AI6123 - Time Series Assignment 2

G2304225K Chong Wai Kiat

Abstract

The main objective of this project is to fit the best model to the daily historical Apple stock price data which included Open, Low, High, Close, Adjusted close, and Volumes from February 1st, 2002 to January 31st, 2017 extracted from Yahoo Finance website. Several analyses including ACF, PACF, and model comparisons are done in this report.

1. Dataset

In this project, the first task is to analyze the daily historical Apple stock prices data which included Open, Low, High, Close, Adjusted close, and Volumes from February 1st, 2002 to January 31st, 2017 extracted from the Yahoo Finance website. The main objective is to discover an interesting trend in Apple stock prices over the past 15 years and to design and develop the best-suited model for forecasting.

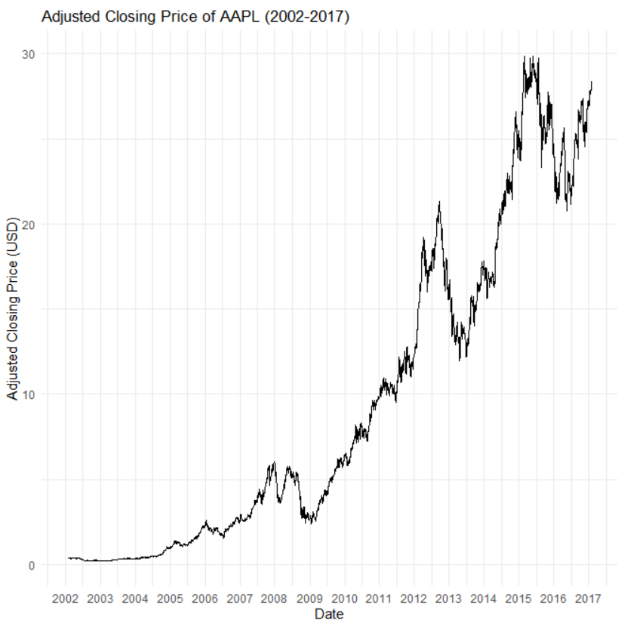

Figure 1: AAPL data from 01-02-2002 to 31-01-2017

Table 1 shows the statistics of all features included in the original dataset. Since Adjusted Close price (Adj.Close) is more complex measurement that modifies the closing price of the stock to reflect all the dividends, stock splits and etc compared to the Close (Close) price which is only the raw price in the stock market, we will choose Adjusted stock price as our main objective or forecasting values. Based on the statistic shown, we can observe that there is a potential trend in the original data as the minimum (0.1983) and maximum (29.8638) gap has a huge gap and the mean value (9.4649) is more skewed to the minimum value. Hence, further analysis is performed in Figure 1. Figure 1 clearly shows that the original data has an increasing trending component. Due to some of the data are not recorded, the data is transformed into the monthly format.

Col	Open	High	Low	Close	Adj.Close	Volume
Min.	0.2320	0.2355	0.2271	0.2343	0.1983	3.93e+07
1st Qu.	1.9700	2.0114	1.9395	1.9638	1.6625	2.50e+08
Median	6.6360	6.6916	6.5273	6.6205	5.6049	4.17e+08
Mean	10.783	10.8876	10.6673	10.7798	9.4649	5.30e+08
3rd Qu.	19.182	19.3208	18.9762	19.1426	16.6883	6.95e+08
Max.	33.615	33.6350	32.8500	33.2500	29.8638	3.37e+09

Table 1: Summary Statistics for Stock Prices

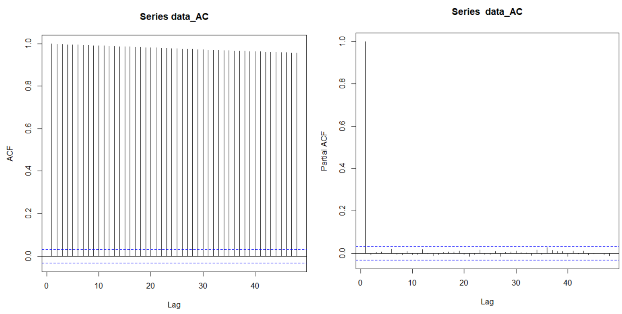

Figure 2: ACF and PACF of original dataset

Parameter	Value
Dickey-Fuller	-2.2214
Lag Order	1
p-value	0.4846
Alternative Hypothesis	Stationary

