Forecasting the S&P500 Index using ARIMA and VAR model

Doan Khanh (DK), Kar Yan Ong, and Federico Chung

December 16, 2019

Abstract

The work presented in this research is a contribution to modeling and forecasting the S&P 500, the financial health indicator for US Markets, by using both ARIMA and VAR models. Using the ARIMA model, our study predicted and estimated a slight increase of S&P 500 values over the following five day period. Using the VAR model with other US financial indexes (Russell 2000, Willshire 5000), a global market index (MSCI World), and a Korean index (KOSPI), we also obtain a slight increase in the prediction. The limitations of the use of historical data to forecast the future value of the S&P 500 due to the lack of discernible patterns are addressed in this research paper. The results can contribute to the existing estimates of future S&P 500 values.

I. Introduction

Predicting the stock market has been an important topic for both academia and the financial industry for many decades now. In recent years, the importance and value of stock's price history have reached enormous proportions in assisting and understanding the behavior of its future prices. Mainly, forecasting the S&P 500 Index has been of interest since its inception to several parties both inside and outside the financial world.

Forecasting the S&P 500 level is a critical analysis tool for both investors and economists. From an investor's perspective, the Standard & Poor's 500 Index is one of the most influential indices in the world. From an economist's view, the index is a broad indicator of the U.S. economy, in some instances, a global economic indicator as well.

There have been a considerable amount of studies trying to explore the future value of the S&P 500 Index. Due to the complexity and interest of this topic, the research approaches can be categorized into two main techniques: time series analysis and artificial neural network. According to (Khashei, Bijari, and Ardali 2009), Artificial neural networks (ANNs) are soft computing technology that is widely used as forecasting models in many areas, including social, engineering, economic, business, finance, foreign exchange, and stock problems. Youngohc Yoon and George Swales demonstrated in (Yoon and Swales 1991) that the neural network approach is capable of learning a function that maps inputs to output and encoding it in magnitudes of the weights in the network connection. It is indicated that the Neural Network approach can significantly improve the predictability of stock price performance. On the other hand, the ARIMA model, also known as the Box-Jenkins model or methodology, is commonly used in analysis and forecasting (Khashei, Bijari, and Ardali 2009). It is widely regarded as the most efficient forecasting technique in social science and is used extensively for time series. Additionally, the Vector Autoregressive (VAR) model is also one of the most famous empirical models for forecasting purposes. The VAR model can tackle the problem that multiple financial indexes may impact each other at the same time instead of treating a univariate time series (S&P500 index) as an independent.

This study aims to forecast the S&P 500 Index in the future shortly after the end of September in 2019 by using both ARIMA and VAR models. In order to achieve this goal, for ARIMA model, we use the historical data of the S&P 500 Index from 01/03/2000 to 09/27/2019. Several ARIMA models were developed and evaluated by Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC). Furthermore, the research also includes the Russell 2000, MSCI World, Wilshire 5000, and KOSPI Indexes to perform multivariate time series analysis in the VAR model. For VAR model, we use historical data of all indices from 2010, after the financial crisis.

The rest of the study is organized as follows: Section II and III reviews of relevant literature and methodologies used in this study. Section IV presents and discusses the results obtained in this study, while limitations and conclusions are provided in sections V and VI.

II. Literature Review

For the time series analysis on the S&P 500 Index, Paul Eulogio (Eulogio 2018) used 21 years (1995-2015) of S&P 500 Stock Index data at a monthly frequency (a total of 252 observations) from Yahoo Finance and the Adjusted Close. Through the Augmented Dickey-Fuller Test, he found that the data was nonstationary and first differenced the data to make it stationary. After taking a look at the ACF and PACF of the differenced data, he decided that ARIMA(0,1,1) was best.

Ionides(Ionides 2018) used the daily S&P500 data from 2013-03-05 to 2018-03-02. Then, he took differences for the log transformation of the Adjusted Close prices for S&P500 to get the log returns for the stock to eliminate the nonstationary characteristics. After that, by AIC criteria, he chose to fit ARMA(1,1) model for log returns. He then used ARIMA(1,1) model for the weekly S&P500 log-returns and found that the residuals performed better but were a less accurate model.

