Forecasting TQQQ

Sterling Hayden

Dr. Giovanni Petris

STAT 4013: STAT FORECASTING PREDICTION

Overview

This project revolves around the application of an ARIMA model to analyze real-world stock closing price data from TQQQ from Yahoo Finance, which represents the 3X leveraged daily returns of the NASDAQ-100 index (QQQ). The primary objective is to assess and evaluate the effectiveness of the ARIMA model in capturing the underlying patterns within the stock closing prices. The forecasts of this analysis will generate forecasts for the upcoming 30 days' closing prices, providing valuable insights into the potential future trends of TQQQ.

Dataset

A table of numbers and letters

Description automatically generatedUtilizing R's Quantmod package, I’ve downloaded TQQQ's daily open, high, low, close, volume, and adjusted prices over the past three years, spanning 753 trading days’ worth of data. A glimpse of the dataset is provided below:

To maintain consistency throughout the analysis, our emphasis will center on the closing price (TQQQ.Close), as this closely aligns with what investors typically observe in their brokerage accounts. Below, I present a visual time series of the closing prices for TQQQ:

A graph with lines and numbers

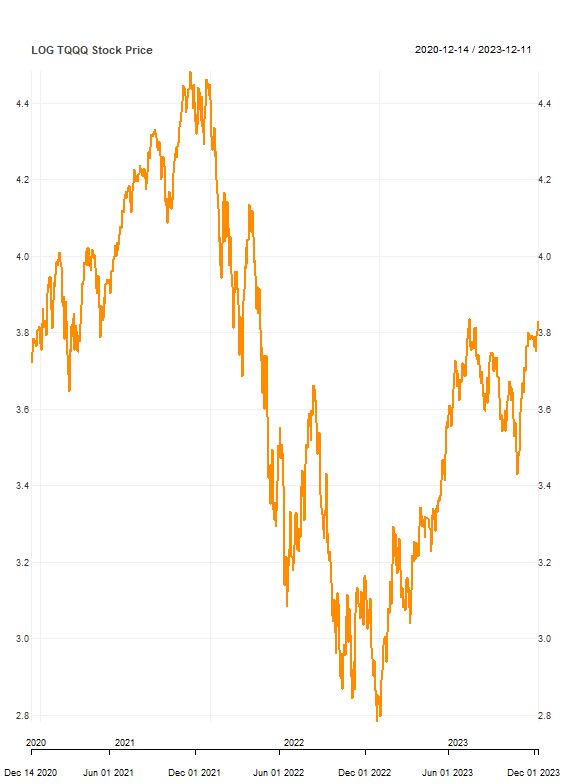
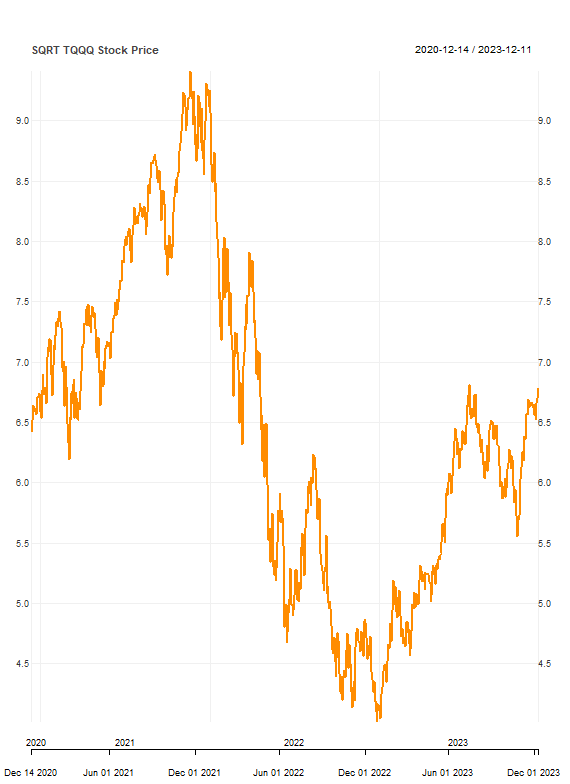
Description automatically generatedA graph of a graph

Description automatically generated with medium confidence

Alternatively, we can use qauntmods built in plot to get a better picture of TQQQ’s closing price, as shown above and to the right.