Table 2: Augmented Dickey-Fuller Test on original data_AC

To find out the seasonal components in the data, seasonal decomposition is performed with the function stl() and the result is shown in Figure 3. It shows an increased seasonality variance

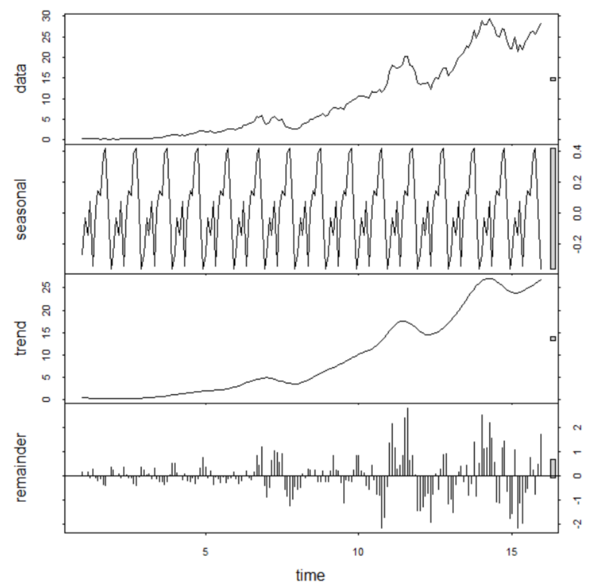

Figure 3: Seasonal Decomposition of monthly data

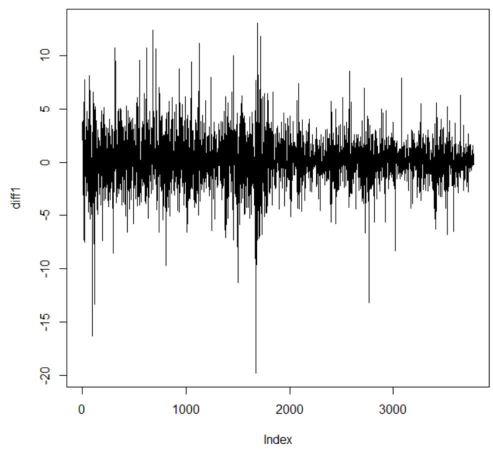

Figure 4: Data after First Order Differencing and Log Transformation

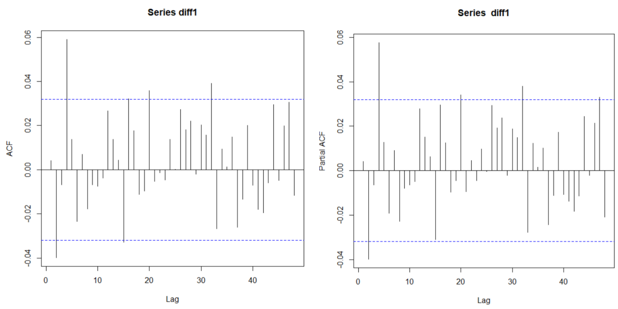

Figure 5: ACF and PACF after Log Transformation and First-Oder Differencing

from the data plot and a repetitive pattern in the remainder plot in the Figure. Based on the ACF and PACF plot in Figure 2, we also observed that the data is non-stationary and non-linear. The Augmented Dickey-Fuller (ADF) test shown in Table 2, the p-value is 0.4846 which is higher than 0.05 and further proves the data is non-stationary.

2. Transformation

Log transformation can stabilize the variance in the data to avoid fluctuations of variance across different timestamps. log() is used to transform the data which is default with base 2. Subsequently, the condition variance of the return of a financial asset is often adopted as a measure of the risk of the asset. Hence, calculating the return is the key component. Differencing diff() is used to calculate the return with the multiplication of 100 to show the percentage of return. It also removes the trending component and makes the data stationary. Figure 4, it shows that the data is more stationary. To further prove this, the ADF test shows a p-value of 0.01 shown in Table 3 which is lower than 0.05 and is evident the data is stationary.

