Yahoo Stocks Project

Dalina Sinn (dalinasinn@ucsb.edu)

2023-03-26

Abstract

This report will analyze Seasonal Autoregressive Integrated Moving Average (SARIMA) and Generalized Autoregressive Conditional Heteroskedasticity (GARCH) models as well as forecasting techniques for those predictions. The data being used was daily and differencing and log transformations will also be used to make the data stationary. Then, ACF and PACF graphs were created and the SARIMA model used is SARIMA$(1, 1, 2)$ x $(4, 2, 1)_{12}$. After analyzing the outcomes, SARIMA$(1, 1, 2)$ x $(4, 2, 1)_{12}$ generated the best fit and with this model, a 12 month forecast prediction was created. The model chosen is not a perfect prediction of Yahoo's stock prices but the model constructed a forecast that has similar patterns as the actual Yahoo stocks. For the GARCH model, ACF and PACF graphs were also made to find the parameters of the model. After testing a few different models, the best model was ARMA$(2, 2)$ + GARCH$(3, 5)$.

1. Introduction

Yahoo is a web service provider that contains services such as email, search engines, etc. Stock prices may rise or fall and prices fluctuate making it hard to figure out the best time to invest. Stock price forecasting is important to help determine these trends. The purpose of this project is to explore predict stock prices so people can recognize patterns to see how well Yahoo is doing and figure out the best time to make an investment. The methods used for this project are SARIMA models and GARCH models. When analyzing this data set, a decrease in 2018 is shown. In 2018, there was a stock market crash that affected many companies which explains why Yahoo stock prices also experienced a drop. In addition, there is a noticeable decrease in stock prices in early 2020, which would be caused by the COVID-19 pandemic. Like most stocks during this time, the prices dropped and Yahoo stocks was one of them. Besides these significant changes, Yahoo stock prices have shown to have a steady increase over the years.

2. Data The Yahoo stocks data set chosen was from Kaggle, which said that it was founded using Yahoo Finance databases. The link can be found here: https://www.kaggle.com/datasets/arashnic/time-series-forecasting-with-yahoo-stock-price. The time range for this data started on November 23, 2015 and ended on November 20, 2020. There is daily information for the stock prices and it utilized the closing price information. For this particular project, the main focus is the `Adj Close` which is "adjusted values factor in corporate actions." There are 1825 observations and it is important to look at this data because it will allow people to understand stocks and predict its movement for financial trading.

3. Methodology

SARIMA SARIMA models are used to take into account seasonality in the data and the model is built as SARIMA(p, d, q) x (P, D, Q). To find the values in the SARIMA model, it is necessary to use differencing and log transformations and then analyzing ACF and PACF graphs as well. This model looks at both seasonal and non-seasonal aspects of the data.

GARCH GARCH models are used in the financial market to indicate volatility in the data. This model contains the conditional mean as well as conditional variance to model the volatility of Yahoo stocks.

4. Results

SARIMA:

```
library(readr)
library(astsa)
library(xts)
```

```
## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

##
## ################################### WARNING ###################################
## # We noticed you have dplyr installed. The dplyr lag() function breaks how    #
## # base R's lag() function is supposed to work, which breaks lag(my_xts).      #
## #                                                                             #
## # If you call library(dplyr) later in this session, then calls to lag(my_xts) #
## # that you enter or source() into this session won't work correctly.         #
## #                                                                             #
## # All package code is unaffected because it is protected by the R namespace   #
## # mechanism.                                                                  #
## #                                                                             #
## # Set `options(xts.warn_dplyr_breaks_lag = FALSE)` to suppress this warning.  #
## #                                                                             #
## # You can use stats::lag() to make sure you're not using dplyr::lag(), or you #
## # can add conflictRules('dplyr', exclude = 'lag') to your .Rprofile to stop   #
## # dplyr from breaking base R's lag() function.                                #
## ################################### WARNING ###################################
```

```
library(fGarch)
```

```
## NOTE: Packages 'fBasics', 'timeDate', and 'timeSeries' are no longer
## attached to the search() path when 'fGarch' is attached.
##
## If needed attach them yourself in your R script by e.g.,
##         require("timeSeries")
```

```
library(forecast)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo

##
## Attaching package: 'forecast'
```

