Time Series Analysis of $SPY

By Alex Wako (awako@ucsb.edu)

Abstract

The $SPY ETF or the Standard and Poor's 500 ETF is described by State Street Global Advisor as an ETF composed of "selected stocks from five hundred (500) issuers, all of which are listed on national stock exchanges and spans over approximately 24 separate industry groups." The price of $SPY corresponds with the price and yield of the S&P 500 Index. It is a popularly traded stock in the market, consistently traded daily with high volume.

The goal of the project is to statistically analyze the price of $SPY using time series models and possibly forecast future prices. The methods used in this project are a Seasonal ARIMA model and GARCH model. The results of the analysis found from each model showed that the high volatility and the influence of unquantifiable variables within the stock market made it difficult to create a model with a good fit.

1. Introduction

The business of a stock trader is a constant use of modeling to understand the direction the stock will move towards. Time series modeling is a tool used by many large institutions to gain an advantage over other groups. Having an advantage means having a stronger understanding of the market and partial control of the stock; therefore, for the long run, it would result in more profits. The goal of the project is to understand how complicated a model must be to benefit the trader, discover the effectiveness of simpler models such as SARIMA, and possibly predict future prices of $SPY.

This project takes a look at the historical weekly closing price of $SPY and applies statistical analysis using time series models with the R programming language. Libraries used in R code for the project are quantmod for pulling $SPY data out of Yahoo Finance and astsa, xts, and fGARCH for time series analysis. These tools are open for anybody to use, so learning if they are effective in predicting future stock prices can make almost anybody an elite trader. The truth is the stock market and programming existed for a while, and the tools have always been open to use, so anything more than suboptimal is an unexpected result. Nevertheless, the main purpose of this project is to understand why a simple model in R does not work with volatile stock data and figure out what factors a model has to consider when analyzing the stock market.

2. Data

The $SPY data used in this project is found in Yahoo Finance and provided by the quantmod library in R. The quantmod library contains a function getSymbol() which allows for different sets of data to be pulled from various sources around the internet. The time range used in this project is from the first trading day of $SPY, January 29th, 1993, to the present day. Another set of analyses is applied to $SPY from March 12, 2021, to the present day. The data's frequency is split into the weekly closing price of the stock, excluding the dividends distributed in the underlying stock. There are 1571 data points in the first set of data and 104 data points in the second set of data. The reason for choosing this data is to properly analyze $SPY and discover any trends or seasonality that a trader can follow. The stock market is a difficult place to properly invest in money, and any kind of information can potentially give an investor an advantage over their competition. With that advantage, it is an opportunity for a trader to earn considerable money and achieve the financial freedom that so many people desire. As quantmod collects data directly from Yahoo Finance, and Yahoo Finance collects data directly from the NYSE, the data is reliable.

3. Methodology

SARIMA - The SARIMA model or the Seasonal Autoregressive Integrated Moving Average is a time series model that extends the ARIMA model to include seasonality. The model takes in all components of the ARIMA model and adds a seasonal trend. The different parts of SARIMA are p, d, q, P, D, and Q, where the upper case variables represent the seasonal component. p and P or the Autoregressive component of the model understands the relationship between the effect of past values with the current value. The value of p and P depends on the number of lag that correlates with the current value. d and D or the Integrated component of the model is the difference to remove any trend and/or seasonality within the data. q and Q or the Moving Average component of the model understands the residuals of past values to create a linear regression for the current value. For the case of this project, the Akaike Information Criterion is used to determine the best model for the data.