Over the past years, there have been many studies that have been made to investigate the co-movement of stock markets in different countries and how they could affect one another using Vector Autoregression Models.(Eun and Shim 1989) investigated the international transmission mechanism of the stock market by estimating a nine international market vector autoregression system using daily rates of return on the stock market. Their study found "that the U.S. stock market is the most influential in the world" (Eun and Shim 1989, pg.8).

In a more recent study, (Yang, Kolari, and Min 2003) examined the long and short-run relationships among the U.S., Japanese, and ten Asian emerging stock markets, particularly during the Asian financial crisis using the VAR model. He found that "The U.S. substantially influenced the Asian markets in all three sample periods but was almost unaffected by the Asian markets" (Yang, Kolari and Min 2003, pg. 17).

Mukherjee and Bose (Mukherjee and Bose 2008) used techniques of cointegration, vector autoregression, vector error-correction models, and Granger causality to find that there is significant information leadership from the U.S. market to all Asian markets. Japan also plays a role in the integration of Asian markets. They also found that the U.S. market is also influenced by information from most of the major Asian markets.

III. Methods

Data & Variables of Interest

The financial indexes in this research were collected from the Yahoo Finance database. The S&P 500 dataset includes daily data on adjusted close, opening, and closing values and the trade volume for every day the stock market was open since "03/01/2000" (3rd of January, 2000). This report used the adjusted close value of the S&P 500 as it reflects the actual value of the S&P 500. The adjusted close values adjust for corporate actions, such as dividends, that affect the actual value of a stock/ index. Additionally, they adjust for historical differences in corporate actions, and it can be used to examine historical returns as it provides an accurate representation of the real equity value over time.

Unlike regular publicly traded stocks, the S&P 500 is a benchmark index that helps investors understand the health of the U.S. economy. The index accounts for 80% of all U.S. market capitalization. The S&P 500 value is calculated by adding the price of all 500 companies that encompass it weighted by their market capitalization. This means that fluctuations in the stock price of companies with higher market capitalization have a more significant effect on the S&P 500 compared to fluctuations in the stock price of companies with low market capitalization. The research also collected other U.S. financial indexes (Russell 2000, Willshire 5000), a global market index (MSCI World), and a Korean index (KOSPI) to perform multivariate time series analysis.

Russell 2000 Index follows the performance of around 2,000 U.S. small-cap firms. Like the S&P, the index is weighted and regularly serves as a benchmark index. Willshire 5000 is a market-capitalization-weighted index of the market value of all US-stocks actively traded in the United States. This goes far beyond the reach of the S&P 500, which does not cover stocks with market caps under around $6 billion. On the other hand, the

MSCI World is a market-cap-weighted stock market index of 1,655 stocks from companies throughout the world. KOSPI is the primary stock market index of South Korea, which represents all common stocks traded on the Korea Exchange. By taking into consideration of other indexes on a global scale, from another market outside of the U.S., the research hopes to provide a more accurate time series analysis in predicting the S&P 500 index.

The collected data have time discontinuity. In order to fix the discontinuity, the research uses the Karman filter to find predictions for missing values. Kalman filtering is an algorithm that provides estimates of some unknown variables given the measurements observed over time. This process is necessary for reducing problems when measuring errors.

Statistical Methods

The following approach of the research is to explore the data and diagnose for trend, seasonality, heteroskedasticity, and stationarity. A stationary time series is one whose properties do not depend on the time at which the series is observed. The research uses the differencing approach to adjust for non-stationarity in the data. This approach helps stabilize the mean of the time series by removing changes in the level of a time series and therefore eliminating (or reducing) trend and seasonality.

After capturing both the trend and all possible seasonalities of the data, the Auto-Correlated Functions (ACF) and Partial Auto-Correlated Functions (PACF) are used to evaluate possible ARIMA models that can provide the best model the errors. Ideally, a good model should suggest normality of standard residuals and has high p-values for Ljung-Box statistic, indicating that the errors are independent of each other.

