Fitting ARIMA Models on Hang Seng Index

**NAME**: Mohamed Yusuf

**REG:** I20/1520/2018

a.) You are provided with data for the HSI share index data.

1. Plot the time series of the high, low and closing prices for HSI share index

Chart, line chart

Description automatically generated

# Zoom In to a specific Date

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

The observations in three time series of close value, low value, and high value for the monthly HSI in the time period from January 2006 to January 2020.

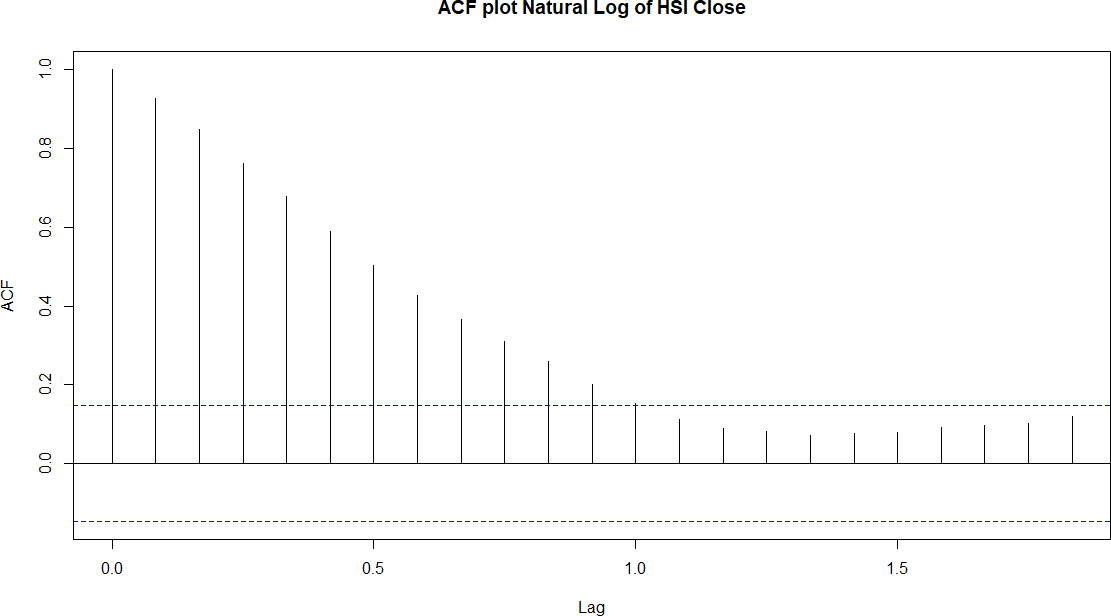
If the stock price forecast is lower than the closing price of the day, it indicates that the stock price may fall in the future

