Title
Predicting the Log Returns of Illumina, Inc. Stock

Author
Chopra, Shelly

Publication Date
2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

Predicting the Log Returns of Illumina, Inc. Stock

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Applied Statistics

by

Shelly Chopra

2022

ABSTRACT OF THE THESIS

Predicting the Log Returns of Illumina, Inc. Stock

by

Shelly Chopra
Master of Applied Statistics
University of California, Los Angeles, 2022
Professor Hongquan Xu, Co-Chair
Professor Nicolas Christou, Co-Chair

Corporations belonging to the biotechnology sector have the potential to provide investors with high returns; between January 2020 and January 2021, the average share price for European and US biotechnology companies increased at more than twice the rate of the S&P 500, outperforming sister industries such as pharmaceuticals and many other industries which witnessed dwindling returns during the COVID-19 pandemic[1].Thus, this research paper aims to formulate time series models that accurately predict the log returns of the stock price of Illumina, Inc. Seven time series models were formulated, including three naive models, two Autoregressive Integrated Moving Average models *(ARIMA)*, one simple exponential smoothing model, and one *ARIMA-GARCH* model. Among the models, it was determined that the simple exponential smoothing model outperformed the other models on the basis of root mean square error (RMSE). This simple exponential smoothing model was then applied to three competing biotechnology companies to assess its applicability to other companies within the industry.

[1]Cancherini.

The thesis of Shelly Chopra is approved.

Mark Handcock

Frederic Schoenberg

Nicolas Christou, Committee Co-Chair

Hongquan Xu, Committee Co-Chair

University of California, Los Angeles

2022

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

The stock market is frequently perceived to be a sentiment indicator that can impact key economic measures such as gross domestic product (GDP) negatively or positively. From 1928 to 2016, the average annual return from stock investment was approximately 8 percentage points higher than on 3-month Treasury Bills[1], leading to sizeable return gaps between investment in stocks compared to other investment vehicles. In particular, corporations belonging to the biotechnology sector have the potential to provide investors with high returns; between January 2020 and January 2021, the average share price for European and US biotechnology companies increased at more than twice the rate of the S&P 500, outperforming sister industries such as pharmaceuticals and many other industries which witnessed dwindling returns during the COVID-19 pandemic[2]. Thus, a topic of interest that is central to this research paper is formulating models to predict the stock price returns of Illumina, Inc., an American biotechnology company that develops, manufactures, and markets integrated systems for the analysis of genetic variation and biological function. The company provides a line of products and services that serves the sequencing, genotyping and gene expression, and proteomics markets. The company announced its initial public offering on July 28, 2000 at $16.00 and reached its all-time high share price of $524.84 on August 16, 2021.

The Efficient Market Hypothesis states that any new information is immediately re-

[1]Santoli.

[2]Cancherini.

flected in stock prices and thus neither technical nor fundamental analysis can generate excess returns[3]. Thus, rather than reviewing company financials to predict future stock prices, this research paper utilizes time series analysis, namely naive models, Autoregressive Integrated Moving Average Models (denoted ARIMA), Exponential Smoothing techniques, and an ARIMA-GARCH model to predict future stock prices based on daily historical adjusted stock price closing data. The quality of the models is assessed primarily based on the root mean square errors (RMSE)), since it penalizes larger errors more than other accuracy metrics and is expressed in the same unit as the forecasted values; however, other accuracy metrics are also noted, including mean error (denoted ME), mean absolute error (denoted MAE), mean percentage error (denoted MPE), and mean absolute percentage error(denoted MAPE). These other accuracy metrics are not as efficient if extreme values are present.

The computation for all of the aforementioned accuracy metrics can be found in the equations below, where $\hat{x_i}$ reflects the i^{th} predicted value, x_i reflects the i^{th} actual value, and n reflects the total number of observations.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{x_i} - x_i)^2}{n}} \tag{1.1}$$

$$ME = \frac{1}{n} \sum_{i=1}^{n} (\hat{x_i} - x_i) \tag{1.2}$$

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |\hat{x_i} - x_i| \tag{1.3}$$

$$MPE = \frac{100\%}{n} \sum_{i=1}^{n} \frac{(x_i - \hat{x_i})}{x_i} \tag{1.4}$$

$$MAPE = \frac{100\%}{n} \sum_{i=1}^{n} \frac{|x_i - \hat{x_i}|}{x_i} \tag{1.5}$$

[3]https://www.investopedia.com/terms/e/efficientmarkethypothesis.asp

Upon finding the model with the highest prediction accuracy, we apply this model to the stock price data to that of 3 competing biotechnology companies, namely Thermo Fisher Scientific, Agilent Technologies, and Qiagen.

CHAPTER 2

Data

Daily stock price data corresponding to Illumina, Inc. (ticker: ILMN) was scraped from Yahoo! Finance[1] from January 2nd, 2018 through June 2nd, 2022, consisting of 1,112 observations, or stock price data corresponding to 13 complete fiscal quarters[2]. While the data set comprises of six continuous variables (opening price, high price, low price, closing price, trade volume, and adjusted closing price), our response variable is adjusted closing price, which reflects the closing price after adjustments for all applicable stock splits and dividend distributions, adhering to Center for Research in Security Prices (CRISP) standards[3]. In the case of Illumina, Inc., however, the closing price is equal to the adjusted closing price for all observations, as no dividend distributions nor stock splits have occurred during the time frame in question. Prior to formulating models, preliminary data analysis was performed on daily stock price data of Illumina, Inc. to understand high-level metrics and preemptively identify potential modeling challenges. Daily stock close price movement demonstrated record high closing prices of $524.84 on August 16th, 2021 and record low closing prices of $209.20 on March 18th, 2020. The stock close price on June 1st, 2022 is $235.20, which is -$172.25 lower (or -73% lower) year-over-year. Table 2.1 shows the annual stock price highs and lows by year to demonstrate range.

[1] https://finance.yahoo.com/

[2] https://finance.yahoo.com/quote/ILMN/history?p=ILMN

[3] https://help.yahoo.com/kb/SLN28256.html

	2018	2019	2020	2021	June 2022 YTD
Max Adj Closing Price	$367.06	$378.23	$400.74	$524.84	$423.80
Min Adj Closing Price	$209.54	$266.84	$209.20	$347.28	$213.05
Range	$157.52	$111.39	$191.54	$177.56	$210.75
Max vs Min, % Difference	75.2%	41.7%	91.6%	51.1%	98.9%

Table 2.1: Illumina, Inc. - Adjusted Closing Price Ranges

Figure 2.1: Time series data of original, untransformed Illumina, Inc. Stock Price Data

The volatile nature of the company's stock price may suggest non-stationarity, which hinders our ability to make accurate price predictions through time series models and would therefore require transformation of the original data set. This subject is explored in greater depth in the subsequent chapter.

For modeling purposes, the data was subset into training and testing data and corresponds to the time frames in table 2.2; the training set was intended to incorporate the 1st fiscal

quarter of 2018 through the 3rd fiscal quarter of 2021 (85% of the data), while the testing data set was intended to incorporate the 4th fiscal quarter of 2021 through the most recent week of data in June of 2022 (15% of the data).

Type	Observations Frequency	Time Frame
Training	944(85%)	1/2/2018 – 9/30/2021
Testing	168(15%)	10/1/2021 – 6/1/2022

Table 2.2: Training Testing Data Splits

CHAPTER 3

Time Series Features

A time series, defined as a sequence of data points that occur in successive order over some period of time[1], can be decomposed into three parts as defined by the equation below:

$$Y_t = S_t + T_t + R_t, \text{for additive models,} \tag{3.1}$$

and

$$Y_t = S_t \times T_t \times R_t, \text{for multiplicative models} \tag{3.2}$$

In the equation above, S_t represents seasonality, which is defined as cycles that repeat regularly over time; T_t represents trend, which is defined as a pattern in the data that shows the movement to higher/lower values over time; R_t is defined as random, irregular influences on the time series.

A key assumption of time series models is stationarity, which implies that the mean, variance, and covariance between the i^{th} and $(i+m)^{\text{th}}$ term of the time series are constant and not a function of time. As a result, a stationary time series should not exhibit trend nor seasonal components. Figure 3.1 shows the additive decomposition of our time series data.