GARCH - The GARCH model or Generalized Autoregressive Conditional Heteroscedasticity is applied to a data on the assumption that the variance of is not constant. The model takes an ARMA component and the variance to describe the data. GARCH models are derived from ARCH models which add on the lagged conditional variance to the model. The GARCH model is commonly used in financial data sets as it attempts to understand the volatility present in the stock market and apply its effect to future values.
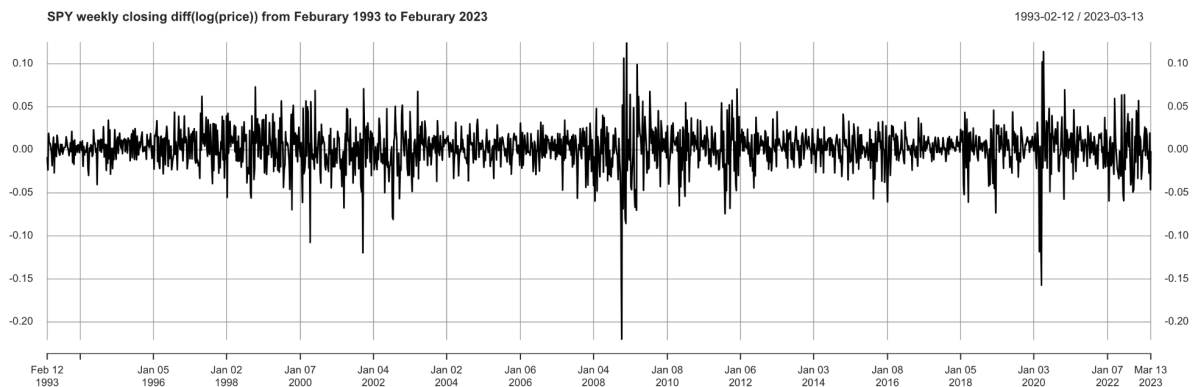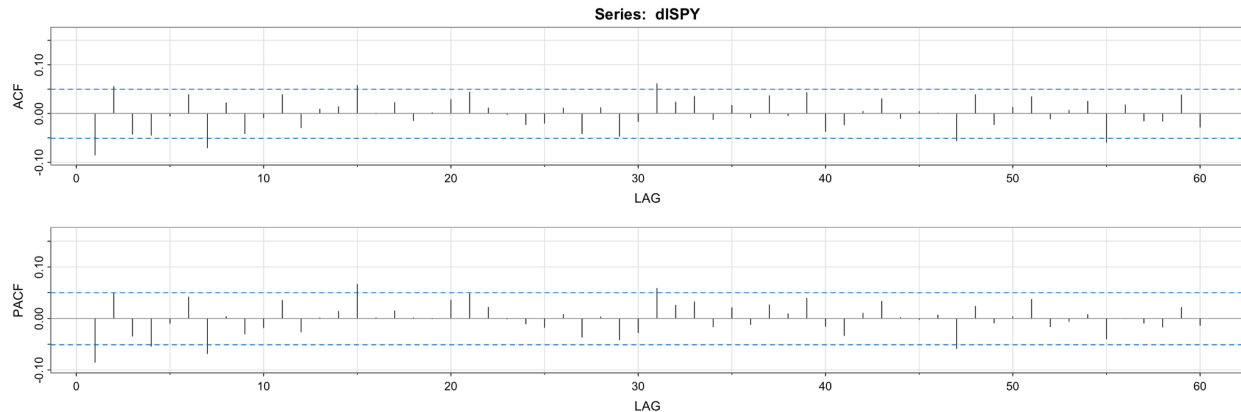
4. Result

SARIMA Model

The original plot of the data showed an inconsistent trend that increased the price of $SPY as time passed.

The original plot shows that there was no visible seasonality in the data. By transforming the data by using log differencing, the data can reveal possible seasonality or new trends.
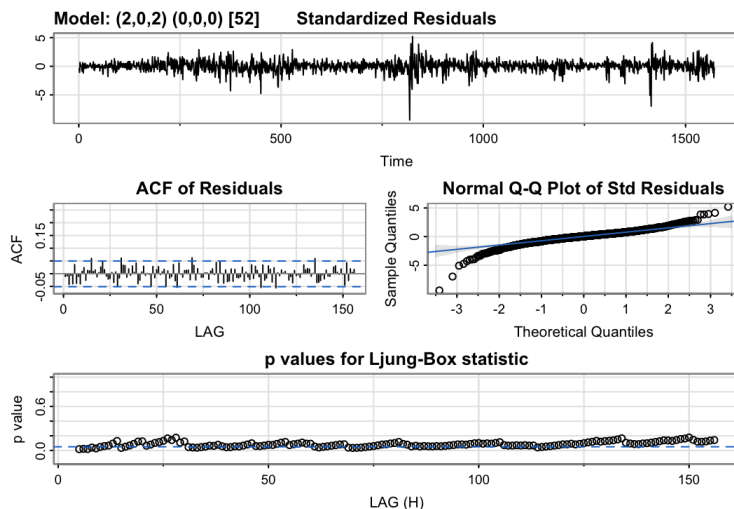
Series: dISPY

After log differencing and plotting the ACF and PACF of the transformed data, the data shows stationarity with AR1 and MA2. Seasonality does not show in the plot, so seasonality is ignored during the process of choosing the best model.

