MM905: Financial Econometrics

Report: Forecasting Stock returns and volatility through Time series models

Environment used – R Studio

Prepared by
Rishabh Gupta
(MSc in Quantitative Finance – 21/22)

Contents

List of Figures and tables 3

1. **Introduction** 4
2. **Data Analysis** 4
   - 2.1 Testing for Stationarity 5
   - 2.2 Data transformation 5
   - 2.3 Distribution of returns 5
   - 2.4 Value at Risk 6
3. **ARIMA Modelling** 6
   - 3.1 Determining a possible candidate model 6
   - 3.2 Fitting ARIMA model 7
   - 3.3 White Noise check 7
   - 3.4 Box cox transformation 8
   - 3.5 Outlier removal 8
   - 3.6 Refitting ARIMA model and White noise check 8
   - 3.7 Results overview 9
   - 3.8 ARIMA Out of sample forecasts 10
4. **GARCH modelling** 10
   - 4.1 GARCH Results overview 10
   - 4.2 GARCH Out of sample forecasts 11
5. **References** 12

List of figures and tables

Figures

1. PepsiCo closing price chart 4
2. Boxplot and QQ plot of log returns 5
3. ACF and PACF of log returns 6
4. ACF and QQ plot of ARIMA (1,0,1) residuals 7
5. Final residuals plot after fitting ARIMA (1,0,1) 8
6. Forecast of log returns for the next 50 weeks 10
7. Residuals after fitting GARCH (1,1) 11
8. Forecast of volatility for the next 50 weeks 11

Tables

1. ACF of closing price 5
2. VaR estimates 6
3. AIC estimates of ARIMA 7
4. ARIMA model final results 9
5. Confidence intervals of ARIMA coefficients 9
6. GARCH (1,1) results 10

1. Introduction

Estimating forecasts of stock prices along with their volatility through time series algorithms has become one of the most popular topics nowadays. This report aims to **predict future stock returns and volatility** of the ticker **PepsiCo Inc (PEP)**. PepsiCo is an American company engaged in the business of manufacturing food, beverages and snacks. The software used for programming and data analysis is R Studio. For forecasting the prices, we'll use ARIMA model, which is based on the idea that future values of a time series can be predicted solely based on its own past values (lagged price and lagged error terms).

Closing price has been used for fitting time series models. We'll mainly focus on the four main tasks below for the forecasting exercise:

1. Stationarity check through statistical tests and plot observation. Differencing, if needed
2. Deciding the order of AR and MA models to be fitted.
3. Fitting the ARIMA model and white noise check for errors
4. Estimating volatility through GARCH engine.

2. Data Analysis

The timeframe of our data ranges from 3rd Jan, 2000 to 31st March, 2022(Weekly data). Let's first look at the time plot of our price data.

