In addition, the summary statistics may also include information about the squared residuals, such as their mean, standard deviation, and autocorrelation.

Interpreting these summary statistics can provide insights into the adequacy of the GARCH (2,2) model for the time series data and the properties of the residuals and the conditional variances.

- The coefficient values for the conditional variance equation represent the autoregressive and moving average parameters that govern the volatility of the time series data. These values can be used to interpret the strength and direction of the relationships between past conditional variances and future conditional variances.
- The residual χ^2 value represents the sum of the squared standardized residuals, which is used to assess the goodness of fit of the GARCH (2,2) model. A low residual χ^2 value suggests that the model fits the data well, while a high value suggests that the model may need to be revised.
- The p-value represents the probability of observing a test statistic as extreme or more extreme than the one observed, assuming the null hypothesis is true. In the GARCH (2,2) modeling context, a low p-value indicates that the coefficient is statistically significant and provides evidence against the null hypothesis of no relationship.
- The degrees of freedom (df) represent the number of observations minus the number of parameters estimated in the model. This value calculates the residual χ^2 value and the associated p-value.
- The summary statistics for the squared residuals can be used to assess the distribution and autocorrelation of the squared residuals. Normal distribution and no autocorrelation are desirable

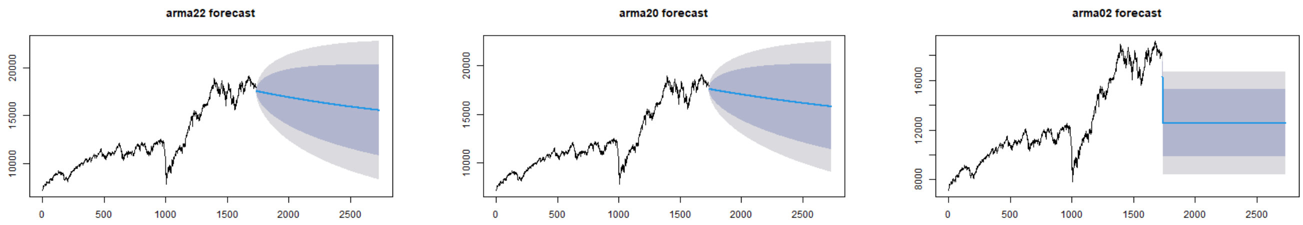
properties for the squared residuals in a GARCH (2,2) model, as they suggest that the model is correctly capturing the volatility of the time series data.

A similar interpretation can easily be made for GARCH(0,2) model.

Forecasting using ARIMA models

Use the selected ARMA/ARIMA model to make future forecasts. This can be done using the forecast() function in R or the predict() function in Python. Visualize the forecasted results and interpret the findings.

Plots for 1000 forecasts using these 3 models



(a) ARIMA(2,0,2) (b) ARIMA(2,0,0) (c) ARIMA(0,0,2)