Multiple SARIMA models are then fit into the data, starting with SARIMA$(1, 0, 0)$ x $(0, 0, 0)_{52}$ and SARIMA$(0, 0, 2)$ x $(0, 0, 0)_{52}$. Using the AIC, the best model is found to be SARIMA$(2, 0, 2)$ x $(0, 0, 0)_{52}$. The diagnostics show that the model is stationary and all the predictors of the model are statistically significant. The Q-Q plot does not prove normality as the plot shows that the data is left skewed and the Ljung-Box statistic does not reject the null hypothesis for all points of time, but this shows that the SARIMA model does not work well with the $SPY data and another model approach

$ttable

|	Estimate	SE	t.value	p.value
ar1	-0.8730	0.1424	-6.1318	0.0000
ar2	-0.7217	0.1124	-6.4212	0.0000
ma1	0.8051	0.1403	5.7397	0.0000
ma2	0.7288	0.1050	6.9403	0.0000
xmean	0.0017	0.0006	2.8947	0.0038

had to be applied to return better results.

The forecast proves the statement as the SARIMA model predictions are less than ideal for achieving an advantage over other traders.
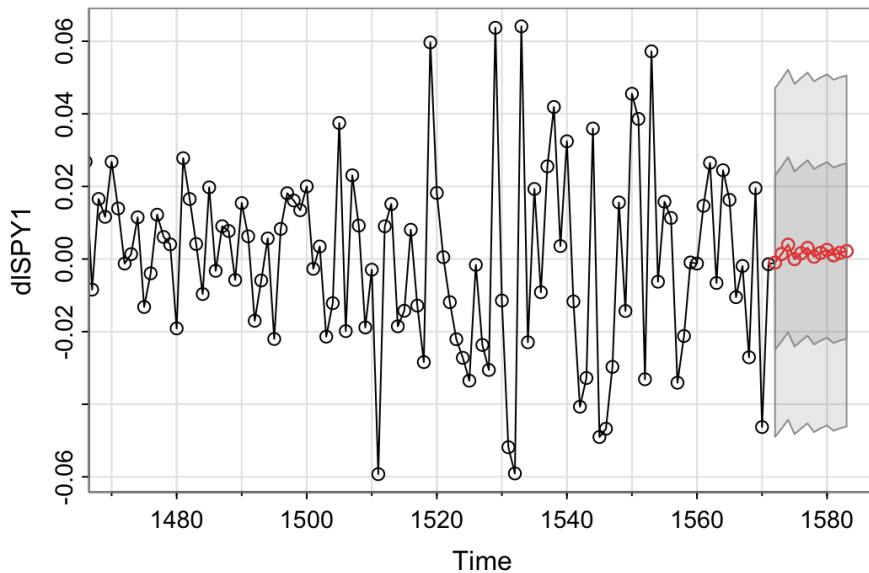
GARCH Model

As shown in the SARIMA model, the data has heteroskedasticity, meaning the GARCH model can be used.

Using results from the SARIMA model, a ARMA component of AR(1,0) and MA(0, 2) is considered when constructing the GARCH model.
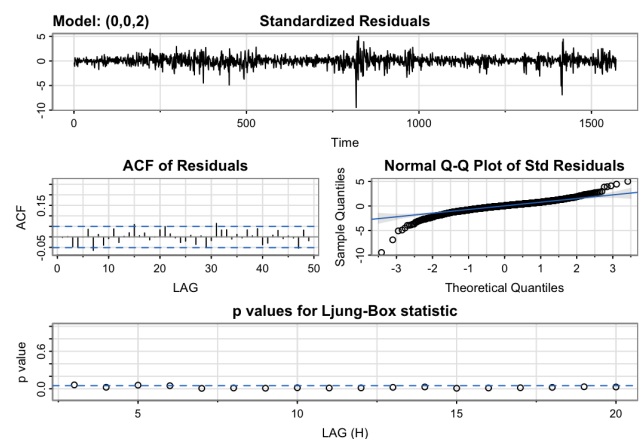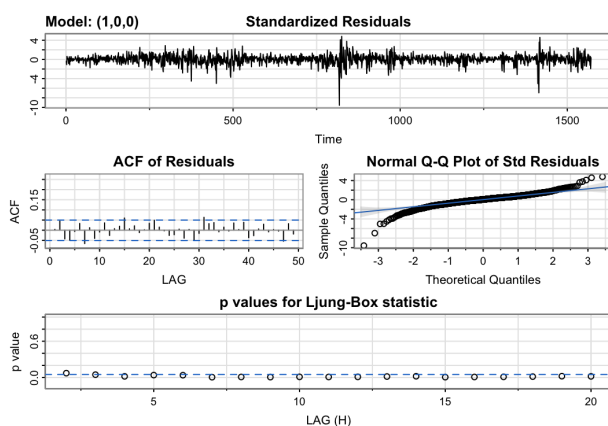
The diagnostics of the two ARMA models also show heteroskedasticity as normality is not shown in the residuals, Q-Q plot, and Ljung-Box test.