# Transforming The Data

The visualizations above suggest the presence of non-stable variance, a condition that does not align well with the assumptions of ARIMA models. In light of this, a crucial step in our analysis involves transforming the data to achieve variance stabilization. This transformation will not only address the non-stationary variance but also enhance the model's ability to discern underlying patterns and trends within the dataset. To accomplish this, we explore log and square root transformations, as illustrated in the visualizations below:



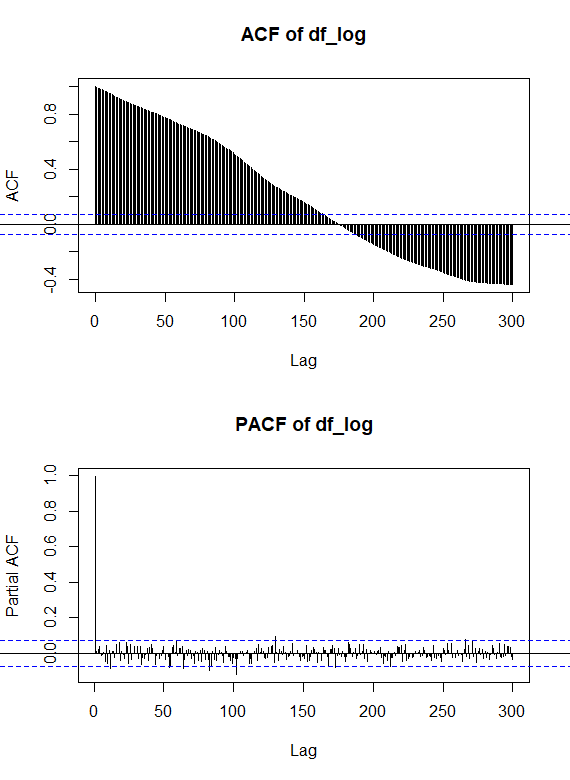
In the plots above, the effects of two different transformations are displayed: the square root transformation on the left and the log transformation on the right. A distinct pattern unfolds, clearly illustrating the superior variance-stabilizing performance of the log transformation, particularly during the period from June 2021 to June 2022. Recognizing this efficacy, I will employ the log transformation as a pivotal step to stabilize the data, aligning it with the assumptions inherent in the ARIMA model. The log transformation will allow for a more robust analysis, ensuring the model's accuracy in capturing the underlying dynamics of the time series data.

Checking Stationarity

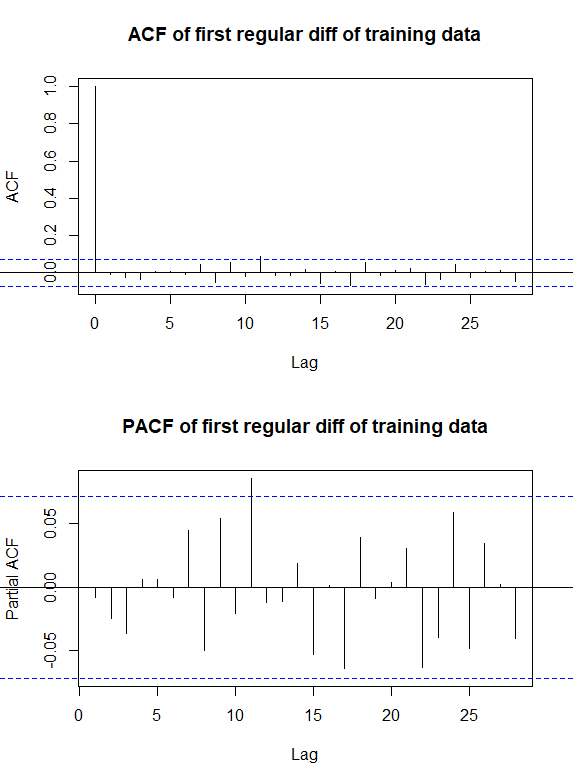
A stationary time series, characterized by a consistent absence of trend, maintains a constant mean and variance over time, rendering it conducive to precise predictions. Following the log transformation of our data, discernible patterns emerge, revealing dynamic fluctuations within specific time intervals. For instance, spanning the year 2020 to late 2021, the stock exhibits an upward trend, transitioning to a downward trend in the early months of 2021 through early 2022. These noticeable trends, coupled with inherent volatility and oscillations, are indicative of a non-stationary stock price movement. It is noteworthy that a majority of financial data inherently tends to be non-stationary, manifesting variances in mean, autocorrelation, and volatility over time. Often adhering to a random walk model, these fluctuations reflect the intrinsic nature of stock behavior, encompassing trends and varying levels of mean and variance within distinct periods.