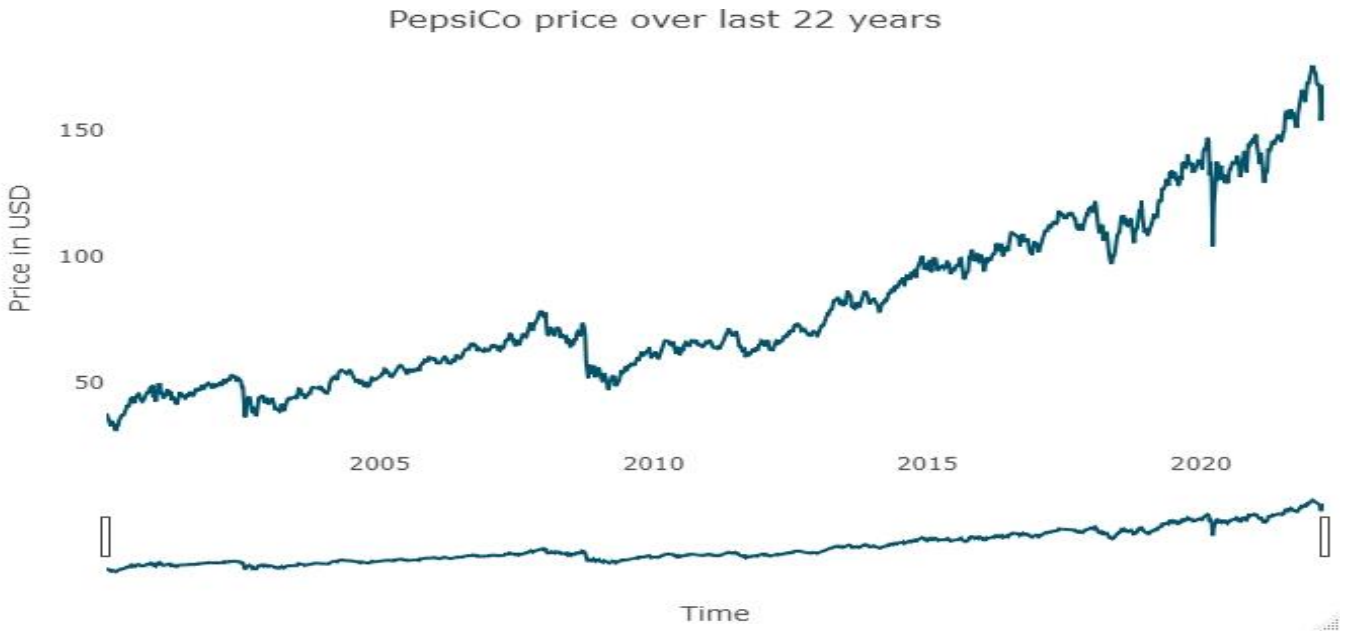

Figure 1: PepsiCo Closing price chart

The plot is showing an increasing trend with significant declines due to 2008 and 2020 crisis and Ukraine Russia war **(Russia is a key market for Pepsi)**.

Whenever we have values in our data where the current value depends on the past value or the past values of the error term, we use linear time series models like Box Jenkins ARIMA model. **The data has to be stationary (completely random) to model it with ARIMA**. Stationarity is achieved when the data is free from trend and it has constant mean and variance.

2.1 Testing for Stationarity – We've applied **Augmented Dicky Fuller test** on price data. Its p value is 0.957, which is greater than 5%, meaning we fail to reject the Null and conclude that our price data is Non stationary.

If we look at the **Autocorrelation function (ACF)** of the stock price, we observe that it decays very slowly and is much closer to 1 at all the lags, which also indicates presence of trend in the data.

Lag	2	4	6	8	10	12	14	16	18	20	22	24	26	28	30
ACF	0.99	0.98	0.97	0.96	0.95	0.94	0.93	0.92	0.91	0.90	0.89	0.88	0.88	0.87	0.86

Table 1: ACF of Closing price

2.2 Data transformation - As the stock price do not follow normal distribution, we've first applied **log transformation and then applied differencing** of the log transformed data, which is called **log returns**.

Now we again test this log return data for stationarity through **ADF test** and we get a **p value of less than 0.01**, meaning our data is now completely random and we are ready to fit ARIMA model.

2.3 Distribution of returns

Now let's look at the below plots of log returns and check if they follow normal distribution or not.