However in Figure 4, we observed a volatility pattern in the data which is called volatility clustering. Although the ACF and PACF after First Order differencing in Figure 5 and ADF test has shown the data is stationary and serially uncorrelated in a linear relationship but it could be not an independent relationship. Hence, we further analyze the ACF and PACF to check the independence of the data. If the absolute and square of the log-return data die down quickly, it means the data is linearly independent. Conversely, if they do not die down quickly while the data shows a volatility clustering pattern, the data is not linearly independent. Figure 6 and Figure 7, it clearly show that the log-return data is not independent and identically distributed. Furthermore, since it is not independent and identically distributed, the data should not be Gaussian distributed although log transformation is applied to stabilize the variance. To verify this, QQplot is used to plot the diagram of the data.

Parameter	Value
Dickey-Fuller	-14.928
Lag Order	15
p-value	0.01
Alternative Hypothesis	Stationary

Table 3: Augmented Dickey-Fuller Test Results after Log Transformation and First Order Differencing

As in Figure 8, it shows the data has a heavy tail distribution and is slightly left skewed. To further prove our analysis, kurtosis, and skewness are performed to further our analysis. Table 4 shows the data tested has 5.441384 of value in kurtosis which proves the data is heavy tail distributed. The skewness value -0.1923283 is also evident that the data is left-skewed. In short, the data is not independent and identically distributed but only uncorrelated. The data is also not Gaussian distributed but a left-skewed heavy tail distributed. Thus, ARCH or GARCH models are best suited for this data.

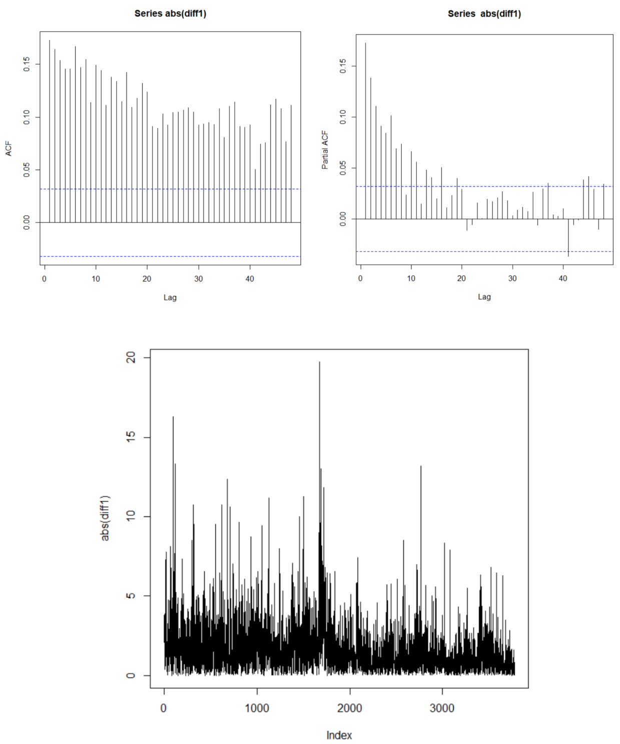

Figure 6: ACF and PACF of Absolute Log-Return data

Table 4: Skewness and Kurtosis of the data

Statistic	Value
Skewness	-0.1923283
Kurtosis	5.441384

3. Model Comparison

To find the best parameters for the model, extended autocorrelation function (EACF) is used to find the best setting of the model. Based on the daily log-return data's EACF shown in Figure 9, it suggests the parameters of (0,2), (1,2), (2,2), and (3,3). Based on the EACF of absolute log-return data shown in Figure 10, it suggests the parameters of (1,1,), (2,2), and (3,3). Based on the EACF of square log-return data shown in Figure 11, it suggests the parameters of (1,1), (2,2), and (3,3). Hence, 5 parameters are experimented and compare the AIC and adequacy of the model. Table 5, it clearly shows that all the Garch models are adequate. However, the Garch (1,1) model has the lowest AIC value (16203.4). To further diagnose the Garch (1,1) model, a few of tests are performed to examine the model. Figure 12