Differencing

In econometrics, it's widely acknowledged that applying a 1-lag difference to a random walk model can effectively render it stationary. Yet, when dealing with the resulting differenced data, it's not uncommon to encounter missing values, requiring the task of addressing those gaps. In my approach, I opted to fill in these missing values by substituting them with the last observed value following the application of a 1-lag difference. Bellow is the non-differenced data’s respective ACF and PACF:



Before differencing we see that the autocorrelation is significant until around lag 160. In other words, the closing price values are correlated with values that are up to 160 trading days apart. Below Is the is the ACF and PACF after being differenced:



Choosing Model Order

For the readers understanding, ARIMA models take in the following parameters:

***p*** – denotes the order of the AR(p) model, identified to capture the short-term correlations of the stationary time series after differencing.

***P*** - denotes the order of the AR(P) seasonal model, identified to capture the seasonal correlations of the stationary time series after differencing.

***d*** – denotes the order of regular differencing needed to remove the contribution of the long-term trend to the nonstationary of a time series.

***D*** - denotes the order of seasonal differencing needed to remove the contribution of the regular seasonality to the nonstationary of a time series.

***q*** – denotes the order of the MA(q) model, identified to capture short-term correlations after differencing.

***Q*** - denotes the order of the MA(Q) model, identified to capture the seasonal correlation after differencing.

***F*** – the frequency of the time series.

The analysis of the Autocorrelation Function (ACF) plot reveals that all lags fall within the confidence interval, indicating the absence of significant correlations at any lag after differencing. This observation suggests a value of 0 for the q parameter in the ARIMA model, as there is no indication of a need for additional Moving Average (MA) terms to capture short-term correlations. Similarly, the Partial Autocorrelation Function (PACF) plot exhibits all lags within the confidence interval, further supporting the absence of significant correlations. This outcome implies a value of 0 for the p parameter, signifying that there is no need for additional AutoRegressive (AR) terms to model the short-term autocorrelations. The graph below provides a visual reference for these conclusions:

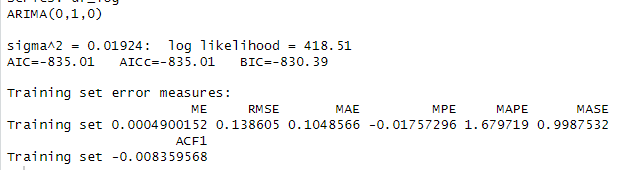
A graph of a graph of a graph

Description automatically generated with medium confidence

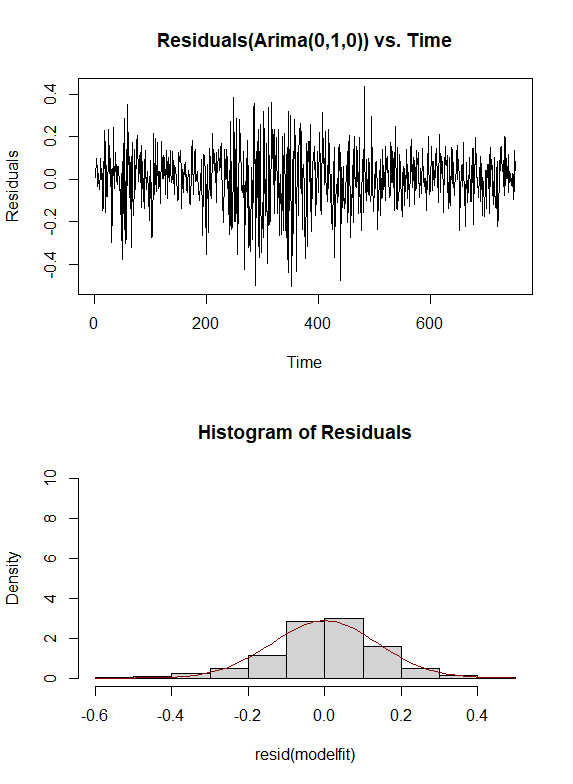
Continuing, given that we have differenced the regular time series only once, it has been established that d = 1. Assuming TQQQ's closing price doesn't follow a seasonal pattern, supported by the ACF and PACF plots displaying all lags within the confidence interval, we opt to set F, P, D, and Q to 0. This decision is based on the observed absence of notable correlations at various lags, indicating that the time series likely lacks any distinct seasonal trends.