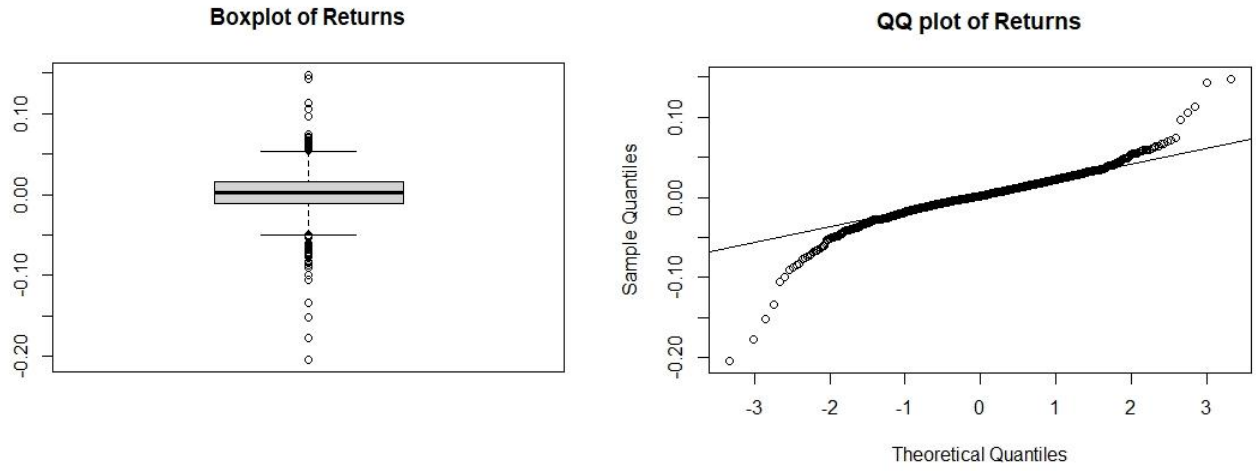

Figure 2: Boxplot and QQ plot of Log returns

From the above QQ plots, we observe that

- The dots in the tails depart upward and downward significantly from the line (If the data would have followed normal distribution the points would fall much near the straight line),
- The actual quantiles in the tails are much higher than the theoretical quantiles, meaning that
- the returns have fat tails and higher peak than the normal distribution.

The boxplot shows a high number of outliers both on positive and negative side. Also, the Kurtosis of the data is coming out to be 11.93, which is much higher than 3. This shows that the returns follow Leptokurtic distribution.

Normality test for Returns – We'll also check the normality through a statistical test known as **Shapiro Wilk Test**. We are getting a p value of less than 5 percent, meaning the data is NOT normally distributed.

2.4 Value at Risk

As stock returns follow stochastic behavior, so Normal distribution does not hold good in real life. The empirical past data is not satisfied by the normal distribution.

For computing VaR through **Parametric approach (Delta normal method)**, our data should follow normal distribution. But our returns data is not normally distributed as proved above.

We get the below values of VaR at difference confidence levels on our data:

Method applied	95% Confidence level	99% Confidence level
Historical	-0.03830509	-0.07461611
Gaussian(Parametric)	-0.04154512	-0.05929241
Modified	-0.04259898	-0.1227768

Table 2: VaR estimates

- In **PepsiCo's current case**, the most relevant is the 99% confidence level VaR through **Cornish Fisher method (this method can take non normal distribution into account)**
- **Interpretation for 99% VaR (Modified)** - If I have a portfolio of worth 100 Million USD, I'm 99% confident that the loss will not increase beyond 12.27 million USD in one week with a 1% chance of being exceeded.

3. ARIMA Modelling

3.1 Determining a possible candidate model

The way to identify the order of ARIMA model is to observe the Autocorrelation and Partial autocorrelation (PACF) graph of the series. Let's have a look at the ACF and PACF plot of weekly log returns of PepsiCo