In addition to the ARIMA model, the VAR model is used for forecasting purposes. In this report, we are interested in examining the structural inference from different indexes with the VAR model. We examine the resulting causal impacts of unexpected shocks or innovations to the S&P 500 from other indices using impulse response functions and forecast error variance decompositions. The impulse response function indicates that if there is a one-unit shock (a one-standard-deviation increase) in the error term of the variable (j) at a specific time (t) $\varepsilon_{j,t}$, how would the response variable change accordingly. The impulse response is defined as the derivative of the response value for $\varepsilon_{j,t}$.

A drawback of Impulse Response Functions is that it cannot be used to analyze contemporaneous reactions of the variables. The Impulse Response Functions are under the assumption that a shock will only affect $\varepsilon_{j,t}$ while other variables remain independent. This is not realistic because some of our indices are rather correlated, and a real-time shock to $\varepsilon_{j,t}$ will also lead to a contemporaneous shock to other indices as well. Hence, it will be difficult for us to see the actual response $Y_{j,t}$ when $\varepsilon_{j,t}$ is shocked.

We use the Orthogonal Impulse Response Function(OIRF) to solve this issue. The OIRF allows us to observe a shock in one index, which has no shock in other indices, so we can indeed observe the real-time effects of a shock on separate indices. The OIRF makes sure that each impulse does not affect other components because they are orthogonal.

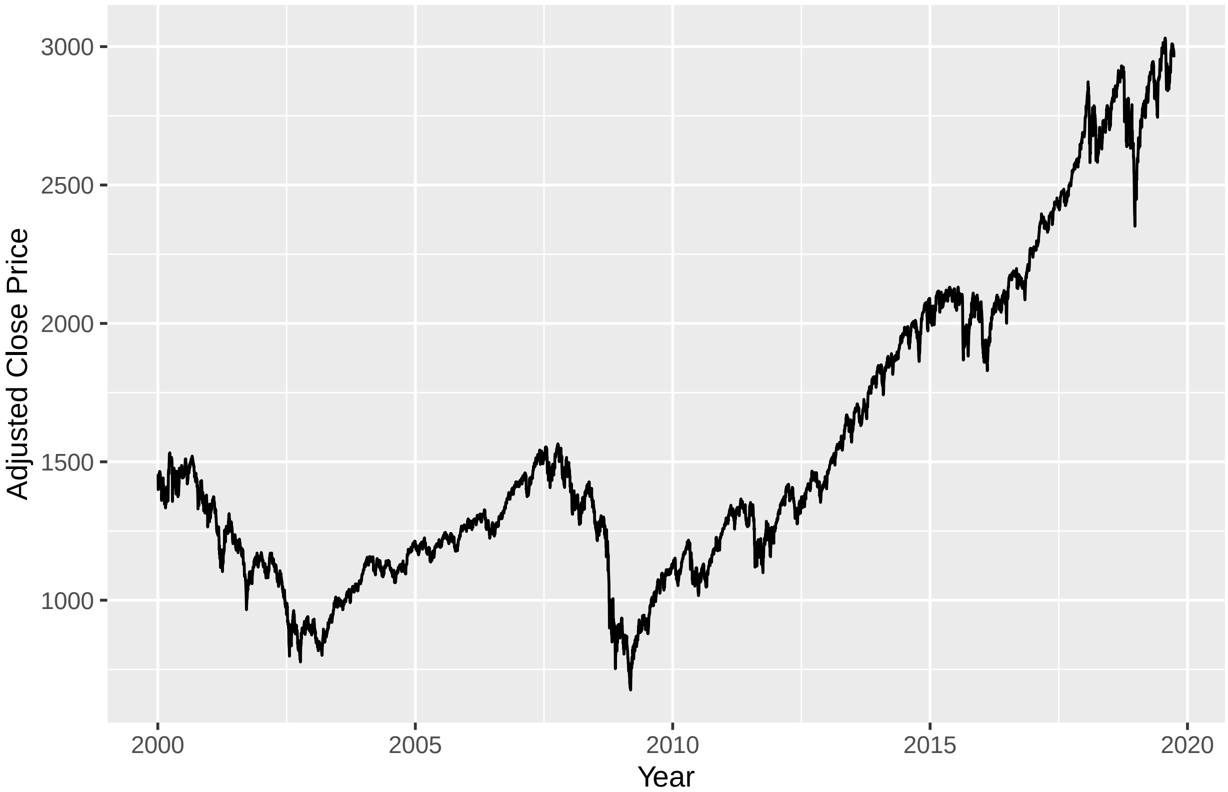
To accompany the OIRF, we use the Forecast Error Variance Decomposition. The FEVD measures the fraction of forecast variance for a variable that can be attributed to each of the driving shocks. In other words, we can see how vital each shock is through the variation of that particular index.