Multiple GARCH models are then fit into the data, using either GARCH(1, 1) or GARCH(1, 0) with the AR or MA components found earlier. Using BIC as the main criterion proposed by Javed(2011), the AR(1)-GARCH(1, 1) returned with the best result. The summary of the model returned with all variables being statistically significant, and the model passing all the tests used.

```
Error Analysis:
         Estimate  Std. Error  t value Pr(>|t|)
mu       3.148e-03  4.439e-04    7.092 1.32e-12 ***
ar1     -1.027e-01  2.772e-02   -3.706  0.00021 ***
omega    2.758e-05  6.280e-06    4.392 1.12e-05 ***
alpha1   2.362e-01  2.890e-02    8.175 2.22e-16 ***
beta1    7.348e-01  2.862e-02   25.675  < 2e-16 ***

Standardised Residuals Tests:

                             Statistic p-Value
 Jarque-Bera Test    R    Chi^2  402.9023  0
 Shapiro-Wilk Test   R    W      0.972077  0
 Ljung-Box Test      R    Q(10)  8.687931  0.5619539
 Ljung-Box Test      R    Q(15)  14.62435  0.4787985
 Ljung-Box Test      R    Q(20)  21.81468  0.3506504
 Ljung-Box Test      R^2  Q(10)  6.450921  0.7760639
 Ljung-Box Test      R^2  Q(15)  7.66551   0.9364918
 Ljung-Box Test      R^2  Q(20)  10.80781  0.9510552
 LM Arch Test        R    TR^2   6.208724  0.9051967
```
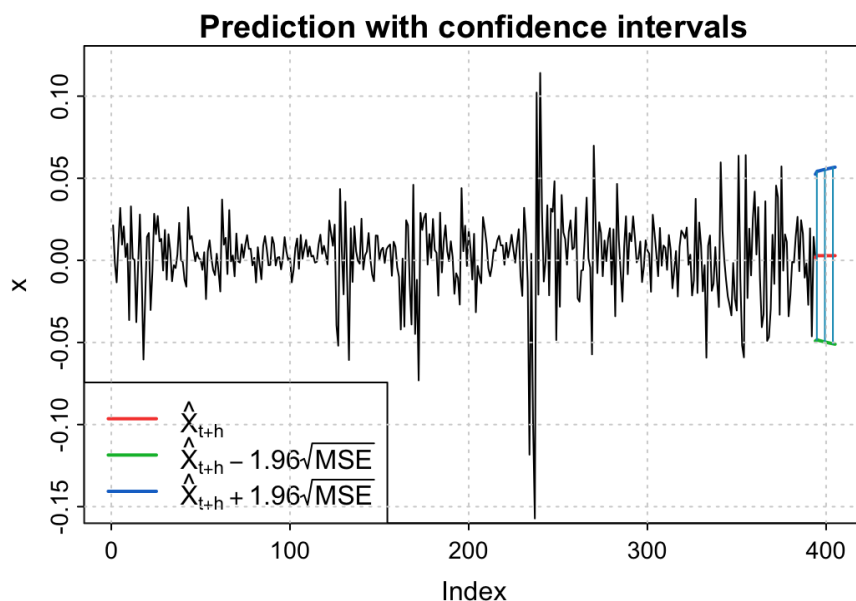
As a result of the prediction of 12 weeks ahead, the plot returned as a horizontal linear line. Same as the SARIMA model, the results are less than ideal for predicting future prices of $SPY, but this shows that one time series model is not enough to solve something as complicated as the price of $SPY.

Prediction with confidence intervals

5. Conclusion

The results of the time series analysis do not show significant information toward predicting the future price of $SPY. In fact, the results show that volatility and outside factors are difficult to predict with just one model.

Some factors that may have affected the outcomes were the size and details of the data set. From 1993 to 2023, a total of three stock market crashes occurred: the dot com bubble from 2000-2002, the global financial crisis from 2008-2009, and the Covid-19 pandemic in 2020. These events likely affected the capability of the models to predict future prices as they brought increased volatility to the already volatile stock market and may have ruined seasonal trends that occurred.