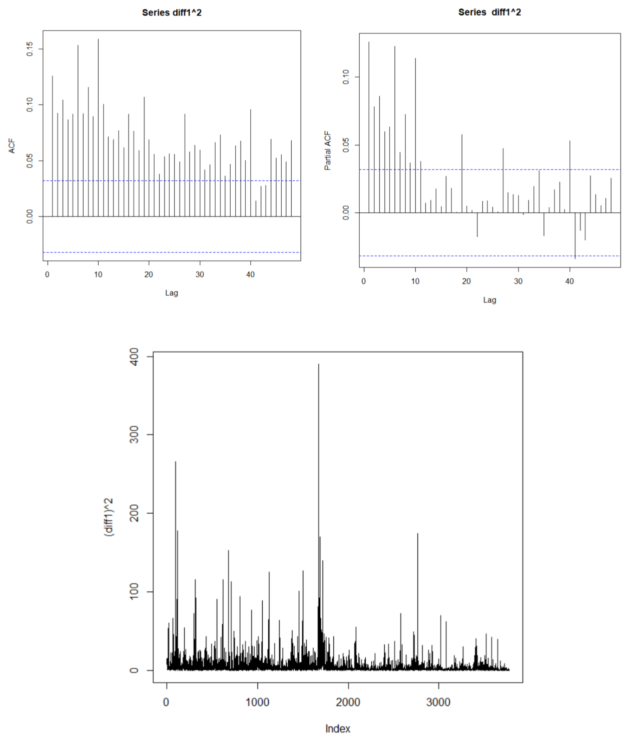

Figure 7: ACF and PACF of Square of Log-Return data

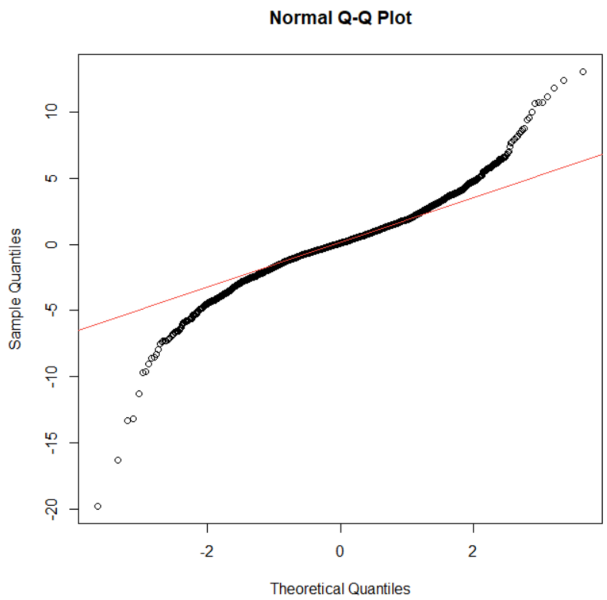

Figure 8: QQplot of the data

shows the QQplot of the Garch (1,1) model. It shows that the model has strike a balance for both left and right tails to avoid the heavy tail distribution on the model. Figure 13 shows the residuals checking of the Garch (1,1) model. It clearly shows that the model does not show any specific patterns on the model which shows the model is independent and stationary. Hence, this project will be using Garch (1,1) for the ARMA model.

AR/MA	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	o	x	o	x	o	o	o	o	o	o	o	o	o	o
1	x	x	o	x	o	o	o	o	o	o	o	o	o	o
2	x	x	o	x	o	x	o	o	o	o	o	o	o	o
3	x	x	o	o	o	x	o	o	o	o	o	o	o	o
4	x	x	x	x	o	x	o	o	o	o	o	o	o	o
5	x	x	x	x	x	x	o	o	o	o	o	o	o	o
6	x	x	o	x	x	x	o	o	o	o	o	o	o	o
7	x	x	x	x	x	x	x	o	o	o	o	o	o	o

Figure 9: EACF of the log-return data

AR/MA	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	x	x	x	x	x	x	x	x	x	x	x	x	x	x
1	x	o	o	o	o	o	o	x	x	o	o	x	o	o
2	x	x	o	o	o	o	o	o	x	o	o	x	o	o
3	x	x	o	o	o	o	o	o	o	o	o	o	o	o
4	x	x	o	o	o	o	o	o	o	o	o	o	o	o
5	x	x	x	x	x	o	o	o	o	o	o	o	o	o
6	x	x	x	x	x	x	o	o	o	o	o	o	o	o
7	x	x	x	x	x	x	x	o	o	o	o	o	o	o