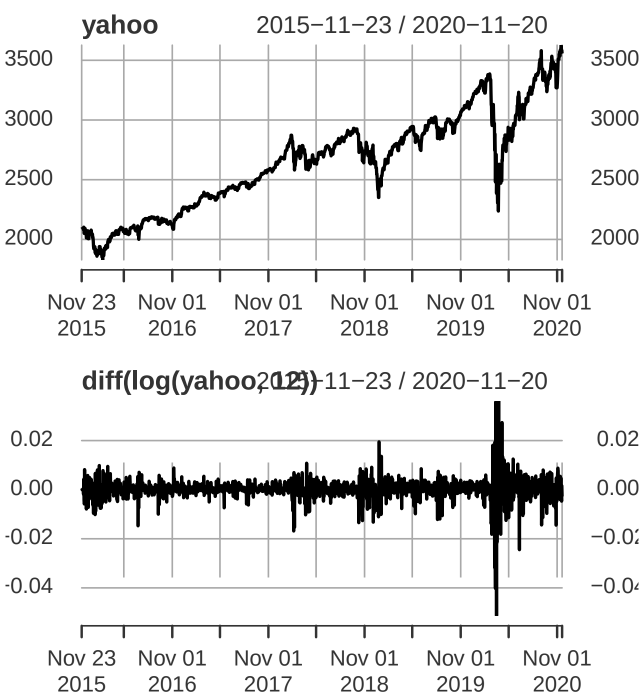
```
## The following object is masked from 'package:astsa':
##
##     gas
```

```
yahoo_stock <- read_csv("yahoo_stock.csv", show_col_types = FALSE)
```

Plots:

Plotting the data allows us to recognize any transformations that we may need to do. The first plot is our original data with no changes. As we can see, there seems to be a noticeable expontential graph which indicates that a log transformation can help counteract this. Then, to detrend the data and to make it stationary, the difference is taken and make it seasonal as well.

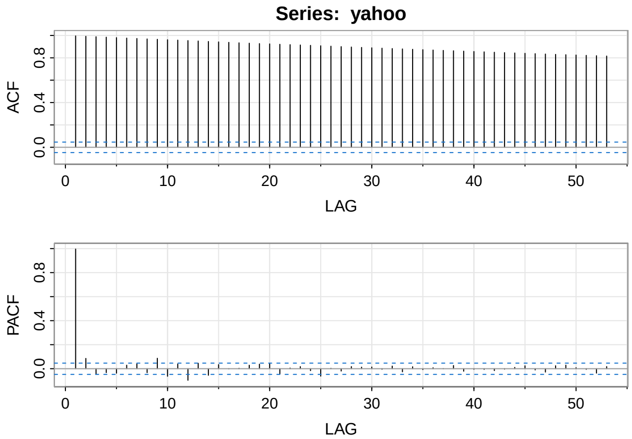
```
yahoo <- xts(yahoo_stock$`Adj Close`, order.by = yahoo_stock$Date)
par(mfrow = c(2,2))
plot(yahoo)
plot(log(yahoo))
plot(diff(log(yahoo, 12)))
```



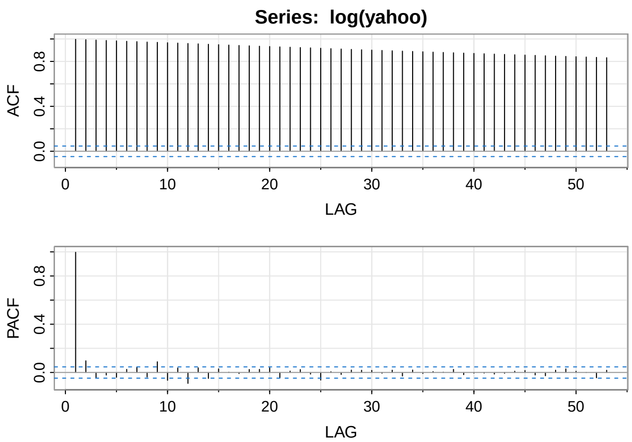
ACF/PACF Results:

After plotting and transforming the data, it is now time to plot the ACF and PACF plots. We will use the first difference ACF and PACF graph to determine the values for the non-seasonal aspect of the SARIMA model. When adding the value 12 to create our seasonal component, we can see that there is a decreasing pattern in the ACF portion. Since we have this decrease, we would take the second difference.