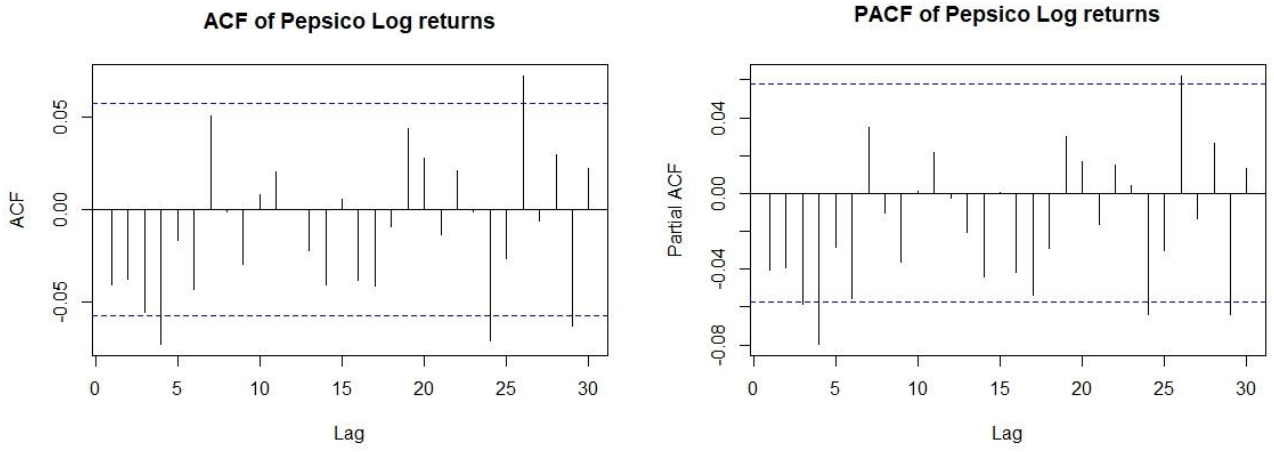

Figure 3: ACF and PACF of Log returns

Observations:

- There are some significant lags (lines cutting off the horizontal dash line) in both the plots
- Both ACF and PACF plots are exhibiting decay in a sinusoidal manner till lag 20 and then expansion also in a sinusoidal manner.

- We also cannot see any seasonality in the data as there are no vertical lines cutting off at a repeated number pattern of lag.
- It is difficult to decide whether it is an ONLY AR or ONLY MA model. It seems like a mixed ARMA model.

Moreover, among the three parameters of ARIMA – p (AR parameter), d (number of differencing) and q (MA parameter), **d will be considered as 0 as log returns already include impact of first order differencing.**

3.2 Fitting ARIMA model

We are getting ARIMA (1,0,1) as the **best model** by applying the auto.arima function. It searches for the best combination and then gives the best one based on lowest AIC. The results are as under:

ARIMA Order	AIC
ARIMA(2,0,2)	-5178.37
ARIMA(0,0,0)	-5169.55
ARIMA(1,0,0)	-5169.47
ARIMA(0,0,1)	-5169.64
ARIMA(1,0,2)	-5179.7
ARIMA(0,0,2)	-5170.25
ARIMA(1,0,1)	**-5181.43**
ARIMA(2,0,1)	-5179.72
ARIMA(2,0,0)	-5169.27

Table 3: AIC estimates of ARIMA

Thus, we are getting the **least AIC for the combination (1,0,1),** meaning the current returns depend on last one week's returns and last one week's error terms.

3.3 White Noise check

After fitting ARIMA (1,0,1) on the log returns, we are getting the below plots for residuals:

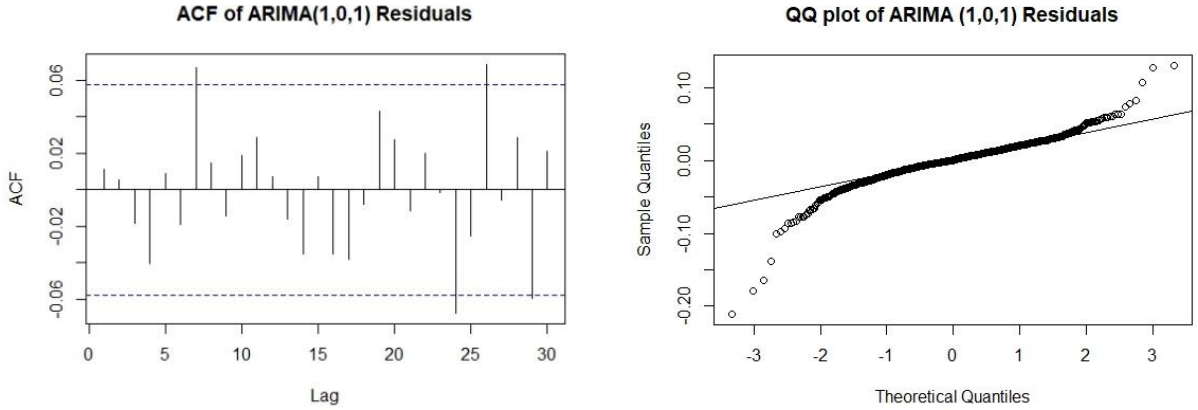

Figure 4: ACF and QQ plot of ARIMA (1,0,1) residuals

We can clearly see that the errors are **not pure white noise** as there are **some significant lags** in the ACF plot and is also **not independent and identically distributed**. There is some information left in the error terms which needs to be pulled out.

- ***So for finding a good fit model, lets transform our data to normal distribution or remove outliers and then apply the ARIMA again.***

3.4 Box Cox transformation

- Box cox method requires input data to be free of negative values. Now as per an article (Link) we can add a constant to our data to deal with this. We tried adding the largest negative value to the data so as to make all the values positive.
- We then applied the Box cox model on our new data and got the transformed data. But the transformation did not work as the model has significantly modified the values of returns. They range from -78 % to -57% which does not hold good in real life.

3.5 Outlier removal

- We firstly applied **tsclean** function, which significantly transformed our data but still there are many outliers left.
- We then applied second method where we eliminated data points from our data whose value is less than "Q1(first quartile) – 1.5(Inter quartile range) and whose value is greater than "Q3(Third quartile) + 1.5(Inter quartile range). After removing those data points, we got a new dataset free from major outliers. **Total number of 55 data points have been removed as outliers.**

3.6 Refitting ARIMA model and White noise check

We again got the same order as the best fit model i.e. ARIMA (1,0,1). We are getting the below plot of residuals:

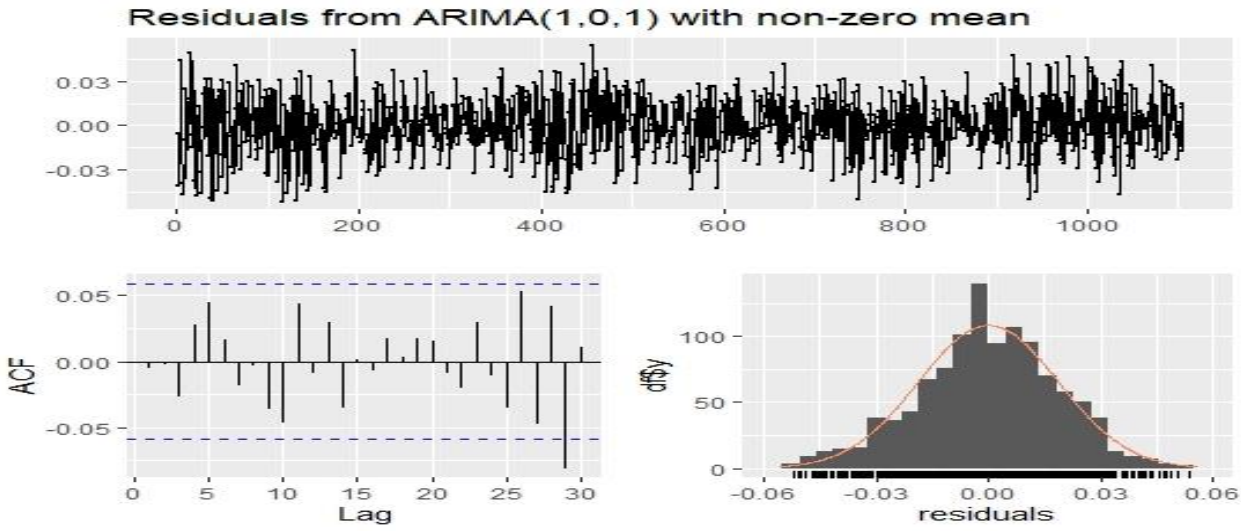

Figure 5: Final residuals plot after refitting ARIMA (1,0,1)

The error terms are completely random now and the ACF plot also looks much better with only one significant lag. Regarding the Ljung box test, we are getting a p value of 0.8612, meaning the error terms are now completely random and IID.