Figure 10: EACF of absolute log-return data

AR/MA	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	x	x	x	x	x	x	x	x	x	x	x	x	x	x
1	x	o	o	o	o	x	x	o	o	x	o	o	o	o
2	x	x	o	o	o	x	o	o	o	x	o	o	o	o
3	x	x	o	o	o	x	o	o	o	x	o	o	o	o
4	x	x	x	x	o	x	o	o	o	x	o	o	x	o
5	x	x	x	x	x	o	o	o	o	x	o	x	o	o
6	x	x	x	x	x	x	o	o	o	x	o	o	x	o
7	x	x	x	x	x	x	x	o	o	x	o	o	x	o

Figure 11: EACF of squared log-return data

To choose the best parameters for the model, the ARMA model's order also experimented with the Garch (1,1) model. Before experimenting with different parameters of the ARMA model, the data is split into train, and test set. The test set is a 30-day actual data for the further evaluation of the forecasting performed by the best-suited model. Hence, finding out the best model required finding the best ARMA model with the Garch (1,1) model for the forecasting. Table 6 shows various AIC models with the combination of Garch (1,1) model forecasted on the test data. It clearly shows that the ARMA (2,2) has the lowest AIC value. Hence, ARMA(2,2)

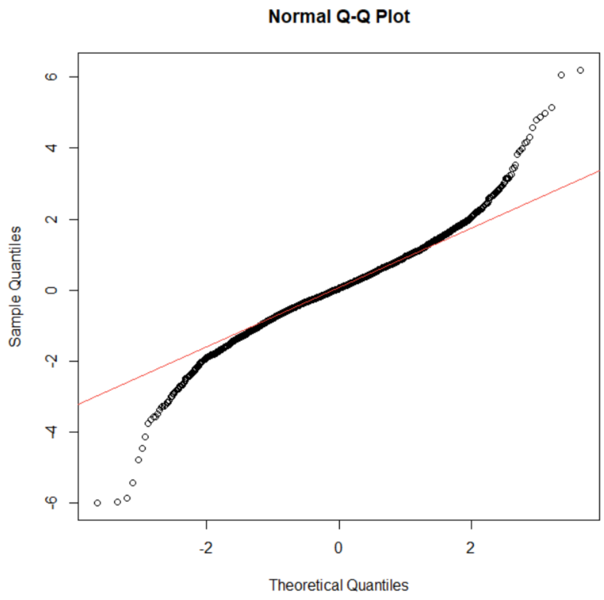

Figure 12: QQplot of Garch (1,1) model

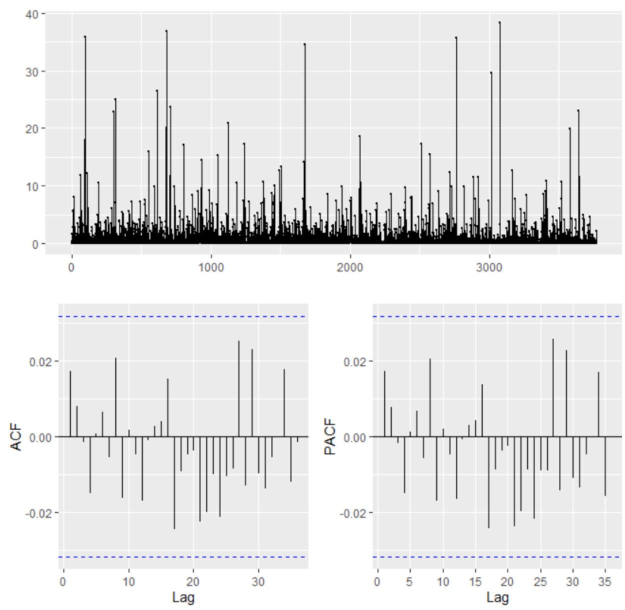

Figure 13: Residual Check of Garch (1,1) model

and Garch(1,1) models are used for the forecasting.

4. Forecast

In forecasting, the model is fitted on both the training and test set to show the model performance and prediction on the future trend. Figure 14 shows that the model has a good prediction on the test data which has successfully predicted the trend.