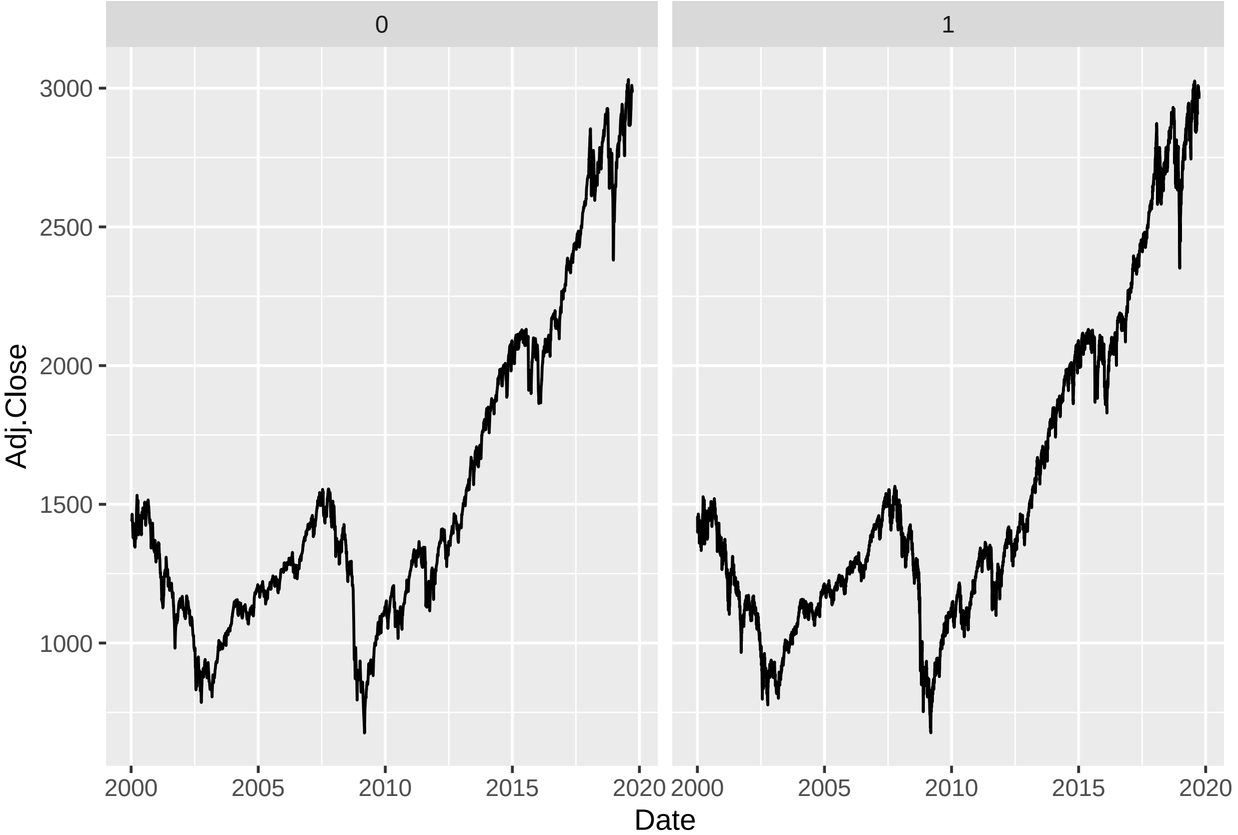
IV. Results and discussions

ARIMA Model

The report starts with the initial preprocessing of the data to make errors stationary. Looking at the first visualization, we notice that the S&P 500 is highly volatile, going up and down across times (similar to other stocks, indexes). Due to the characteristics of the financial data, we can observe that there is no clear seasonality from the visualization. However, there has been a high return from 2000 to 2019. We also observe two big drops due to recessions in 2003 and 2009. The time discontinuity problem is fixed in the data as the market closed during weekends and holidays with the use of the Kalman filter. As can be seen from the second visualization, the Kalman filter is relatively good in predicting the days that the market is closed.

Figure 1: S&P 500 Index from 2000 to 2019





The differencing method is then used to create stationary errors. The plot represents the residual series as well as ACF/ PACF of the errors. The first order differencing allows the study to remove the trend in error to adjust for the non-stationary problem. There is a drop in the ACF plot, while the PACF plot slowly decay. These plots help us confirm that we have stationarity and helps us decide which model we will use.