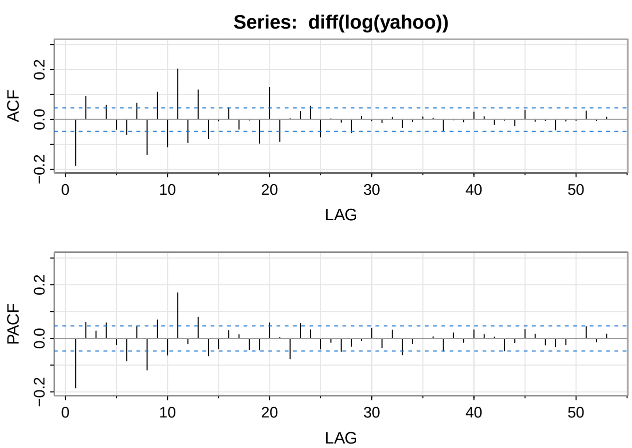
```
acf2(yahoo)
```



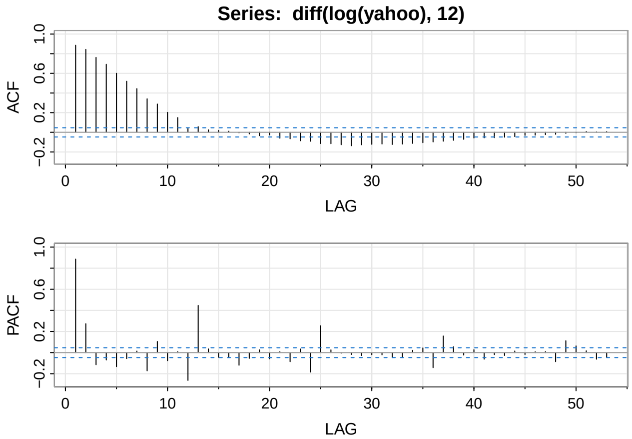

```
acf2(log(yahoo))
```



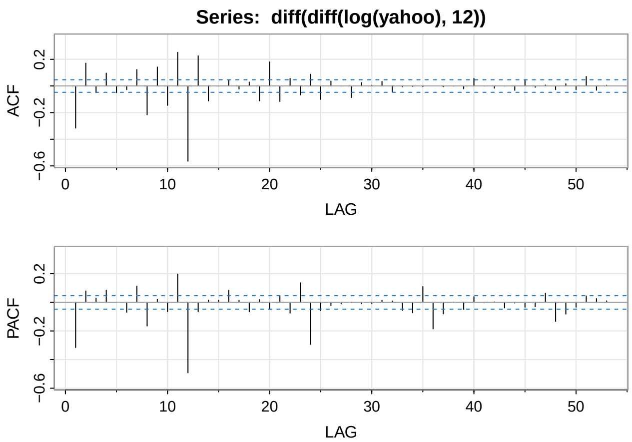

```
acf2(diff(log(yahoo)))
```




```
acf2(diff(log(yahoo), 12))
```

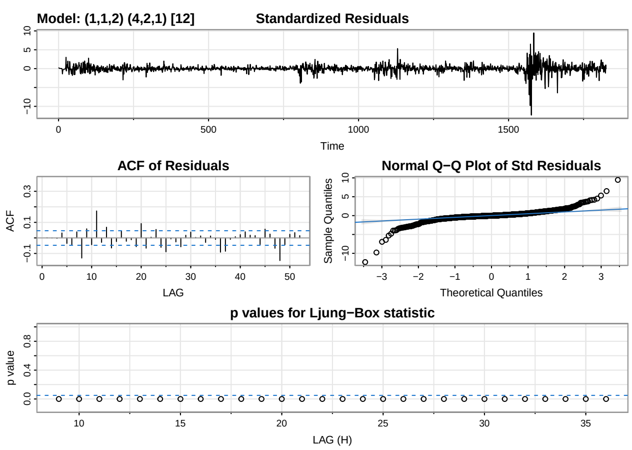



```r
acf2(diff(diff(log(yahoo), 12)))
```



From these plots, we can determine that SARIMA$(1, 1, 2)$ x $(4, 2, 1)_{12}$ is the model chosen.