Fitting the ARIMA Model

Given our p, q, P, D, Q, and F are all 0 the model would be the following:

We can validate this assumption by employing the auto.arima() function from R's forecast library. Upon applying our log-transformed and non-differenced data to the auto.arima() function, we consistently obtain the same model. This serves as confirmation that our parameter selection is accurate and aligns effectively with the characteristics of the time series data. Below is the summary of our models fit:  


Furthermore, we can look at the fitted models’ residuals below. The residuals of our fitted model show a mean of 0, indicating that the model effectively captures the overall trend or pattern in the data, with residuals centered around zero on average. However, the persistence of residual variance suggests that the model might not entirely account for all the variability in the data, leaving room for unexplained fluctuations or noise. This observation makes sense due the complex nature of using stock data; let alone 3X leveraged QQQ data. Moreover, it's noteworthy that these residuals also follow a normal curve, indicating a distribution resembling the familiar bell-shaped curve. This conformity to a normal distribution suggests that, on average, the residuals display a symmetric pattern around zero, further validating the appropriateness of our model.

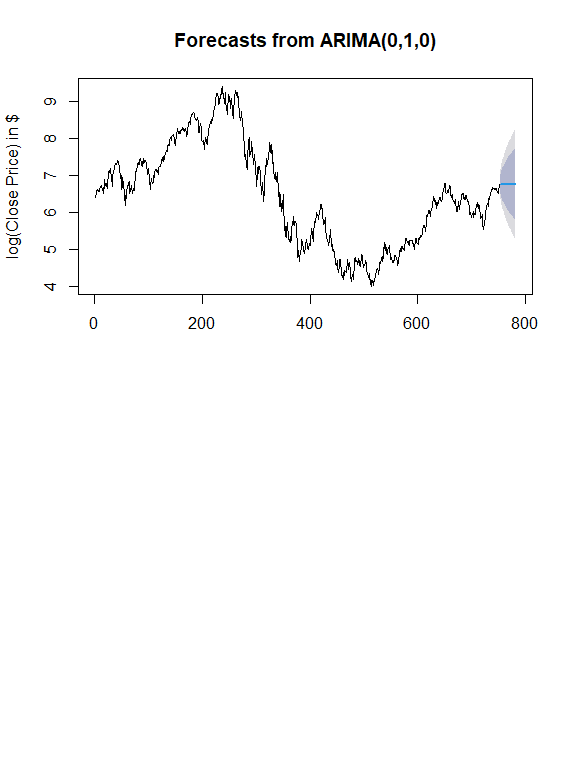


Lastly, we can look at the Ljung-Box statistic to test if the models’ residuals are autocorrelated. Upon testing we’re returned with a p-value of 0.82, thus we fail to reject the null hypothesis and do not have any reason to believe that the models residuals are non-autocorrelated. Below is the plot of p values for the Ljung-Box plot:

A graph with lines and dots

Description automatically generated

Forecasting

Pictured below is the 30-day forecast of TQQQ. The blue line represents the mean, the blue shaded area represents a 80% confidence interval, and the grey area represents a 95% confidence interval.

Delving further into the analysis, we can closely examine the day-by-day exact values.

A screenshot of a computer

Description automatically generated

These forecasted values offer insights into the potential future close prices of TQQQ. The point forecast, representing the mean, provides an estimated average value for each day in the forecast horizon. Alongside this, the prediction intervals, encompassing lower and upper bounds, offer a range within which the actual TQQQ close prices are likely to fall with a specified level of confidence. For instance, a 95% prediction interval indicates a 95% confidence level that the true close price will reside within that particular range.