1. Plot the time series natural log of closing prices for HSI share index

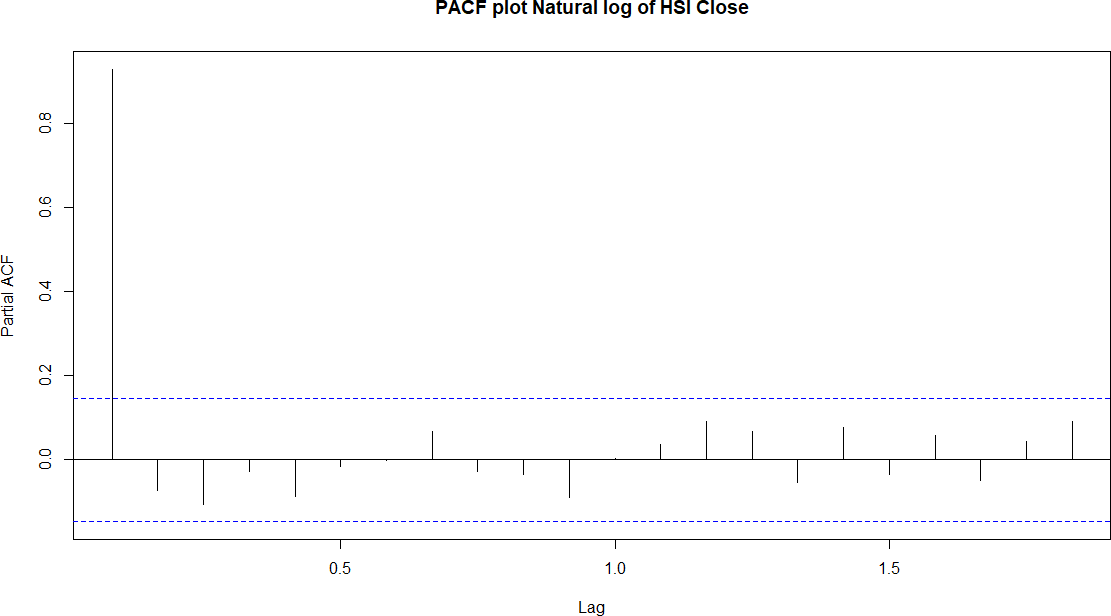
Chart, line chart

Description automatically generated

1. Plot the ACF and PACF of the NATURAL log of closing prices for HSI share index.



**PACF**



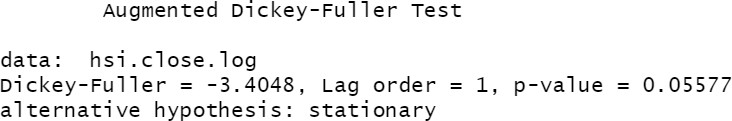
Are the log series stationary?

The ACF chart has a slow decay, indicating that the natural log of the closing price is not stationary.

The PACF curve in the graph above shows a cut off after the first lag, indicating that this is an AR (1) process.

1. Is the data stationary? How do we test for unit root? The data is not stationary. For testing the unit root we use the following tests.

Augmented Dickey Fuller Test

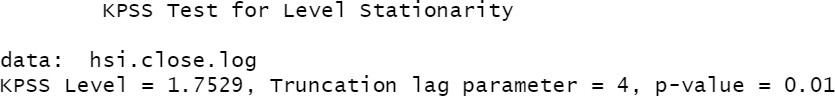


For a time series to be stationary, the p-value obtained from the ADF test must be less than 0.05 or 5%.

If the p-value is more than 0.05 or5%, the time series is said to have a unit root, indicating that it is a non-stationary process.

According to the results of our test, the p-value for the series is0.0557, which is greater than the 0.05 significance level. The null hypothesis is not rejected, thus we conclude that the series has a unit root and is not stationary.

Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test



The null hypothesis is that x is stationary at its current level.

We can see from the output that the KPSS test had a p-value of 0.01, which is less than 0.05, thus we reject the null hypothesis that the series is stationary.

b) Using HSI closing price data, explain and use the Box and Jenkins approach on how to fit an ARIMA model. Clearly explain and illustrate Model identification, Parameter estimation, Diagnostic checking and finally Forecasting.

1. Explain and illustrate with a diagram the Box and Jenkins approach for fitting ARIMA models

Stage 1

Identification

No

Repeat

Stage 2

Estimation

Stage 3

Diagnostic checking

Is model

Adequate?

Forecasting

Yes

# Box and Jenkins approach for fitting ARIMA models

* 1. **Identification.**

# A class of basic ARIMA models is chosen using data plots, autocorrelations, partial autocorrelations, and other information. This entails estimating acceptable p, d, and q values.

# Estimation.

It involves estimation of parameters of the different models (i.e., the coefficients).

# Diagnostic Checking.

The autocorrelations of the residual series are used to check for faults in the fitted model.It also determines whether model estimated is adequate.

1. Using Box-Jenkins approach fit an Integrated Autoregressive Moving Average model ARIMA (p, d, q) for HIS closing price data.

The model ARIMA (1, 1, 1) may be the best fit, based on our findings in the ACF/PACF section. With the forecast package, creating an ARIMA model is simple; we simply call the function 'arima' and enter our parameters.

We'll utilize the 'auto.arima' built-in function in forecast to find the best parameters for our model.