Model	Test	X-squared	df	p-value	AIC
GARCH(0,2)	Jarque-Bera	3649.8	2	2.2e-16	16545.36
	Box-Ljung	0.64531	1	0.4218	
GARCH(1,1)	Jarque-Bera	1980.7	2	2.2e-16	16203.4
	Box-Ljung	1.1388	1	0.2859	
GARCH(1,2)	Jarque-Bera	3971.9	2	2.2e-16	16592.34
	Box-Ljung	0.58491	1	0.4444	
GARCH(2,2)	Jarque-Bera	3919.6	2	2.2e-16	16576.08
	Box-Ljung	0.28295	1	0.5948	
GARCH(3,3)	Jarque-Bera	3778.9	2	2.2e-16	16506.54
	Box-Ljung	0.070822	1	0.7901	

Table 5: Diagnostic Tests and AIC Values for Various GARCH Models

Model	AIC
ARMA(0,0)	4.301526
ARMA(0,1)	4.301778
ARMA(1,0)	4.301786
ARMA(1,1)	4.302272
ARMA(1,2)	4.302639
ARMA(2,1)	4.302739
ARMA(2,2)	4.300350

Table 6: AIC Values for ARMA Models with Garch (1,1) model

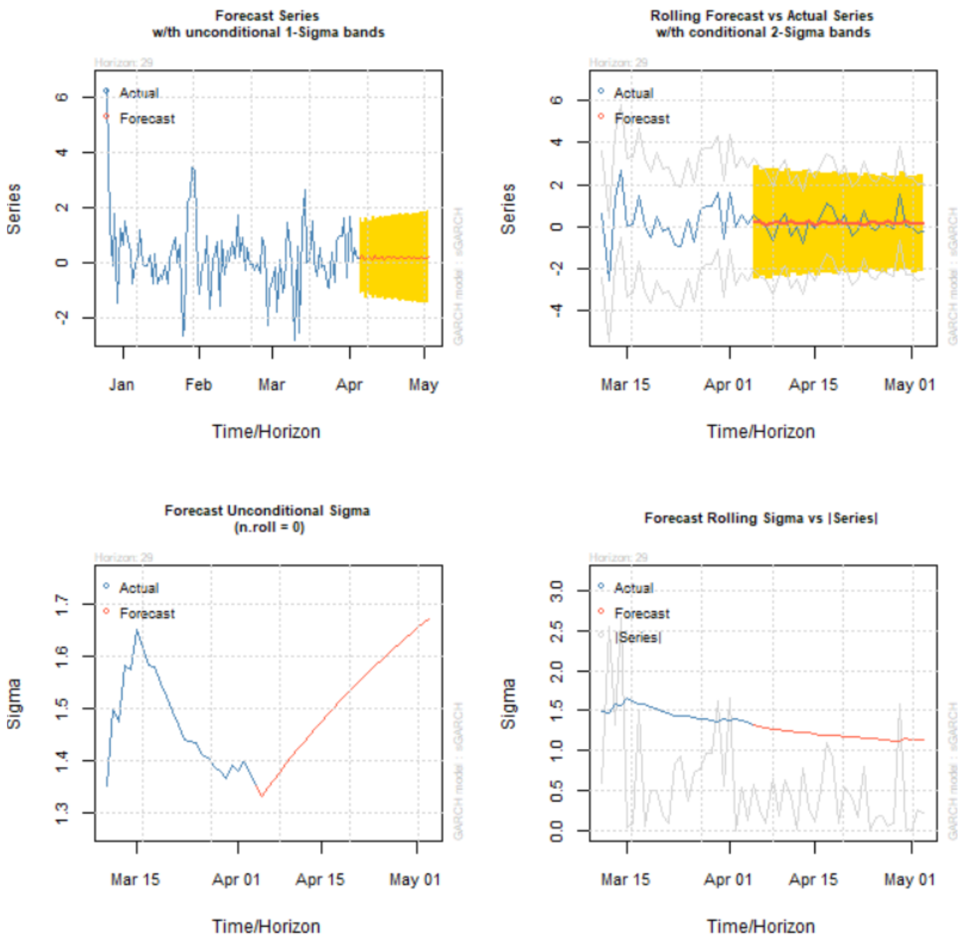

Figure 14: Forecasting of the ARMA (2,2) with Garch (1,1) model