3.7 Results overview

We are getting the below equation from the Final ARIMA (1,0,1) model:

$$\hat{Y}_t = 0.0019 + 0.8746\, Y_{t-1} - 0.9328\, e_{t-1} + e_t$$

Particulars	Value	Interpretation
AR1 coefficient	0.8746	Also known as ϕ_1. The current week's stock return depends on past week's return by a factor of 0.8746.
MA1 coefficient	-0.9328	Also known as θ_1. The current week's return also depends on past week's error term with a factor of 0.9328
Mean (Constant)	0.0019	Average value of Returns i.e. $E(X_t)$
$Sigma^2$	0.0003494	Variance of noise
Log likelihood	2830.96	the logarithm of the probability of the observed data coming from the estimated model. It is used for comparing different model. The higher the number the better the model
AIC	-5653.92	A metric for evaluating strength of a model. The lower the AIC, the better the model. Calculated as: -2(Log likelihood) + 2(p+q+k+1) -2(2830.96) + 2(1+1+1+1) = -5653.92 *(k means number of constant)*
BIC	-5633.89	It is also a metric for model selection, but AIC is preferred more than BIC. Calculated as: AIC + [log (sample size) – 2]*(p+q+k+1) -5653.92 + [log (1105) – 2]*(1+1+1+1) = -5633.89

Table 4: ARIMA model final results

Confidence intervals of coefficients at 95% confidence level (2.5% in each tail)

	Lower limit [Coefficient – (1.96*Standard error)]	Upper limit [Coefficient + (1.96*Standard error)]
AR(1)	0.784234254	0.964933024
MA(1)	-1.000263946	-0.865274343
Intercept	0.001256366	0.002449858

Table 5: Confidence intervals of ARIMA coefficients

Results Overview

- Pepsico returns have high dependency on its past week's returns and also past week's error terms.
- If we want to express the **equation in terms of price**, we can also write it as:

$$\hat{Y}_t - Y_{t-1} = 0.0019 + 0.8746\,(Y_{t-1} - Y_{t-2}) - 0.9328\, e_{t-1} + e_t$$

3.8 ARIMA Out of sample forecasts

We are getting the below plot for next 50 weeks forecast of PepsiCo returns:

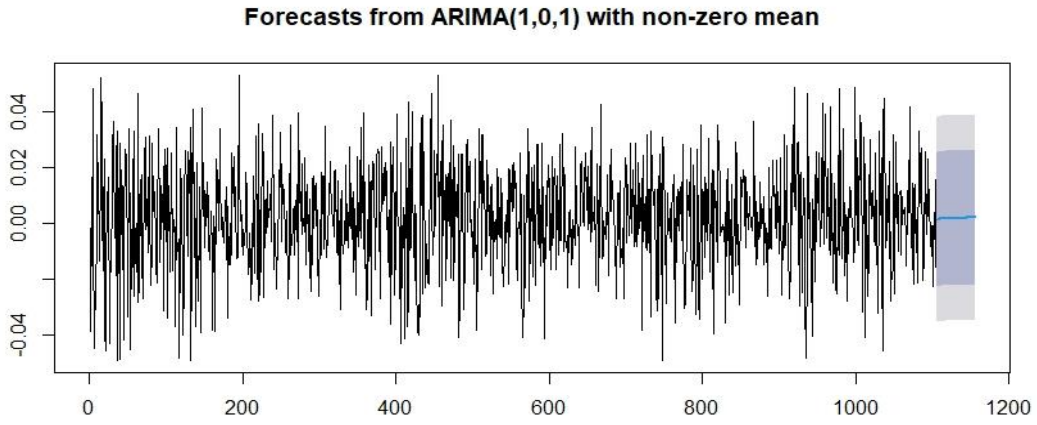

Figure 6: Forecast of log returns for the next 50 weeks

- The dark blue shaded area represents 80% confidence interval of the forecasted returns and the outer light blue shows 95% confidence interval. Point forecasts represents the mean value.