[1]https://www.investopedia.com/terms/t/timeseries.asp

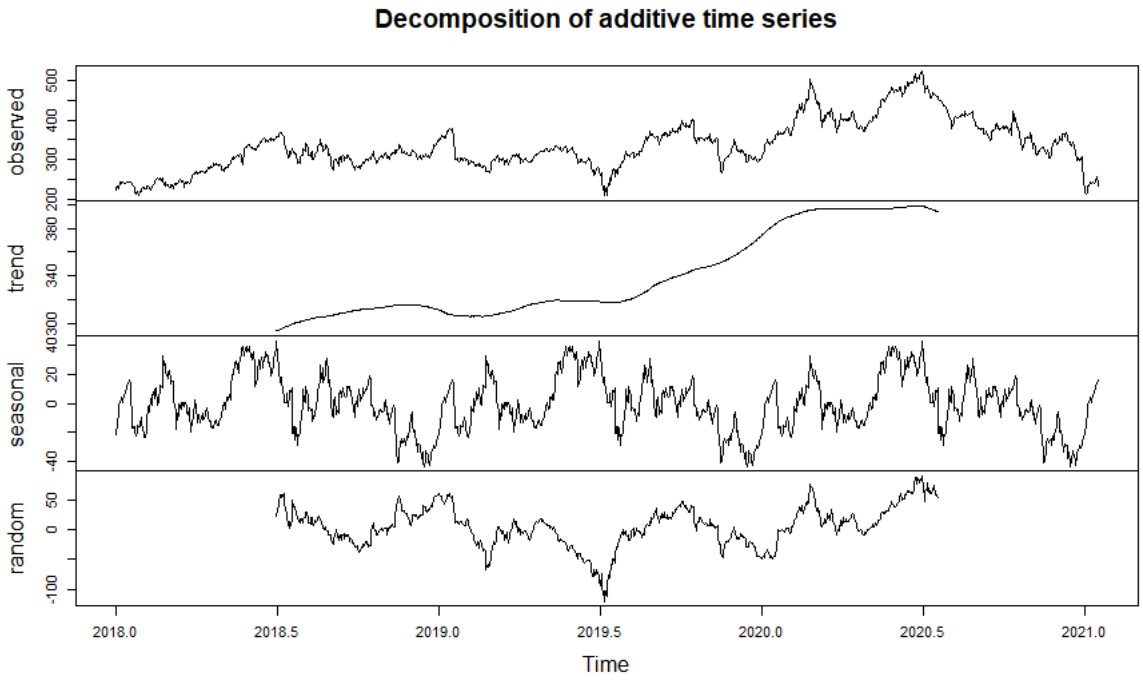

Figure 3.1: Illumina, Inc. - Time series additive decomposition of original stock price data

With respect to trend, the decomposition chart suggests a positive trend in the stock price, particularly from the second half of 2019 through the first half of 2020, followed by a plateau in price until halfway through 2021. With respect to seasonality, we observe seasonal peaks in the middle of each year. Therefore, we predict that the data is non-stationary and will require transformation prior to modeling. The autocorrelation (ACF) plot in figure 3.2 further suggests non-stationary data, since the ACF values are larger and positive (statistically significant at the 95% confidence level, since each bar extends beyond the blue dashed line) and decrease slowly overtime; for stationary data, we expect the ACF to drop to 0 quickly.

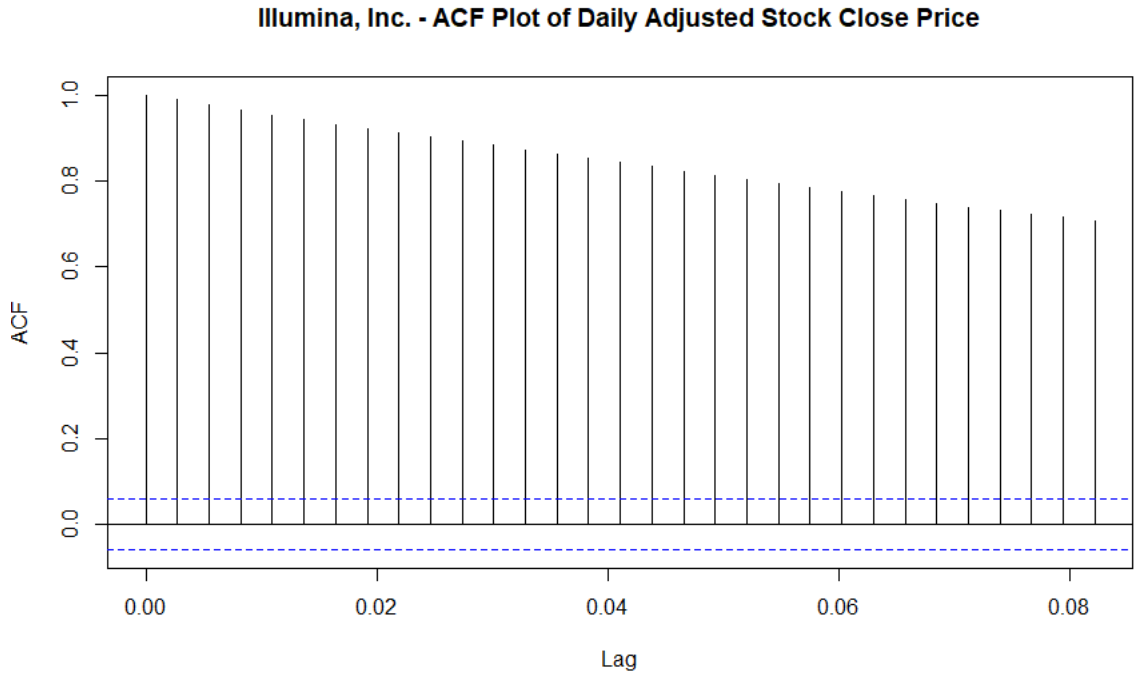

Figure 3.2: Autocorrelation Plot (ACF) for original, untransformed Illumina, Inc. Stock Price Data

To confirm this, we utilize the Augmented Dickey-Fuller Test; for a simple autoregressive (AR) time series model[2] represented as

$$y_t = \rho y_{t-1} + u_t, \tag{3.3}$$

in which ρ reflects the unit root, or stationarity and u_t reflects the intercept term, the Augmented Dickey-Fuller test hypothesizes the following:

$$H_0 : \rho = 0 \text{ (non-stationary data)}$$

$$H_1 : \rho \neq 0 \text{ (stationary data)}$$

[2]Verma, Yugesh.

Upon running this test, we obtain a p-value of 0.74. Thus, we fail to reject the null hypothesis; there is insufficient statistical evidence to claim that the data is stationary. We must therefore identify a transformation in the data to make it stationary prior to formulating any models.

To stabilize the variance, we will first take the *log* of the stock prices, followed by taking the first order difference to detrend the data. This transformation reflects the daily log returns of the stock price. Figure 3.3 below demonstrates that the log returns of the stock price exhibits stationary characteristics, such as constant variance.

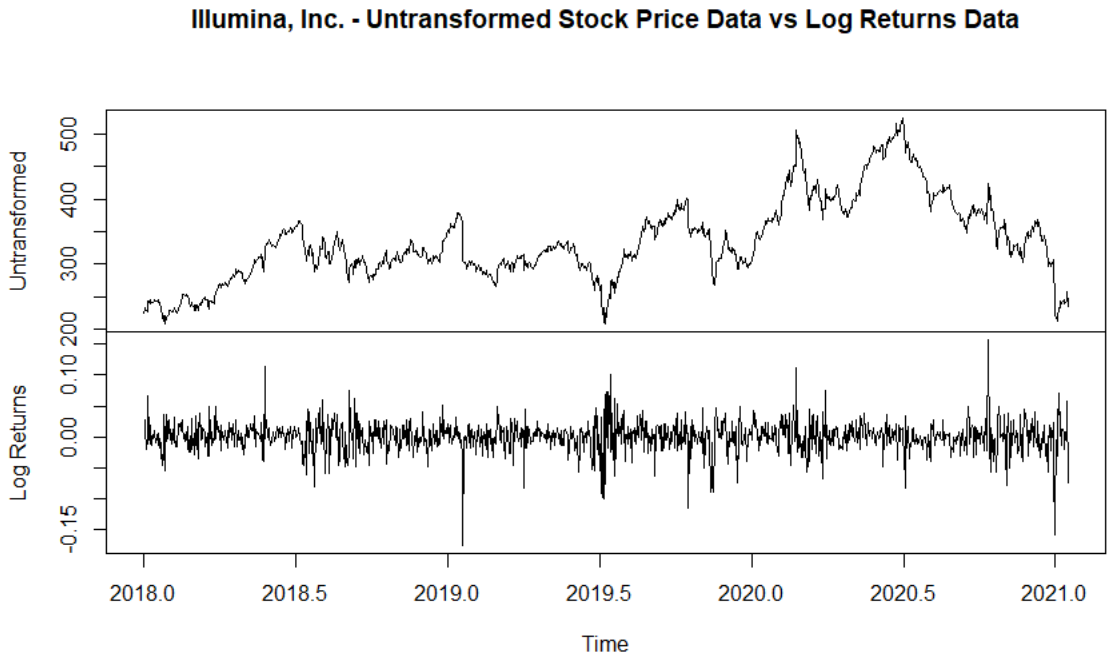

Figure 3.3: Illumina, Inc. - Original Time Series vs Log Returns Time Series

To confirm that the log returns transformation results in stationary data, we re-run the Augmented Dickey-Fuller Test, which yields a p-value < 0.01. Therefore, we reject the null hypothesis and conclude that there is sufficient statistical evidence to assume that the log returns transformation results in stationary data, which we will then use to formulate time series models. We will begin by creating baseline models using naive methods and then attempt to identify models with higher predictive power using Autoregressive Moving

Average models (denoted as ARMA(p, q)), exponential smoothing techniques, and ARIMA-GARCH techniques.