We get that the best ARIMA model is achieved with parameters p=0, d=1, q=0. We will compare ARIMA (0,1,0) and ARIMA (1,1,1)

**ARIMA (0,1,0)**



**ARIMA (1,1,1)**



The Akaike information criterion (AIC) score is a good indicator of the ARIMA model accuracy. The lower the AIC score, the better the model performs.

The AIC of ARIMA (0,1,0) is 3092.36, while the AIC of ARIMA (1,1,1) is 3096.1.

We choose the model with the lower AIC, which is ARIMA (0,1,0), to proceed to forecasting.

1. Perform diagnosis for each model fitted using the Ljung–Box–Pierce Q-statistic, or any other appropriate methods. Interpret.

The Ljung-Box test determines if a time series has autocorrelation.

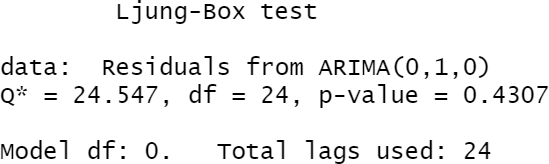
The following hypothesis are used in the Ljung-Box test:

*H0*: The residuals are independently distributed.

*HA*: The residuals are not independently distributed;

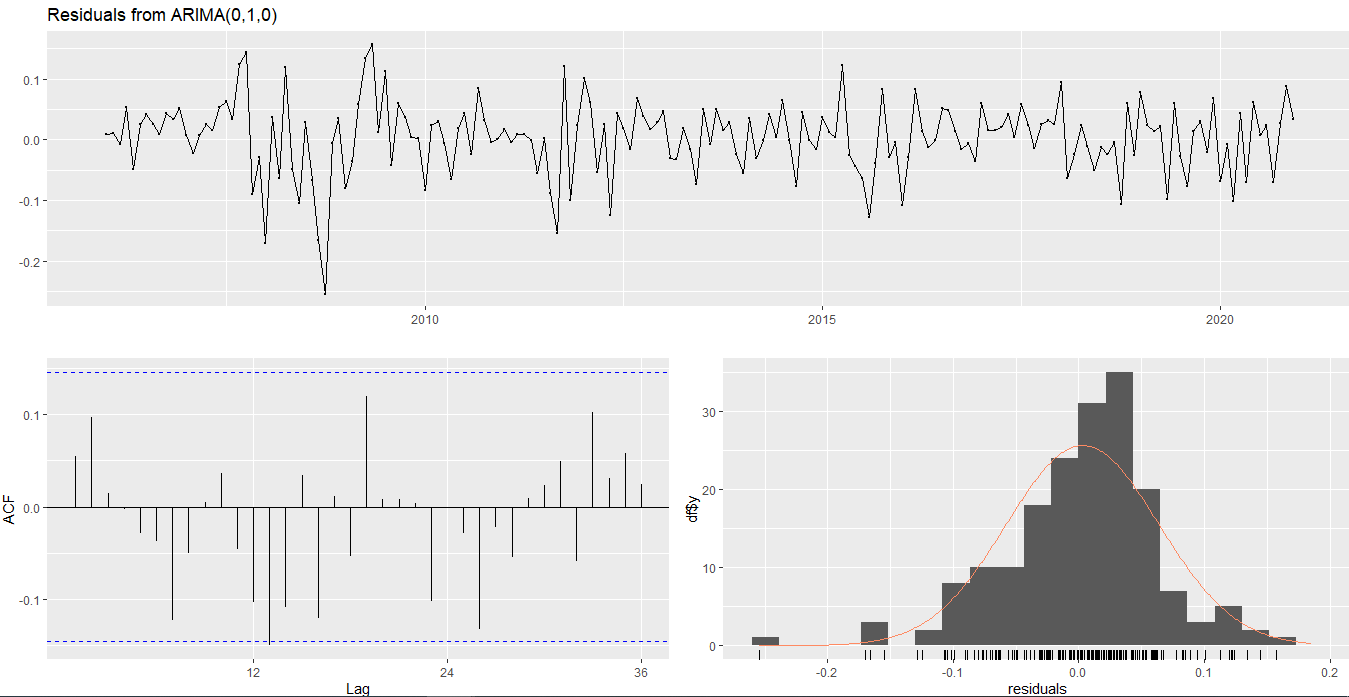
The residuals are independent if the p value is greater than 0.05, which is what we want for the model to be correct.