4. GARCH modelling

- In ARIMA modelling, we were taking the conditional variance. We assumed variance as constant. ARIMA model assumes that conditional variance is constant.
- But when it comes to **real world financial data, generally the variance is not constant**. There may be periods where the volatility is very high and periods where it is low. This gives us the need to model the volatility. GARCH is a model to model volatility.
- We'll take the returns and model GARCH on the returns data

ARCH effect check – The first step is to see if the ARCH effect is present in the data, means whether the variance is changing across different points of time. We tested this by applying **Archtest function (Lagrange multiplier method)** on **squared residuals** of our ARIMA (1,0,1) model. We got p value less than 5%, meaning there is some volatility clustering in the residuals.

4.1 GARCH results overview

After applying the GARCH (1,1) on the returns data, we get the below parameters:

Parameter	Estimate	Std error	t value (Estimate/Std error)	Pr(>|t|)
Mu	0.001908	0.000508	3.7557	0.000173
AR(1)	- 0.051101	0.030545	-1.6730	0.094332
Omega or α_0	0.000007	0.000001	6.4000	0.000000
Alpha1 (ARCH parameter) α_1	0.037557	0.004246	8.8450	0.000000
beta1 (GARCH parameter) β_1	0.942490	0.007674	122.8143	0.000000

Table 6: GARCH (1,1) results

Since p value of Omega, Alpha 1 and Beta 1 is less than 5%, it means that these parameters are statistically significant

GARCH equation

$$\sigma^2_t = 0.000007 + 0.037557\ X^2_{t-1} + 0.942490\ \sigma^2_{t-1}$$

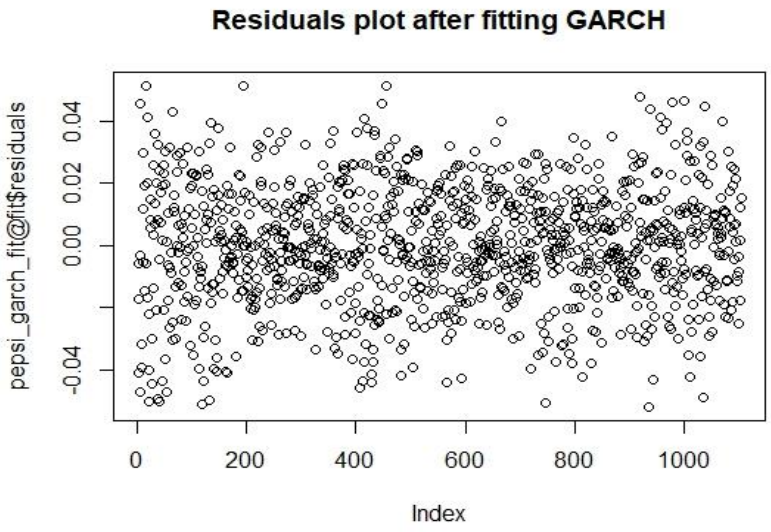

Figure 7: Residuals after fitting GARCH (1,1)

When we look at the above plot, we can say that error terms are now completely random with constant mean and variance.

White noise check - When we perform the **Jarque Bera** test on **residuals**, we get P value > 5%, meaning error terms are normally distributed. But when we perform the **Ljung box** test on **Squared residuals**, The p value comes to less than 5 percent, which means ***there is some autocorrelation left in the error terms and they are not independent.***