CHAPTER 4

Baseline Models

Three naive models are first created to serve as benchmarks for comparison for more advanced models devised in subsequent chapters: naive method, average method, and a simple moving average method. Our response variable is the log returns of the adjusted closing price for Illumina, Inc.

4.1 Naive Model

For naive models, all forecasts are simply equal to the value of the last observations, as delineated by the equation[1] below.

$$\hat{y}_{T+h|T} = y_T \tag{4.1}$$

Our testing data set consists of 168 observations (15% of the total data set) corresponding to fiscal Q3 of 2021 through the most recent of week of fiscal Q2 of 2022. Thus, the log returns from October 1st, 2021 through June 1st, 2022 is -0.015, which is identical to the log return corresponding to September 30th, 2021, which is the last day in our training data set.

[1] Hyndman and Athanasopoulos.

4.2 Average Model

For average models, the forecasts of all future values are equal to the average of the historical data. If historical data is denoted as $y_1, ..., y_T$, the forecasts can be defined as[2]:

$$\hat{y}_{T+h|T} = \bar{y} = (y_1 + ... + y_T)/T \tag{4.2}$$

The log returns forecast from October 1st, 2021 through June 1st, 2022 is 0.0006, which is equal to the average log returns from January 2nd, 2018 to September 30th, 2021(training data set time frame).

4.3 Moving Average Model

The simple moving average model defines the forecast for each future value as the average of the prior m values as defined by the equation[3] below. For our purposes, the moving average took the average of the last 3 observations.

$$\hat{T}_t = \frac{1}{m} \sum_{j=-k}^{k} y_{t+j} \tag{4.3}$$

4.4 Accuracy by Baseline Model

Accuracy metrics on our testing data set can be found in table 4.1 below for each method. Of the models, the average baseline model had the highest accuracy, as indicated by the lowest root mean square error (RMSE) of 0.0309. The mean error metric (ME) corresponding to the average model is -0.0039, which is closest to 0 compared to the other models and thus reiterates that it outperforms the naive and moving average model. We can interpret the mean error as follows: on average, the prediction values corresponding to the average model

[2]Hyndman and Athanasopoulos.

[3]Hyndman and Athanasopoulos.

were -0.0039 units lower than the actual value. The mean absolute error metric (MAE) suggests that, on average, the distance of the forecasts corresponding to the naive, average, and moving average models from the actual values is 0.0235, 0.0205, and 0.0226, respectively (all similar values). Based on the Mean Absolute Percentage Error metric (MAPE), all 3 baseline models were greater than 100%, indicating that each model overstated the actual log returns.

Data/Metric	ME	RMSE	MAE	MPE	MAPE
Naive	0.0114	0.0327	0.0235	92.0378	421.9247
Average	-0.0039	0.0309	0.0205	100.3413	106.7252
MA(3)	0.0091	0.0312	0.0226	92.6956	362.4764

Table 4.1: Naive Time Series Models - Accuracy Metrics

CHAPTER 5

Autoregressive Integrated Moving Average Model (ARIMA)

5.1 Parameter Selection

We attempt to model the log returns of the stock price data using an Autoregressive Integrated Moving Average Model, denoted as ARIMA(p, d, q). Since our variable of interest is the log returns of the data, we have already applied first order differencing to our response variable; thus, the parameter, d, which reflects the degree of differencing, equals 0, which makes this an Autoregressive Moving Average Model (denoted as ARMA(p, q)). We therefore need to identify the parameters, p and q, which reflect the lag order and the size of the moving average window, respectively. To do so, we will observe the autocorrelation (ACF) and partial auto-correlation plots (PACF) of the data in figure 5.1. While the ACF plot demonstrates statistically significant autocorrelations at multiple lag values (5, 6, 7, 17, etc.) as indicated by the extension beyond the 95% confidence interval bands, there remains a 5% chance that the correlation is random. The PACF plot shows similar conclusions to the ACF plot; thus, we hypothesize that the parameters p = 0 and q = 0 would fit the model best. *ARMA(*$0, 0$*)* reflects a white noise model, which assumes no log return dependence between subsequent years.

To select a model, the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) were assessed for various parameter values in tables 5.1 and 5.2 (lowest values of AIC and BIC are most strongly desired). Based on the tables below, the AIC

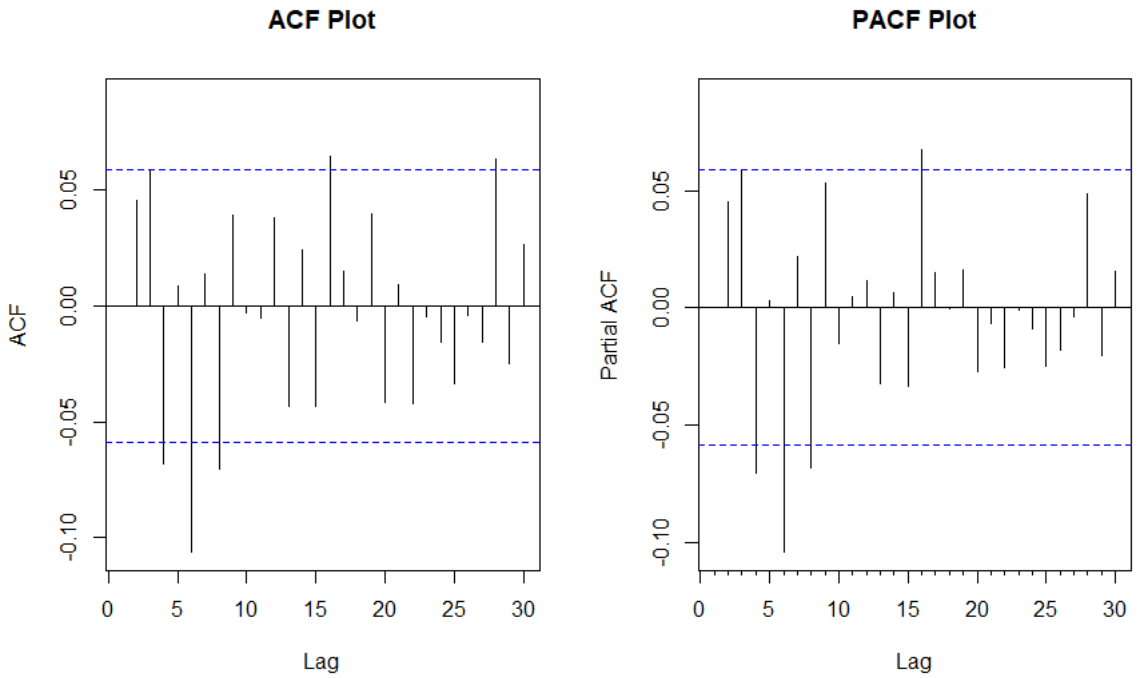

Figure 5.1: Illumina, Inc. - ACF and PACF Plots for Log Returns Data

criteria suggests that $ARIMA(3,0,2)$ is the ideal model, while the BIC criteria suggests that $ARIMA(0,0,0)$ is the ideal model, which is in line with our initial hypothesis. The model coefficients corresponding to $ARIMA(3,0,2)$ are cited in table 5.3 below.

p	d	q	AIC
3	0	2	$-5,017.833$
3	0	3	$-5,016.753$
2	0	3	$-5,016.706$
4	0	2	$-5,014.888$
4	0	3	$-5,014.445$
0	0	0	$-5,002.376$

Table 5.1: ARIMA Models - Parameter AIC Values

p	d	q	BIC
0	0	0	$-4,997.363$
1	0	0	$-4,983.34$
0	0	1	$-4,983.34$
1	0	1	$-4,976.863$
2	0	2	$-4,973.119$

Table 5.2: ARIMA Models - Parameter BIC Values

	AR(1)	AR(2)	AR(3)	MA(1)	MA(2)
Coefficients	-1.598	-0.7195	0.0635	1.6182	0.7646
Standard Error (SE)	0.086	0.1050	0.0337	0.0811	0.0912

Table 5.3: ARIMA(3, 0, 2) - Model Coefficients

Thus, both models will be evaluated for accuracy and diagnostics will be performed on the residuals.

5.2 Model Performance

Performance metrics on the testing data set related to *ARIMA(*$3, 0, 2$*)* and *ARIMA(*$0, 0, 0$*)* can be found in table 5.4 below. For most of the accuracy metrics (ME, RMSE, MAE), both models had nearly identical values. Only Mean Percentage Error (MPE) and Mean Absolute Percentage Error (MAPE) differed, in which *ARIMA(0, 0, 0)* had a slightly higher accuracy rate. The mean error (ME) suggests that on average, the predictions from both models differed from the actual values by -0.0039 units. Figures 5.2 and 5.3 display the forecasted values for each of the models. *ARIMA(*$0, 0, 0$*)* exhibits less variability in the forecast as compared to *ARIMA(*$3, 0, 2$*)*.