Figure 3: Estimated Residuals with differencing approach

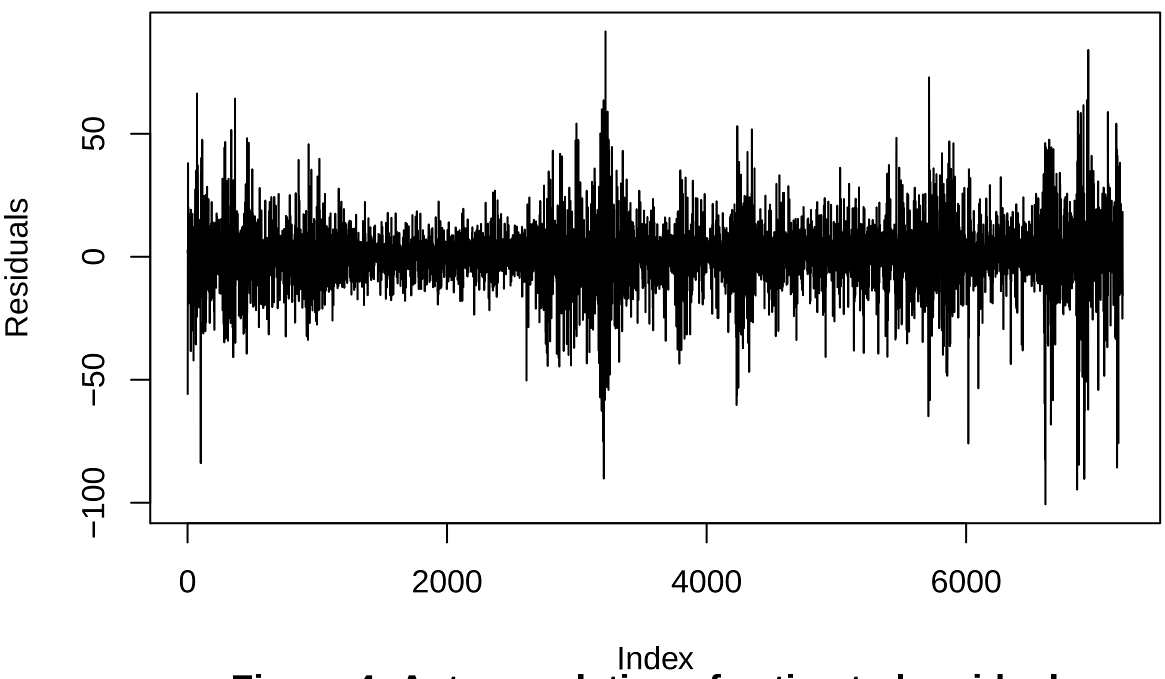