# ARIMA (0,1,0)

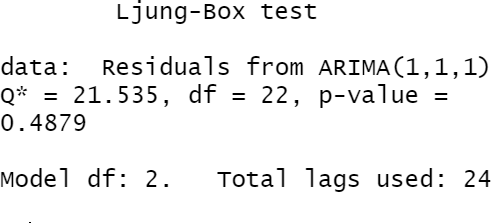


The test's p-value is 0.4307, which is significantly higher than 0.05. As a result, the null hypothesis of the test is not rejected, and we conclude that the data values are independent.

# Examine Residuals from ARIMA (0,1,0)

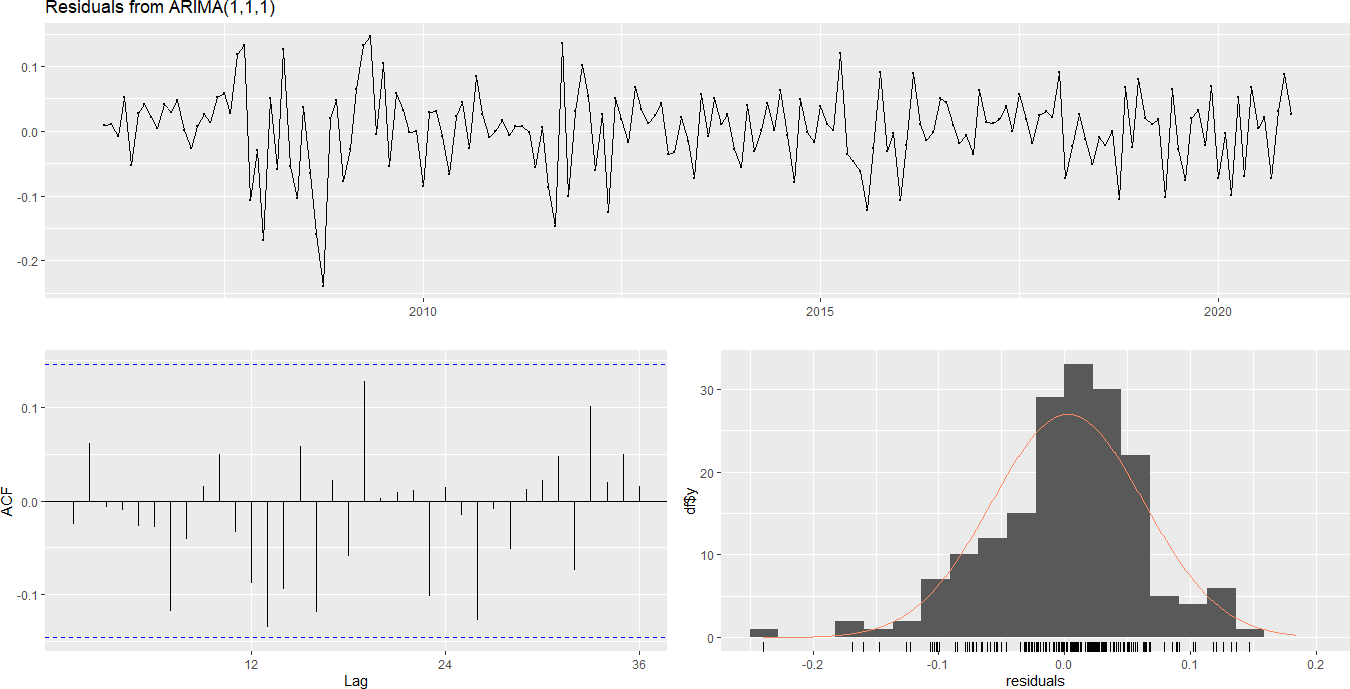


# ARIMA (1,1,1)



The L-Jung Box test has a p-value of 0.4879, which is higher than 0.05. As a result, the null hypothesis of the test is not rejected, and we conclude that the data values are independent.