Data/Metric	ME	RMSE	MAE	MPE	MAPE
$ARIMA(3,0,2)$	-0.0039	0.0309	0.0204	96.2769	110.4428
$ARIMA(0,0,0)$	-0.0039	0.0309	0.0205	100.3413	106.7252

Table 5.4: ARIMA Models - Accuracy Metrics on Testing Data

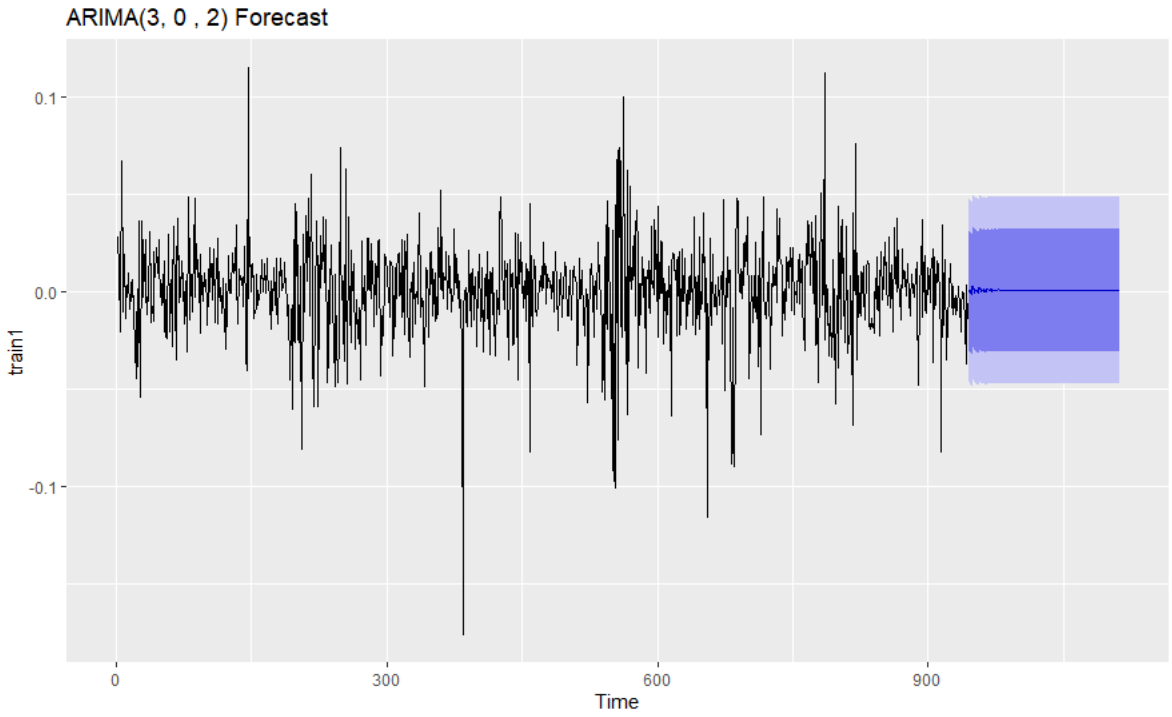

Figure 5.2: ARIMA(3, 0, 2) - Forecast of Log Returns

5.3 Residuals Diagnostics

After generating our 2 models, residuals are assessed to ensure the following criteria are satisfied: residuals are uncorrelated, have a mean of 0, exhibit homoscedasticity, and are characterized by normal distribution.

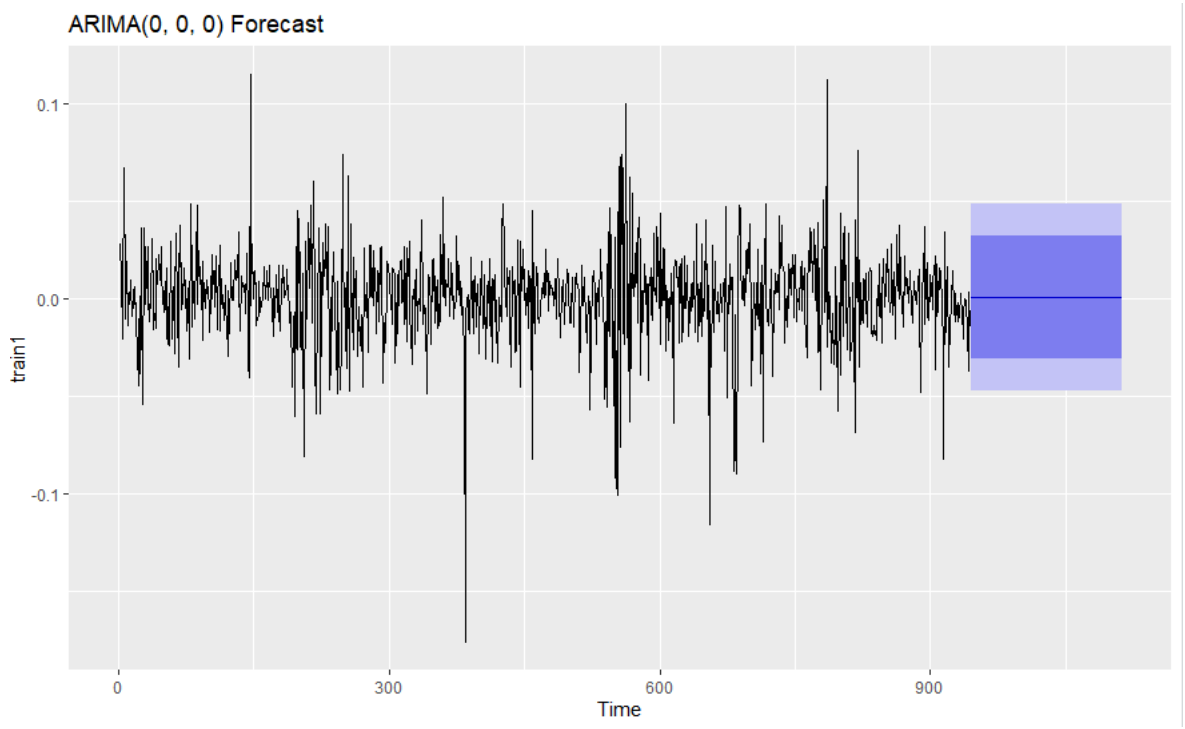

Figure 5.3: ARIMA(0, 0, 0) - Forecast of Log Returns

5.3.1 ARIMA(3, 0, 2) - Residual Diagnostics

The plot in figure 5.4 shows the residuals of the model, which appear to be scattered randomly around 0. The mean of the residuals is approximately 0 and satisfies the 0 residual mean criteria. The variance of the residuals appears to be mostly constant and not a function of time.

To assess residual correlation, the Ljung-Box test was performed, which states the following hypotheses:

$$H_0 : \text{Residuals are independently distributed}$$

$$H_1 : \text{Residuals exhibit serial correlation}$$

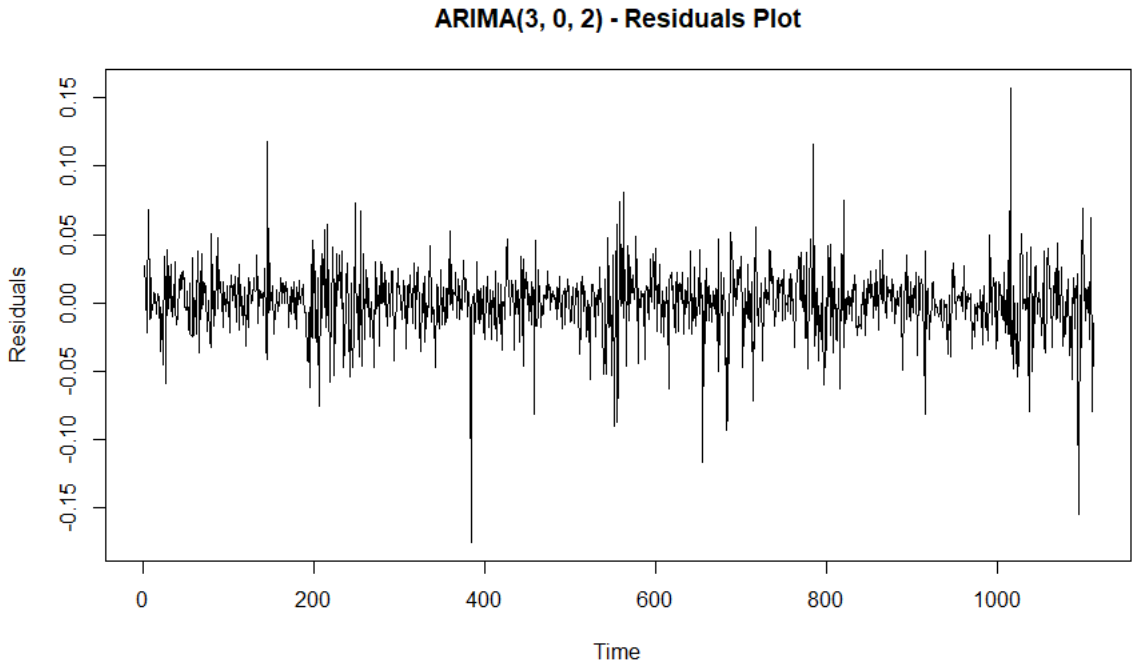

Figure 5.4: ARIMA(3,0,2) - Residuals Plot

The test yielded the results in table 5.5 below.