Figure 4: Autocorrelation of estimated residuals

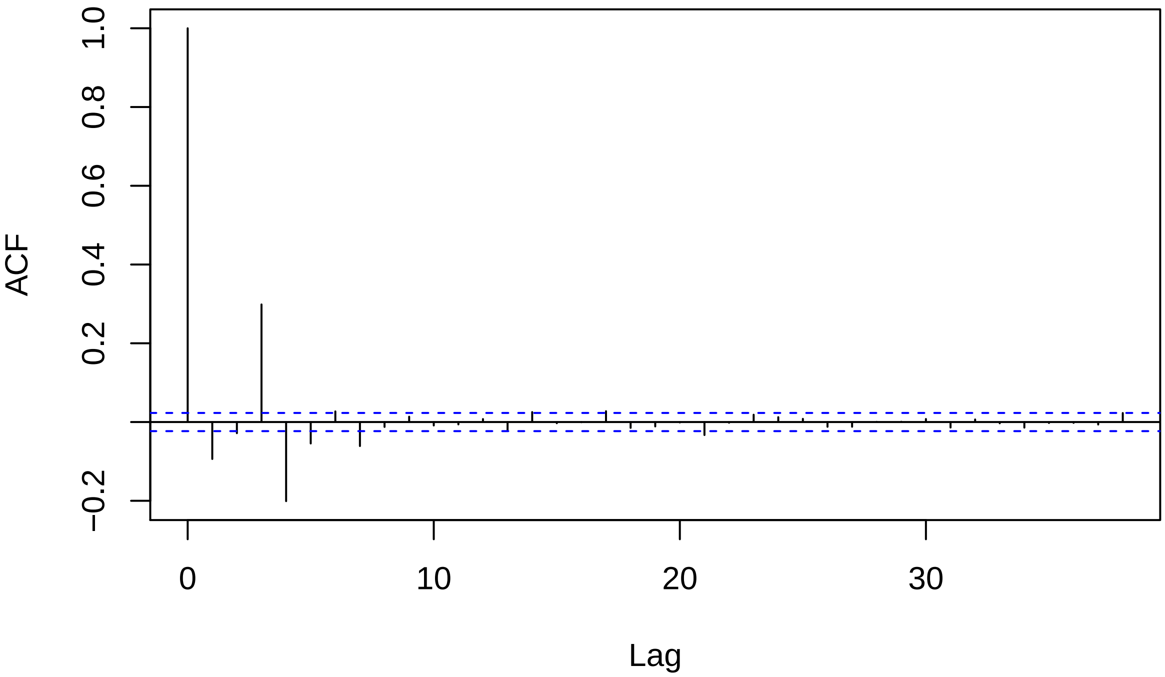