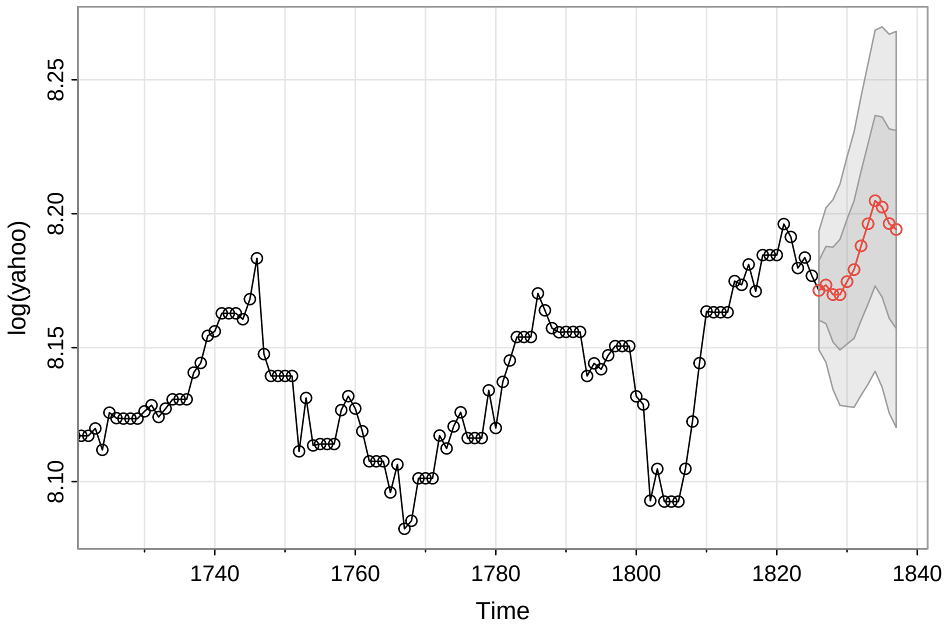
```r
sarima(log(yahoo), p = 1, d = 1, q = 2, P = 4, D = 2, Q = 1, S = 12)
```



Forecasting:

With our SARIMA$(1, 1, 2)$ x $(4, 2, 1)_{12}$ model, we will use it to forecast the next 12 observations. As we can see, it follows a similar pattern to previous points as has an increasing trend and then decreases.

```
sarima.for(log(yahoo), n.ahead = 12, p = 1, d = 1, q = 2, P = 4, D = 2, Q = 1, S = 12)
```