Metric	ARIMA(3, 0, 2)
Total Lags Used	10
Test Statistic, Q*	11.595
p-value	0.02063

Table 5.5: ARIMA(3,0,2) - Ljung-Box Test Results

Assuming a critical value of 0.05, we reject the null hypothesis, since the p-value of 0.02063 is less than the critical value. Thus, there is sufficient evidence to claim that the residuals exhibit serial correlation, which is a violation of the condition that residuals should be independent.

Next, we observe the histogram of the residuals in Figure 5.5. The histogram suggests that residuals are approximately normal, which satisfies our condition of homoscedasticity.

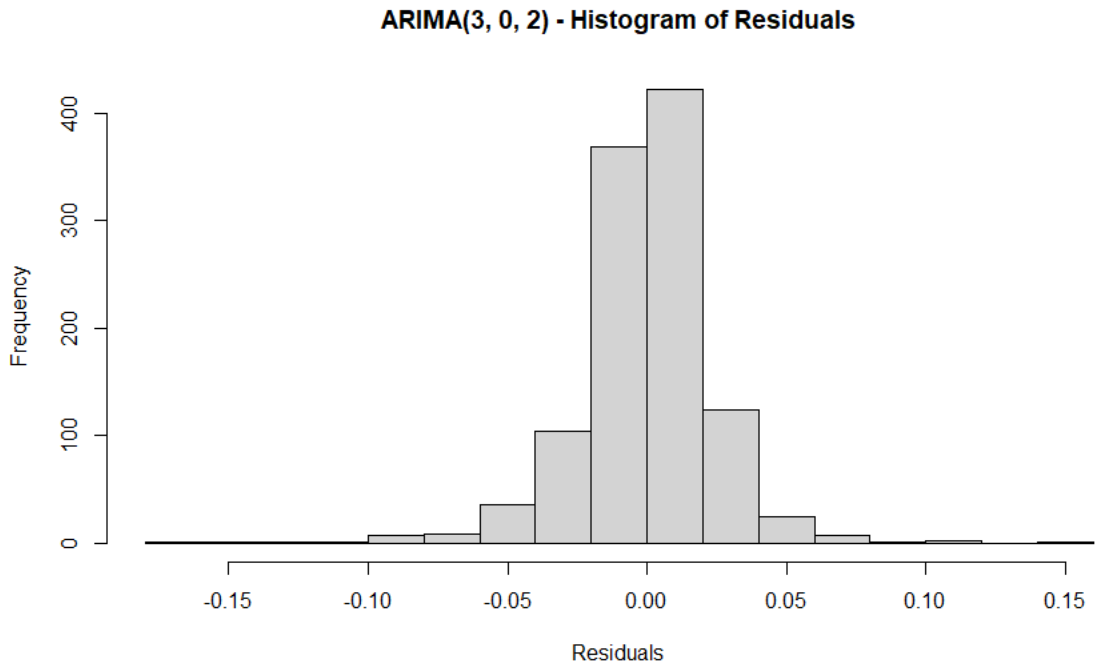

Figure 5.5: ARIMA(3, 0, 2) - Residuals Distribution Histogram

5.3.2 ARIMA (0, 0, 0) - Residual Diagnostics

The residuals plot in figure 5.6 appear to be scattered randomly around 0. The mean of the residuals is approximately 0 and satisfies the 0 residual mean criteria. The variance of the residuals appears to be mostly constant and not a function of time. Figure 5.7 displays the residuals histogram and aligns with the homoscedasticity assumption.

Metric	ARIMA(0, 0, 0)
Total Lags Used	10
Test Statistic, Q*	47.439
p-value	$3.25e-07$

Table 5.6: ARIMA(0,0,0) - Ljung-Box Test Results

We ran the Ljung-Box test and obtained results in table 5.6. Since the p-value is close to 0, we conclude that the residuals exhibit serial correlation.

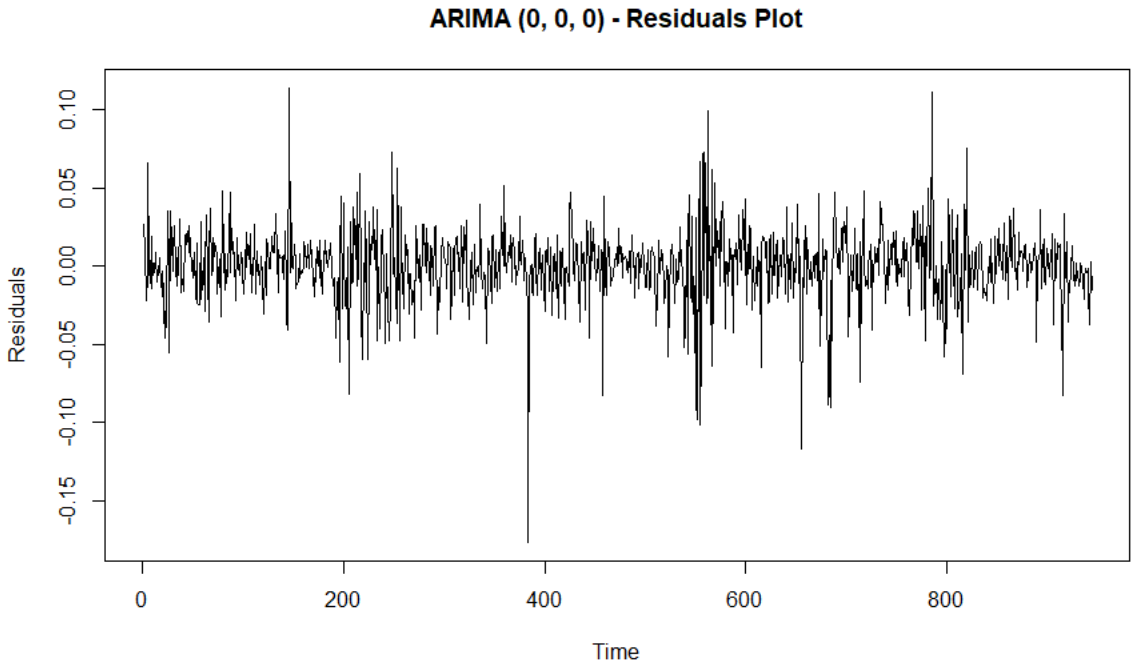

Figure 5.6: ARIMA(0, 0, 0) - Residuals Plot

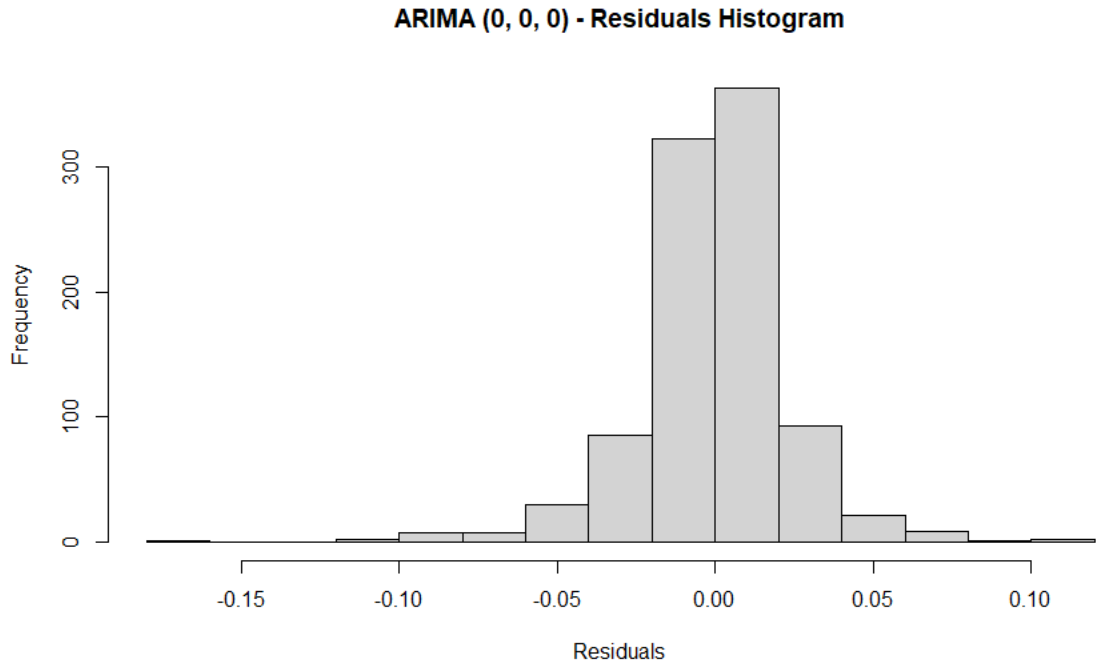

Figure 5.7: ARIMA(0, 0, 0) - Residuals Histogram

CHAPTER 6

Exponential Smoothing Model

For data that does not exhibit trend nor seasonality, simple exponential smoothing (SES) can be an effective univariate time series forecasting method, in which the weights of older observations exponentially decrease through the use of a smoothing parameter, α, which is often set to a value between 0 and 1. The Single Exponential Smoothing (SES) model can be written as[1]

$$F_t = \alpha A_{t-1} + (1 - \alpha)F_{t-1}, \tag{6.1}$$

where F_t reflects the exponentially smoothed forecast at time period t, α is a smoothing constant, A_{t-1} is the actual value in the prior period, and F_{t-1} is exponentially smoothed forecast for period *t-1*.