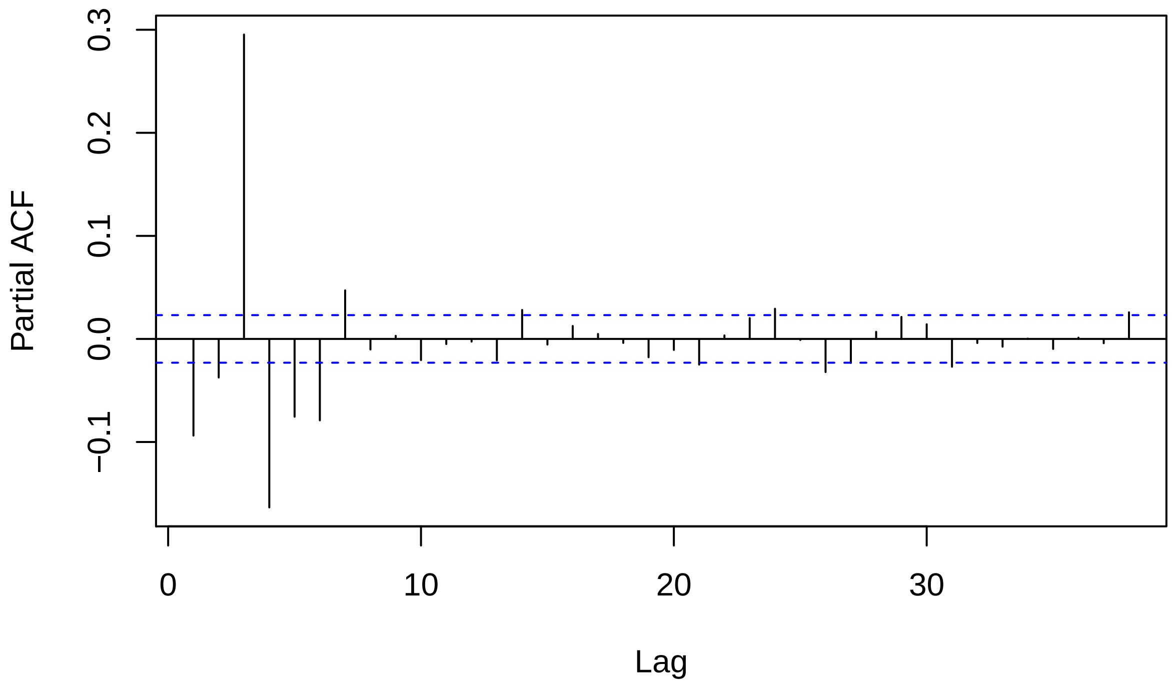
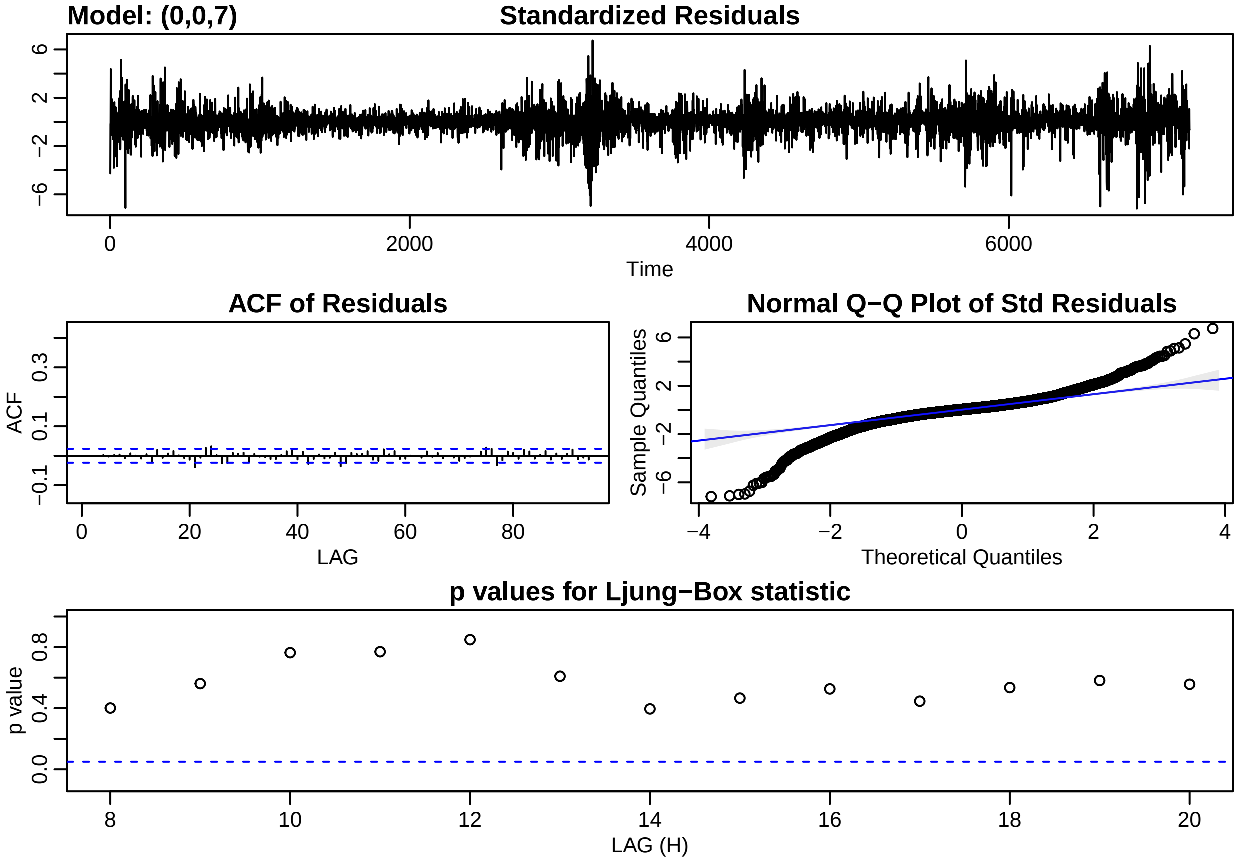