Additionally, weekly closing prices can bring difficulty in understanding the in-depth structure of the stock market and $SPY. $SPY is an index consisting of multiple stocks selected by a group of people. The stocks making up $SPY change quarterly, and the percentage holding of each stock can change daily. If one of the higher holding stocks were to crash or pump and another holding stock was to do the opposite reaction and recover the price of $SPY, the weekly $SPY data will never know about the change and never understand why a stock ended at a said price at the end of the week.

The results of the project showed that a single model could not predict the price of a stock as complicated as $SPY. A more sophisticated model that can understand the effect of the 500 holdings stocks relative to $SPY will more accurately predict the future prices of $SPY. Volatility is also difficult to understand with one time series model and is likely the biggest factor in the outcome of the project.

References

Baker, B. (2022, October 28). *Biggest stock market crashes in US history*. Bankrate. Retrieved March 20, 2023, from https://www.bankrate.com/investing/biggest-stock-market-crashes-in-us-history/

Javed, Farrukh, Garch-Type Models and Performance of Information Criteria (May 31, 2011).

State Street Corporation. (n.d.). *Spy: SPDR® S&P 500® ETF Trust*. State Street Global Advisors. Retrieved March 20, 2023, from https://www.ssga.com/us/en/intermediary/etfs/funds/spdr-sp-500-etf-trust-spy

Shumway, R. H., & Stoffer, D. S. (2017). In *Time series analysis and its applications: With R examples*. essay, Springer.

```r
library(quantmod)
library(astsa)
library(xts)

getSymbols("SPY", from = "1993-01-31")

SPY <- to.weekly(SPY$SPY.Adjusted)
tsSPY1 <- SPY$`SPY$SPY.Adjusted.Close`

# Plotting data

plot(tsSPY1, ylab = "Closing price of SPY", main = "SPY Weekly Closing Price from Feburary 1993 to Prese
plot(tsSPY1, ylab = "Closing price of SPY", main = "SPY Weekly Closing Price from Feburary 1993 to Prese

# Transformation

dlSPY1 = diff(log(tsSPY1))[-1]

# Plotting transformation

plot.ts(dlSPY1, main = "SPY weekly closing diff(log(price)) from Feburary 1993 to Feburary 2023")

# Spotting trends

lag1.plot(dlSPY1, 20)
acf2(dlSPY1, 60)
try(decompose(dlSPY1)) # Returns an error that the time series has no or less than 2 periods

# A seasonal trend does not seem to exist and the ACF and PACF show to be stationary
# The ACF and PACF show the model is likely has AR(1) and/or MA(2) for the first set of data
sarima(dlSPY1, 1, 0, 0, 0, 0, 0, 52)
sarima(dlSPY1, 0, 0, 2, 0, 0, 0, 52)
sarima(dlSPY1, 2, 0, 0, 0, 0, 0, 52)
sarima(dlSPY1, 2, 0, 2, 0, 0, 0, 52)
sarima(dlSPY1, 0, 0, 1, 0, 0, 0, 52)
sarima(dlSPY1, 1, 1, 0, 0, 0, 0, 52)

# The Seasonal ARIMA (2, 0, 2) x (0, 0, 0) shows the best results
# Forecasting 12 weeks ahead
forecast <- sarima.for(dlSPY1, 12, 2, 0, 2, 0, 0, 0, 52)

library(fGarch)

# Heteroskedasticity is seen in the time series data so a GARCH model is deployed
GARCH1 <- sarima(dlSPY1, 1, 0, 0)
GARCH2 <- sarima(dlSPY1, 0, 0, 2)

summary(m1AR <- garchFit(~arma(1, 0)+garch(1, 1), dlSPY1))
summary(m2AR <- garchFit(~arma(1, 0)+garch(1, 0), dlSPY1))
summary(m1MA <- garchFit(~arma(0, 2)+garch(1, 1), dlSPY1))
summary(m2MA <- garchFit(~arma(0, 2)+garch(1, 0), dlSPY1))
summary(m1ARMA <- garchFit(~arma(1, 2)+garch(1, 1), dlSPY1))
summary(m2ARMA <- garchFit(~arma(1, 2)+garch(1, 0), dlSPY1))
```

```r
summary(m1 <- garchFit(~garch(1, 1), dlSPY1))
summary(m2 <- garchFit(~garch(1, 0), dlSPY1))

predict(m1AR, 12, plot = TRUE)
```