6.1 Parameter Selection

Our first step in formulating the model is identifying an optimal smoothing parameter, α. To do so, the root mean square error (RMSE) was evaluated for levels of α for increments of 0.01 between 0 and 1, inclusive. Figure 6.1 the RMSE for these various levels of α. Table6.1 below also shows RMSE for a sample of α values between 0.01 and 0.1. Through this method, we identify that the minimum RMSE value of 0.0306 corresponds to $\alpha = 0.03$.

[1]Hyndman and Athanasopoulos.

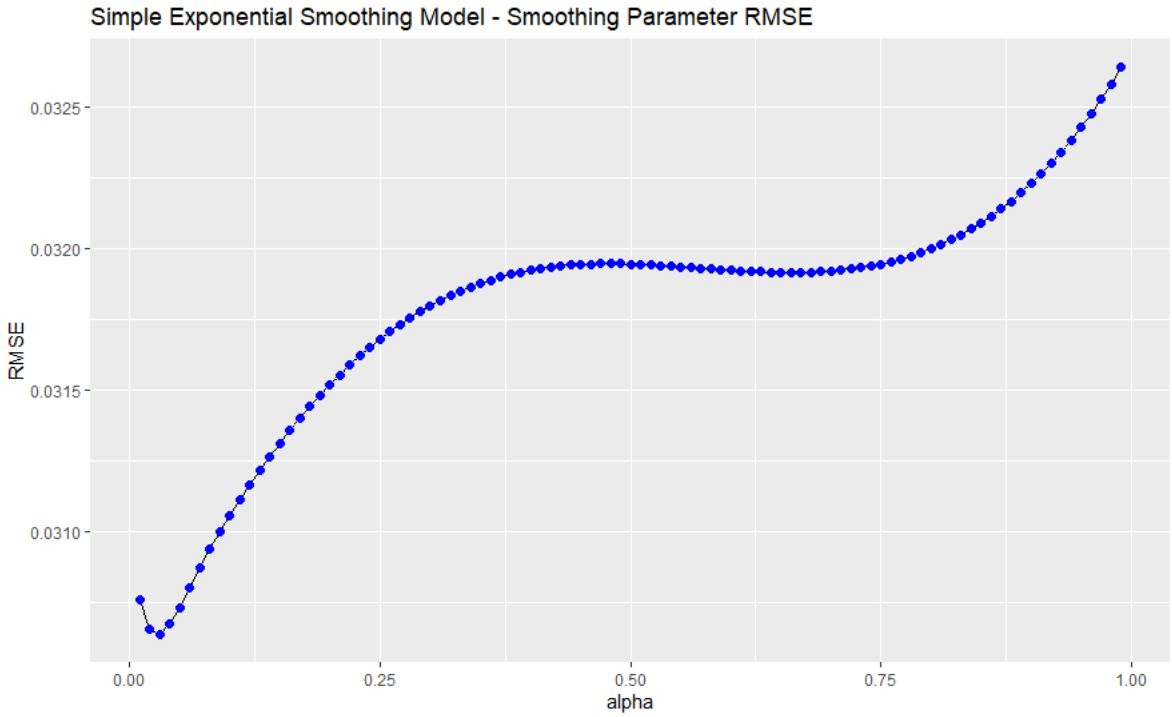

Figure 6.1: Exponential Smoothing Model - RMSE for various smoothing parameter values

α	RMSE
0.01	0.03076021
0.02	0.03065679
0.03	0.03063760
0.04	0.03067242
0.05	0.03073295
0.06	0.03080201
0.07	0.03087129
0.08	0.03093755
0.09	0.03100000
0.10	0.03105880

Table 6.1: Exponential Smoothing Parameter - RMSE Values

Utilizing a smoothing parameter of $\alpha = 0.03$ in the model yields accuracy metrics on the testing data set in table 6.2 and provides forecast values in figure 6.2. While model comparisons will be explored in greater depth in a later chapter, it should be noted that the RMSE corresponding to the simple exponential smoothing model is lower than that of all prior models, including the *ARIMA(3,0,2)*, *ARIMA(0,0,0)*, and our 3 baseline models (naive, average, and moving average models); however, residuals must be diagnosed to assess potential drawbacks in prediction quality.

Data/Metric	ME	RMSE	MAE	MPE	MAPE
Testing	0.0003	0.0306	0.0205	98.0711	157.4763

Table 6.2: Exponential Smoothing - Accuracy Metrics on Testing Data

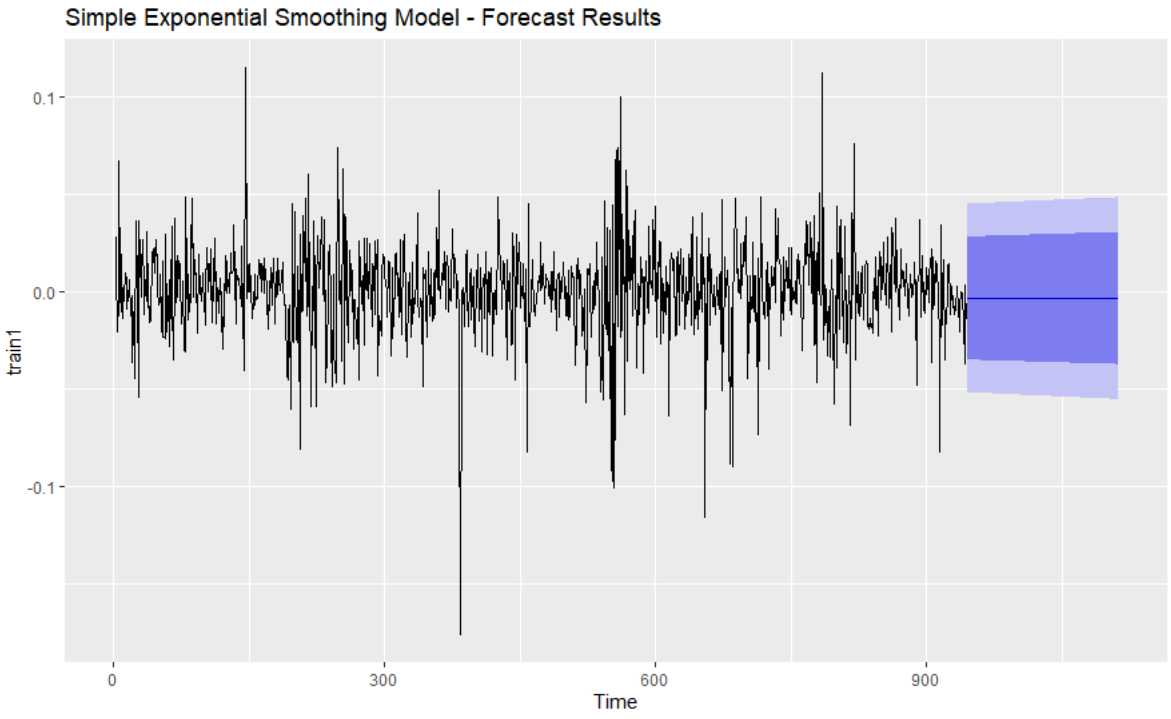

Figure 6.2: Simple Exponential Smoothing - Forecast Results

6.2 Residuals Diagnostics

Visually, the plot of the residuals in figure 6.3 demonstrates even fluctuations around 0; we confirm that the mean of the residuals is approximately 0, indicating that predictions will not necessarily be biased; however, there appears to be some fluctuations in the variance of the residuals(i.e. less variance between $t = 300$ and $t = 400$).the histogram of the residuals in figure 6.4 suggests alignment with normality assumptions, indicating that prediction intervals may be accurate.

The Ljung-Box test was also performed, obtaining a p-value of approximately 0, indicating that there is sufficient statistical evidence to conclude serial correlation among residuals. Thus, additional transformation of the data could aid in reducing serial correlation; however, this is beyond the scope of our current analysis.