4.2 GARCH Out of sample forecasts

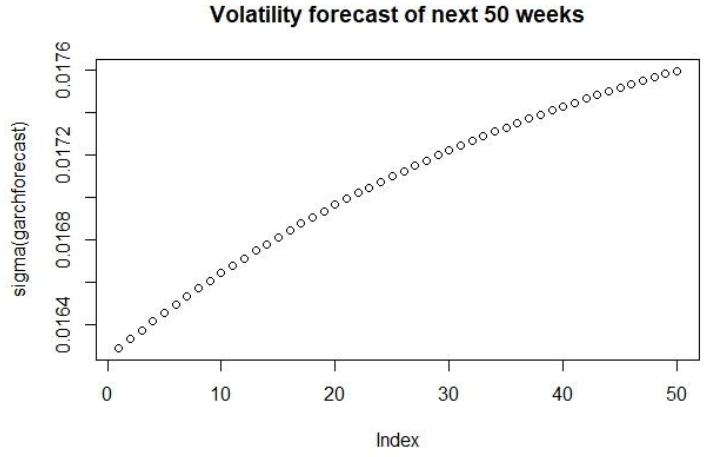

Figure 8: Forecast of volatility for the next 50 weeks

5. References

Time Series Analysis with Applications in R, Jonathan and Kung

L stern group, Ly Pham, Time series analysis with ARIMA – ARCH/GARCH model in R

Otexts, 8.7 – ARIMA modelling in R,
URL: https://otexts.com/fpp2/arima-r.html

Otexts, 3.3 – Residual diagnostics,
URL - https://otexts.com/fpp2/residuals.html

Otexts, 8.6 – Estimation and order selection,
URL - https://otexts.com/fpp2/arima-estimation.html

Selva Prabhakaran, 2021, ARIMA Model – Complete Guide to Time Series Forecasting in Python
URL - https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/

Roopam upadhayay, Step by step guide
URL - http://ucanalytics.com/blogs/step-by-step-graphic-guide-to-forecasting-through-arima-modeling-in-r-manufacturing-case-study-example/

Subhasree Chatterjee, 2018, Time series analysis using ARIMA modelling in R
URL - https://datascienceplus.com/time-series-analysis-using-arima-model-in-r/

PennState Eberly College of Science, Lesson 3
URL - https://online.stat.psu.edu/stat510/book/export/html/665

Enes Zvornicanin, 2021, Choosing the best q and p from ACF
URL - https://www.baeldung.com/cs/acf-pacf-plots-arma-modeling

RPubs
URL - https://rpubs.com/Sergio_Garcia/managing_financial_data_r

Shreyashi Saha and Sagarnil bose, Forecasting of a Time Series (Stock Market) Data in R
URL - https://stat-wizards.github.io/Forcasting-A-Time-Series-Stock-Market-Data/

Ranjith kumar, 2020, Time Series Model(s) — ARCH and GARCH
URL - https://medium.com/@ranjithkumar.rocking/time-series-model-s-arch-and-garch-2781a982b448

Florian Mudekereza, 2020, Introduction to GARCH in R
URL - https://rpubs.com/florian1/garchmodeling

PennState Eberly College of Science, Lesson 2.1
https://online.stat.psu.edu/stat510/lesson/2/2.1#:~:text=A%20moving%20average%20term%20in,0%20and%20the%20same%20variance

Github
URL - https://matzc.github.io/tmseries/lecture6.html

Duke university, ARIMA models for time series forecasting
URL - https://people.duke.edu/~rnau/411arim.htm

Github project
URL - https://ionides.github.io/531w18/midterm_project/project38/Midterm_proj.html#test-for-residuals

Vamsidhar Ambatipudi, Time series modelling using ARIMA
URL - https://www.youtube.com/watch?v=40e-6TUbIrM&list=PLheFgWAePSu9LiM0jr8Hb_yPQr-uMH-W9&index=8

Yevonnael Andrew, 2020, GARCH models
URL - https://rpubs.com/yevonnael/garch-models-demo

John, 2020, How to eremove outliers in R
URL - https://www.r-bloggers.com/2020/01/how-to-remove-outliers-in-r/

A QQ plot dissection kit
URL - https://seankross.com/2016/02/29/A-Q-Q-Plot-Dissection-Kit.html#:~:text=These%20%E2%80%9Cthin%20tails%E2%80%9D%20correspond%20to,plot%20across%20the%20X%2DY%20diagonal.