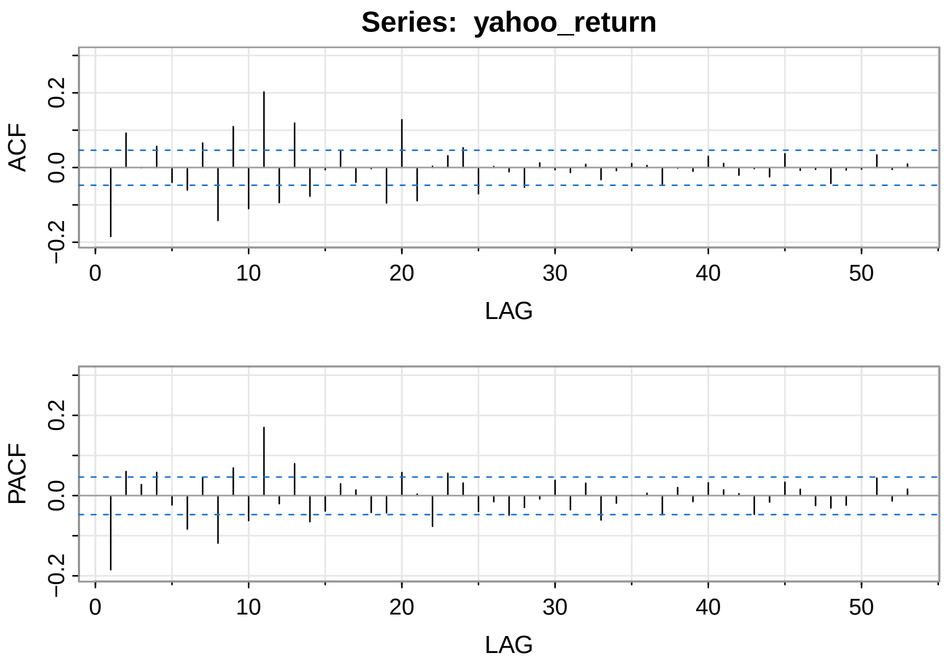


GARCH

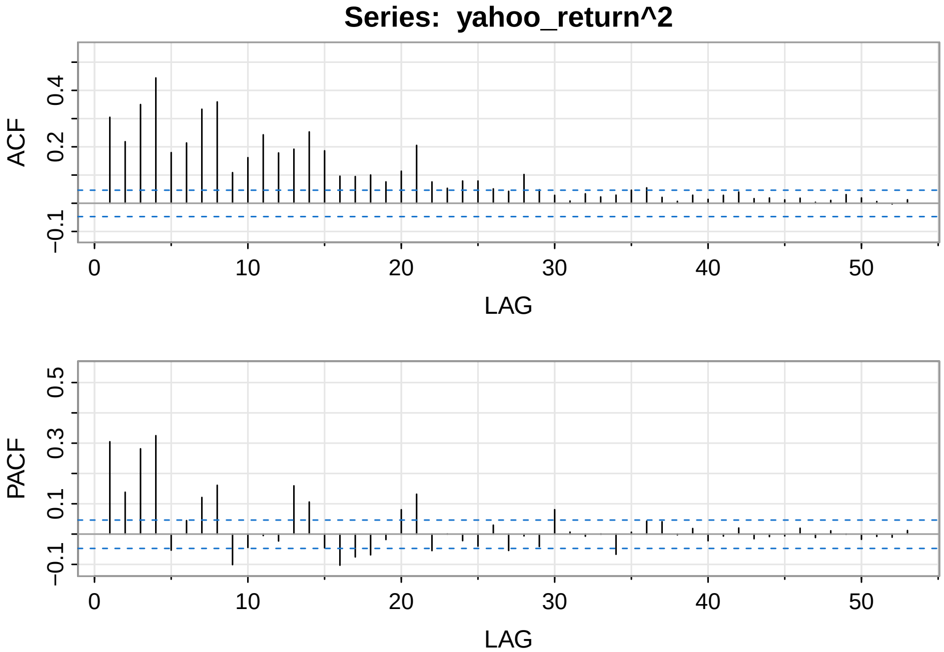
ACF/PACF Results

Since we are working with stocks, the GARCH model will be using Yahoo returns as the data. To determine the values used for the model, the ACF and PACF of the returns will be used. From the plots, ARMA(2,2) and GARCH(3,5) will be used.

```
yahoo_return <- diff(log(yahoo))[-1]
acf2(yahoo_return)
```