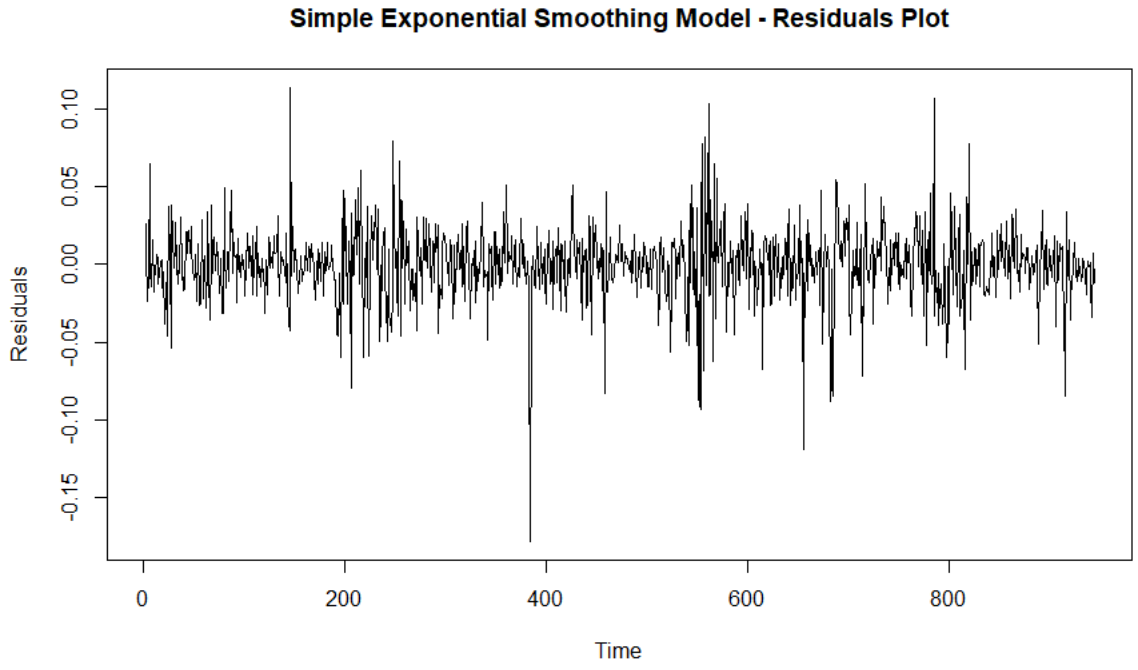

Figure 6.3: Simple Exponential Smoothing - Residuals Plot

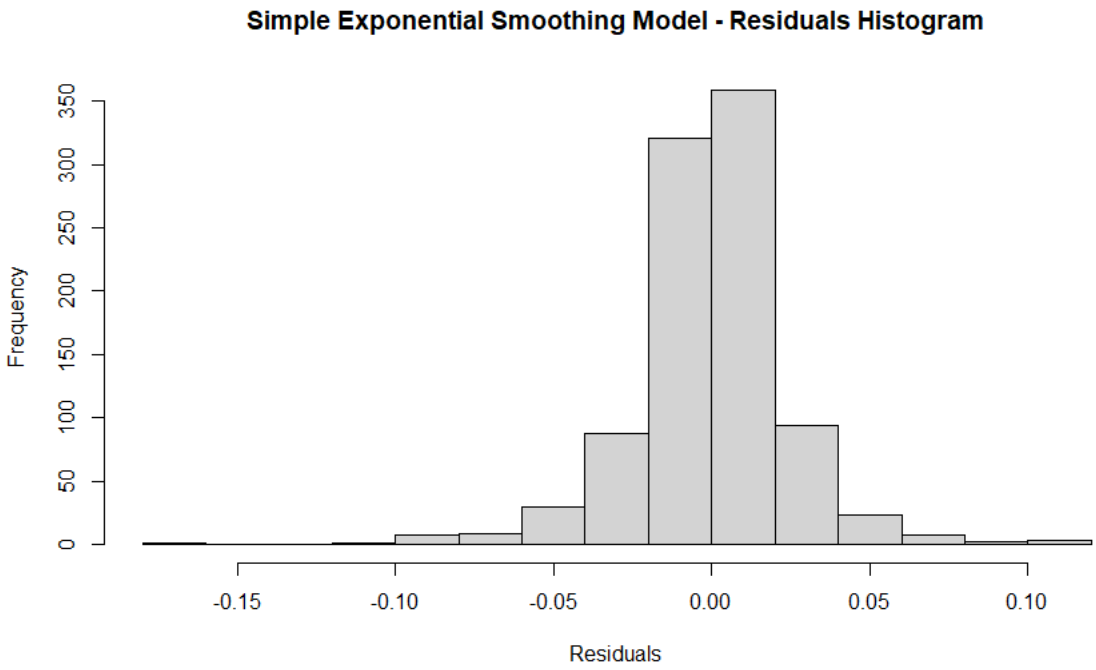

Figure 6.4: Simple Exponential Smoothing - Residuals Histogram

CHAPTER 7

ARIMA - GARCH Model

When visually assessing the time series of the daily log returns of the company's stock in figure 2.1, we notice that the data is characterized by volatility, which can impact a statistical model's prediction accuracy. *ARIMA* models such as those formulated in chapter 5 cannot capture volatility because their conditional variance is constant. As a result, one way to remedy this issue is to incorporate Generalized Autoregressive Conditional Heteroskedasticity (GARCH) with an *ARIMA* model in order to successfully capture some known properties of returns time series such as volatility clustering. A *GARCH(p,q)* process can be defined[1] as:

$$X_t = e_t \sigma_t, \tag{7.1}$$

in which e_t reflects the error term, which is a white noise process that does not necessarily have a constant variance; σ_t reflects this conditional variance, which can be further expressed[2] as:

$$\sigma_t^2 = \omega + \alpha_1 X_{t-1}^2 + ...\alpha_p X_{t-p}^2 + \beta_1 \sigma_{t-1}^2 + ... + \beta_q \sigma_{t-q}^2, \tag{7.2}$$

In the conditional variance equation, p represents the number of lag variances, q represents the number of lag residual errors to include in the GARCH model, and ω represents the variance intercept. It should be noted that GARCH models can only be applied to data that exhibits stationarity, volatility, and ARCH effects, which is explained in further detail

[1]Williams, Brandon.

[2]Williams, Brandon.

in the next section. We already confirmed that our data is stationary in chapter 3 using the Augmented Dickey-Fuller test.

In order to formulate an ARIMA-GARCH model, we must first confirm the existence of ARCH effects, which measures the autocorrelation of squared residuals. Upon confirmation of ARCH effects, we will then select GARCH model parameters and *ARIMA* model parameters and apply this to our testing data set.

7.1 Parameter Selection

We confirm the existence of ARCH effects using Engle's ARCH Test, which examines the autocorrelation parameter, α, of our squared residuals. A model with ARCH effects is one in which α is not 0. Thus, our null hypothesis suggests no ARCH effects, while our alternative hypothesis suggests existence of ARCH effects, which can be stated as follows:

$$H_0 : \alpha_0 = \alpha_1 = ... = \alpha_m = 0$$

$$H_1 : e_t^2 = \alpha_0 + \alpha_1 e_{t-1}^2 + ... + \alpha_m e_{t-m}^2 + e_t$$

Upon running this test, our p-value is approximately 0, which means that we reject the null hypothesis. There is sufficient evidence to conclude that ARCH effects are present in the data and that a GARCH model can be utilized in our case. The next step is to identify the parameters, p and q, which will be selected based on maximum likelihood (ML) estimation, whereby the likelihood function, defined[3] as

$$LF = \Pi_{t=1}^{N} \frac{1}{\sqrt{2\pi\sigma_t^2}} exp(\frac{-X_t^2}{2\sigma_t^2}), \tag{7.3}$$

[3]Hyndman and Athanasopoulos.

is maximized. This selection process yields a GARCH model of order $p = 1$ and $q = 1$ with the conditional variance parameters in table 7.1 below:

ω	α	β
0.000143	0.243060	0.562652

Table 7.1: GARCH(1,1) Model - Coefficients

Next we must select parameters corresponding to our *ARIMA* model. From chapter 5, we identified that *ARIMA(3, 0, 2)* and *ARIMA(0, 0, 0)* were the best fits for our data based on *AIC* and *BIC* criteria; therefore, we will select *ARIMA(0, 0, 0)* for this model in conjunction with *GARCH(1, 1)*. Accuracy metrics can be found in table 7.2 below:

Data/Metric	ME	RMSE	MAE	MPE	MAPE
Testing	−0.0046	0.0310	0.0205	100.7553	118.2443

Table 7.2: ARIMA(0,0,0)-GARCH(1,1) Model – Accuracy Metrics on Testing Data

This demonstrates similar accuracy results to the *ARIMA(0,0,0)* and *ARIMA(3,0,2)* models that did not utilize GARCH effects; the exponential smoothing model, however, has a slightly lower RMSE than this ARIMA-GARCH model.

7.2 Residual Diagnostics

Visually, the plot of the residuals in figure 7.1 demonstrates even fluctuations around 0; we confirm that the mean of the residuals is approximately 0, indicating that predictions will not necessarily be biased. The variance also appears to be mostly constant. The histogram of the residuals in figure 7.2 suggests alignment with normality assumptions, indicating that prediction intervals may be accurate.

The Ljung-Box test was also performed, obtaining a p-value of approximately 0, indicating that there is sufficient statistical evidence to conclude serial correlation among residuals.