# Examine Residuals from ARIMA (1,1,1)



The variation of the forecast mistakes appears to be approximately consistent throughout time, as seen by the time plot of the forecast errors. The forecast errors are generally normally distributed, and the mean appears to be near to zero, according to the time series' histogram. As a result, the forecast errors should be regularly distributed with a mean of zero and a constant variance.

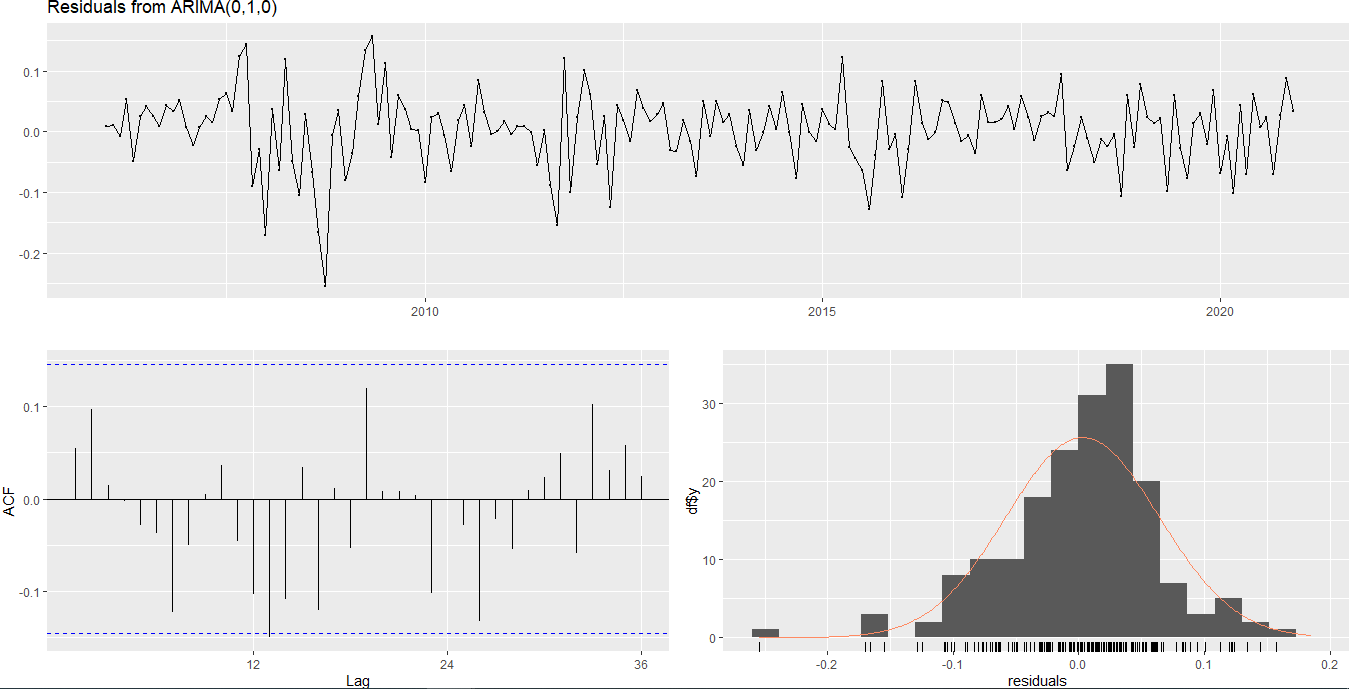
iv) Perform Forecasting and illustrate by plotting the forecast and 95% confidence intervals

Chart, line chart

Description automatically generated

Prediction intervals are represented as a shaded region, with the intensity of the color denoting the likelihood associated with the interval.

**INTRODUCTION**



A time series is a collection of observations of well-defined data items obtained through repeated measurements over time. For example, measuring the value of retail sales each month of the year would comprise a time series. This is because sales revenue is well defined, and consistently measured at equally spaced intervals. Data collected irregularly or only once are not time series.