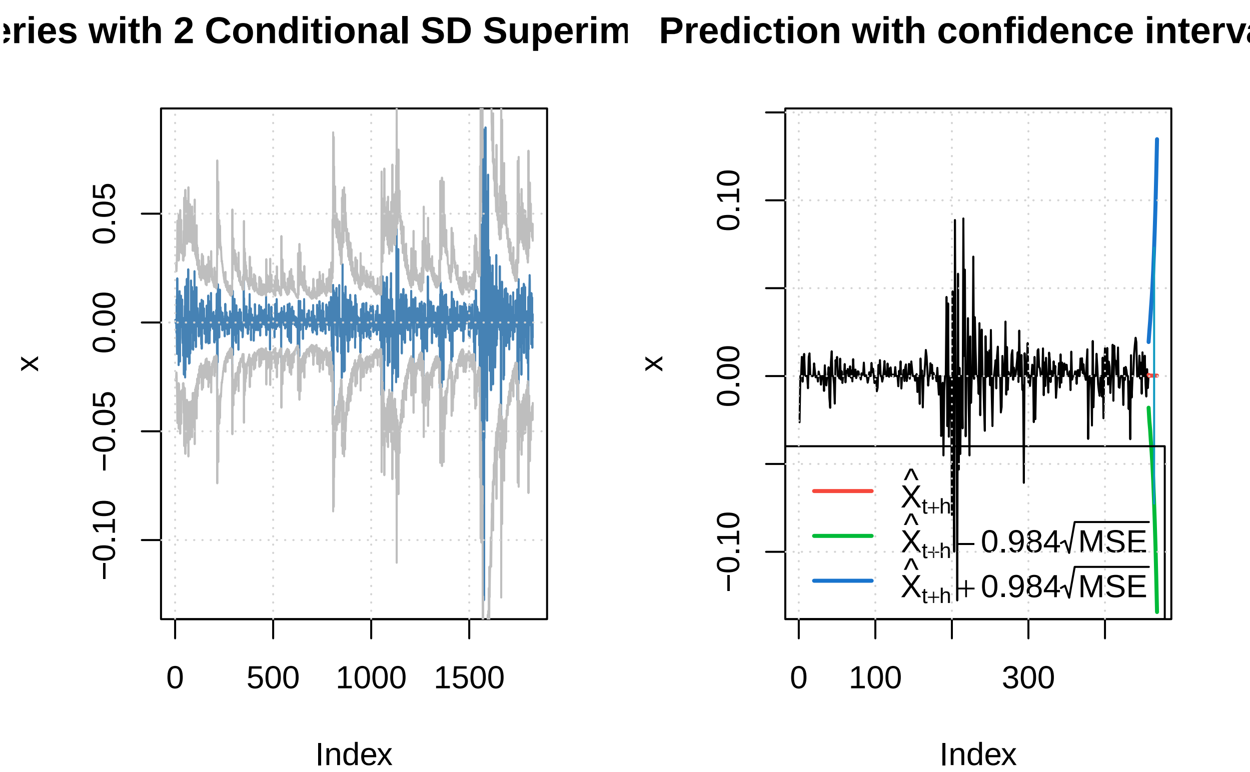


```
acf2(yahoo_return^2)
```



GARCH Plot and Forecasting: In our Series with 2 Conditional SD Superimposed plot, it appears that Yahoo returns are quite volatile which are indicated by the spikes. In addition, we can forecast the next 12 observations just like we did for the SARIMA model.

```
summary(yahoo_return.g <- garchFit(~arma(2,2)+garch(3,5), data=yahoo_return, cond.dist='std'))
par(mfrow = c(1,2))
plot(yahoo_return.g, which = 3)
predict(yahoo_return.g, n.ahead=12, plot = T)
```



eries with 2 Conditional SD Superim
Prediction with confidence interva
x
0.05
0.00
−0.05
−0.10
0
500
1000
1500
Index
0.10
0.00
−0.10
100
300
$\hat{X}_{t+h}$
$\hat{X}_{t+h} - 0.984\sqrt{MSE}$
$\hat{X}_{t+h} + 0.984\sqrt{MSE}$

Conclusion and Future Study

To summarize, we started off by performing differences and log transformations in order to detrend and create a stationary time series to the Yahoo data set. Then, we read ACF/PACF graphs so we can construct our SARIMA model. After choosing our model, we forecast the next 12 observations which followed an increasing then decreasing pattern.

After doing SARIMA models, we moved onto a GARCH model. We used the Yahoo returns data for this portion. We repeated the same steps by analyzing the ACF and PACF plots to find the best values for our GARCH model. From this model and plotting it, we discovered that there is volatility and then we forecast the next 12 observations as well.

In future studies, we can apply the techniques we learned in other data sets. We can explore different stock data as well and use other models to see if they would forecast better outcomes than SARIMA and GARCH

References

The data used was from Kaggle: https://www.kaggle.com/datasets/arashnic/time-series-forecasting-with-yahoo-stock-price

Information about Yahoo was from: https://www.google.com/search?q=stock+market+in+2018&oq=stock+market+in+2018&aqs=chrome..69i57j0i512j0i22i30l8.3745j1j9&sourceid=chrome&ie=UTF-8