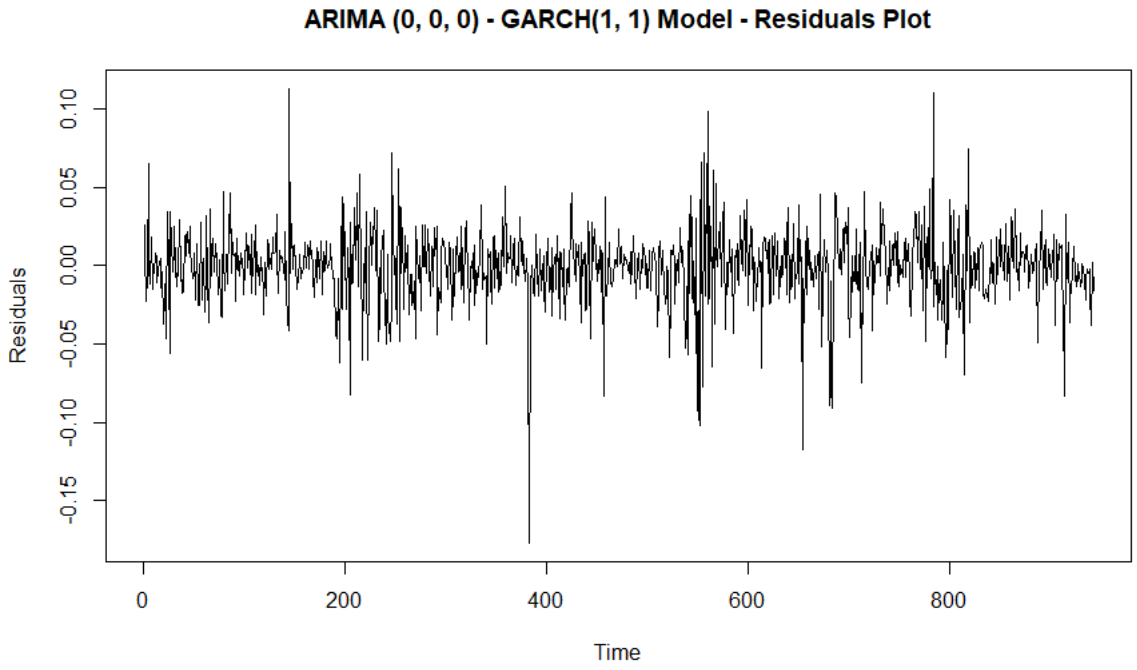

Figure 7.1: ARIMA(0,0,0) - GARCH(1,1) – Residuals Plot

Thus, additional transformation of the data could aid in reducing serial correlation; however, this is beyond the scope of our current analysis.

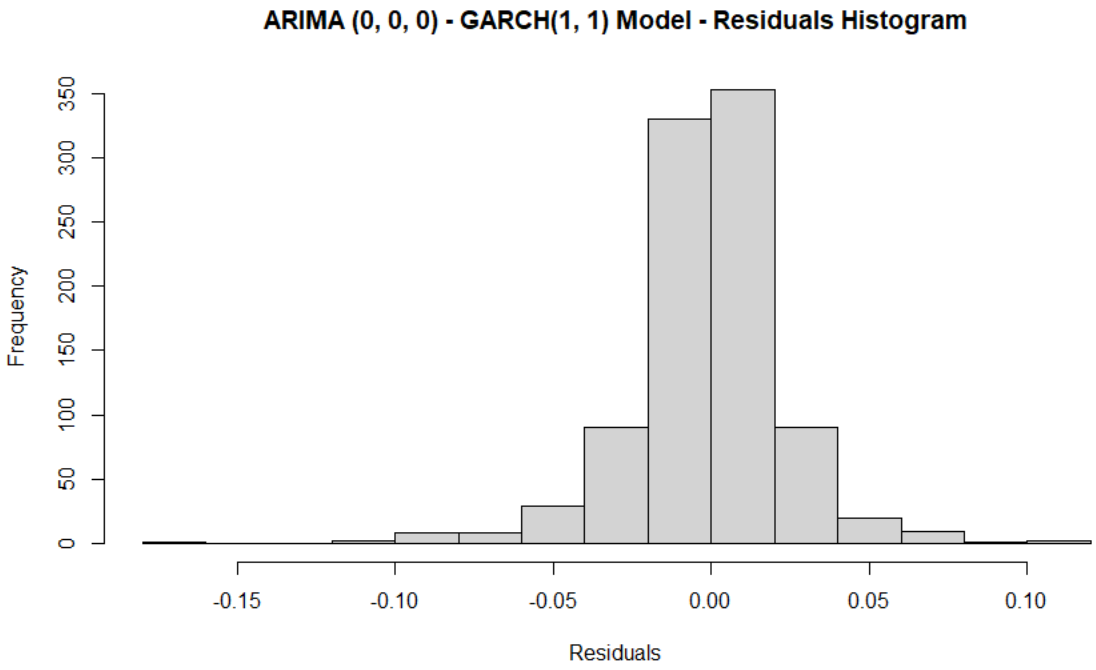

Figure 7.2: ARIMA(0,0,0) - GARCH(1,1) – Residuals Histogram

CHAPTER 8

Conclusion

8.1 Model Comparison

Table 8.1 demonstrates accuracy metrics for each model applied to the testing data set.

Data/Metric	ME	RMSE	MAE	MPE	MAPE
Naive	0.0114	0.0327	0.0235	92.0378	421.9247
Average	−0.0039	0.0309	0.0205	100.3413	106.7252
MA(3)	0.0091	0.0312	0.0226	92.6956	362.4764
$ARIMA(3,0,2)$	−0.0039	0.0309	0.0204	96.2769	110.4428
$ARIMA(0,0,0)$	−0.0039	0.0309	0.0205	100.3413	106.7252
Exponential Smoothing	0.0003	0.0306	0.0205	98.0711	157.4763
ARIMA-GARCH	−0.0046	0.0310	0.0205	100.7553	118.2443

Table 8.1: Model Comparison - Accuracy Metrics on Testing Data

Table 8.1 demonstrates that all of our models performed remarkably similarly, with all RMSE values falling between 0.030 and 0.035. According to the RMSE criteria, the exponential smoothing model, which was selected by identifying the smoothing parameter corresponding to the lowest RMSE, was the most ideal on the testing data set. This model also has the lowest absolute mean error (ME) value of all the models. The model with the highest RMSE and MAPE values is the naive model, which simply set prediction values equal to the value of the last observation in the training data set.

8.2 Applications of Simple Exponential Smoothing Model

Since the simple exponential smoothing model outperformed the other models according to the RMSE criteria, we will apply this model to the log returns data of 3 competitors, namely Thermo Fisher Scientific, Qiagen, and Agilent Technologies to assess how well the model works on other companies within the same industry. It should be noted that all 3 companies have been publicly traded during the same time as our data set for Illumina, Inc., so the same data set sizes and time frames were selected for each company. Stationarity of the log returns of the stock was confirmed for each company, as the p-values from the Augmented Dickey-Filler test were all less than 0.01.

Table 8.2 displays the accuracy metrics for the 4 companies below:

Data/Metric	ME	RMSE	MAE	MPE	MAPE
Illumina, Inc.	0.0003	0.0306	0.0205	98.0711	157.4763
Thermo Fisher Scientific	−0.0001	0.0184	0.0138	83.7316	108.2026
Agilent Technologies	−0.0007	0.0200	0.0163	107.9193	109.7027
Qiagen	−0.0003	0.0172	0.0130	Inf	Inf

Table 8.2: Exponential Smoothing - Accuracy Metrics by Company

Based on table 8.2, it appears that the model actually performed better on all of the other companies, with the RMSE being the lowest for Qiagen.

8.3 Concluding Remarks

While some models performed better than others, it should be noted that all models failed the condition in which residuals should not exhibit serial correlation, which indicates that additional transformation of the data or further information can be added to the models to aid in reducing serial correlation, thereby improving model accuracy and reliability. One

method that could be explored is time series regression, in which our time series is assumed to have a linear relationship with another time series. With the most recent declines in the overall stock market, it could be of value to create a model that includes a relationship with the movement of the S&P 500, for example.

REFERENCES

Verma, Yugesh. "Complete Guide to Dickey-Fuller Test in Time-Series Analysis." Analytics India Magazine, 6 Sept. 2021, https://analyticsindiamag.com/complete-guide-to-dickey-fuller-test-in-time-series-analysis/.

Hyndman, Roy J, and George Athanasopoulos. "Forecasting: Principlesnbsp;andnbsp;Practice (2nd Ed)." 3.1 Some Simple Forecasting Methods, https://otexts.com/fpp2/simple-methods.html.

Williams, Brandon. GARCH(1,1) Models - University of California, Berkeley. 15 July 2011, https://math.berkeley.edu/ btw/thesis4.pdf.

Cancherini, Laura, et al. "What's Ahead for Biotech: Another Wave or Low Tide?" McKinsey amp; Company, McKinsey amp; Company, 15 Sept. 2021.

Santoli, Michael. "The S&P 500 Has Already Met Its Average Return for a Full Year, but Don't Expect It to Stay Here." CNBC, CNBC, 19 June 2017.