An observed time series can be decomposed into three components: the trend (long term direction), the seasonal (systematic, calendar related movements) and the irregular (unsystematic, short-term fluctuations. The categories of time series are; the time domain vs the frequency domain. The time domain approach tells us how does what happened today affect what will happen tomorrow?

The frequency domain approach tells us what is the economic cycle through periods of expansion and recession? Other categories are the univariate and multivariate, linear and nonlinear, discrete and continuous. In general, a time series is affected by four components which are; Trend, seasonal, cyclical and irregular components.

TREND

The general tendency of a time series to increase, decrease or stagnate over a long period of time

SEASONAL VARIATION

This component explains fluctuations within a year during the season usually caused by climate and weather conditions, customs, traditional habits.

CYCLICAL VARIATION

This component describes the medium-term changes caused by circumstances which repeat in cycles. The duration of a cycle extends over a long period of time.

IRREGULAR VARIATION

Irregular variation in a time series are caused by unpredictable influences which are not regular and do not repeat in a particular pattern. These variations are caused by incidences such as war, strike, earthquake, flood, revolution.

There is no defined statistical technique for measuring random fluctuations in a time series

Considering the effects of these four components, two different types of models are generally used for a time series.

1. Additive model.
2. Multiplicative model.

Time Series Analysis considers the fact that data points collected over time may have an internal structure (such as autocorrelation, trend, or seasonal fluctuation) that needs to be considered. Time series variables such as exchange rates have inconsistent behavior, making forecasting illogical. Despite these claims, many multinational firms, foreign exchange dealers, exporters, importers, and speculators continue to hedge their positions based on anticipated prices based on ex-post data.

These hedging decisions assume that patterns in ex-post data exist and that these patterns, at least in the near run, provide a signal of future exchange rate movement. If such patterns exist, modern mathematical techniques can theoretically be used to solve them.

The Autoregressive AR and the Moving Average MA are the two processes that make up the ARMA model. We can represent the level of current observations of a series Xt as a function of the level of its lagged observations. The AR model can depict this.

We may also simulate that the observations of a random variable at time t are affected not just by the shock at time t, but also by the shocks that occurred before time t, using the ARMA model.

The MA model can illustrate this.

The goal of Autoregressive Integrated Moving Average (ARIMA) models is to characterize the current behavior of variables using linear relationships with their previous values. It has an Integrated (I) component (d) that represents the amount of differencing that must be done on the series in order for it to become stationary. The ARIMA's second component is an ARMA model for the series that has been differentiated to make it stationary. The AR and MA components of the ARMA component are further divided, as explained above. To estimate the values of the orders of the AR and MA processes, the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) are used. The data is analyzed using the statistical program R.

**OBJECTIVES**

The purpose of this research is to do statistical analysis on data from the Hang Seng index between the high, low, and close prices. The data's attributes are discussed, and fundamental time series techniques are used to analyze the data. Some of the graphical tools used to study the series include plots of the series, the autocorrelation function, and the partial autocorrelation function. In order to create reasonable projections from the model, we also want to fit ARIMA models to the data.

**LITERATURE REVIEW**

Box and Jenkins: Time series analysis and, forecasting and control

Robert H. Shumway: Time series analysis and its application

**CONCLUSION**

Time series analysis is one of the most important aspect of data analytics for any large organization as it helps in understanding seasonality, trends, cyclicality and randomness in the sales and distribution and other attributes. These factors help companies in making a well-informed decision which is highly crucial for business. Time series analysis and modeling is a well-known mathematical and statistical technique for uncovering hidden characteristics in time-dependent data. ARIMA modeling is one of the most widely used time series approaches. We looked at the low, high, and closing prices of Hang Seng index data in this experiment. The logarithms of the rate returns are employed in the study instead of the actual data due to the nature of the data. This is because the logarithm of the returns has advantageous statistical features for analysis. Time series forecasting focuses on predicting future trends of data. However due to seasonal effects and other factors in real world market data are hard to be stationary

SMA443

PROJECT

TIME SERIES ANALYSIS OF HSI DATA

AUTHOR: MOHAMED YUSUF

REG NO: I20/1520/2018

DATE:23RD APRIL 2022