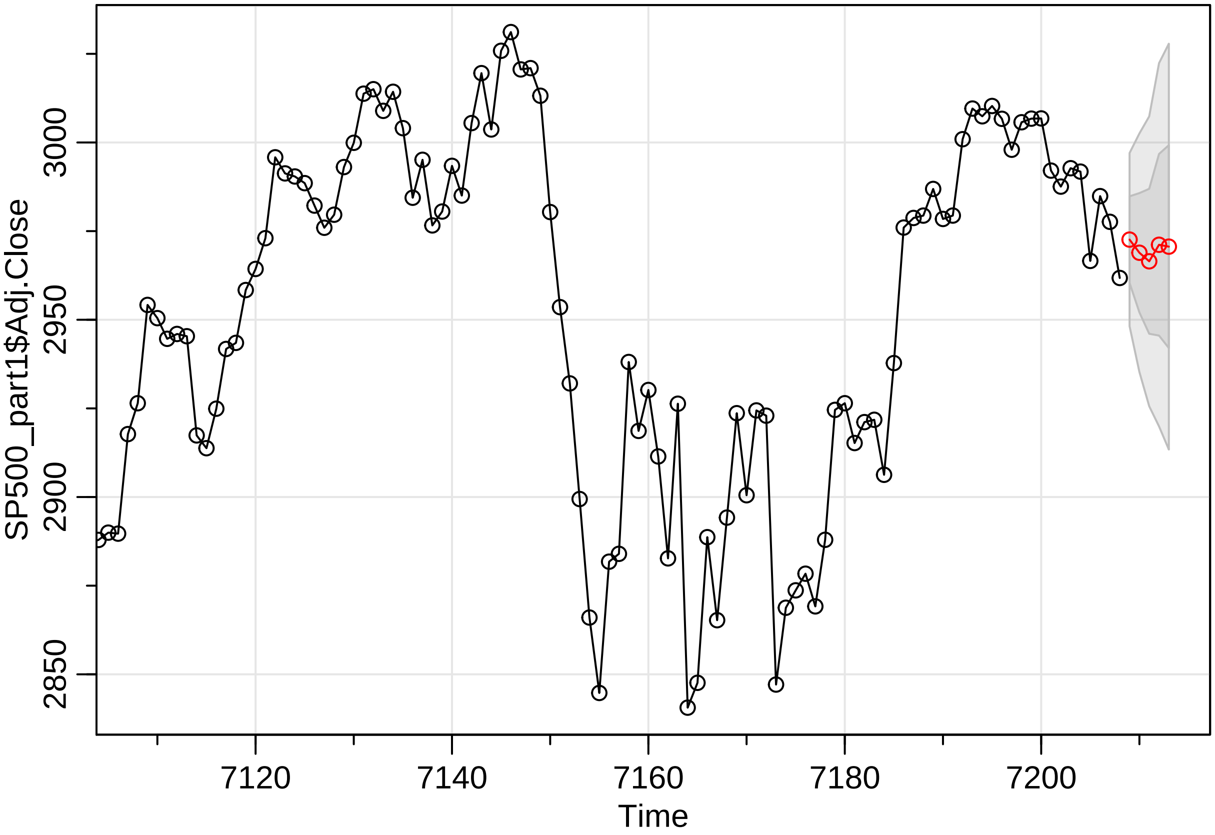
Figure 5: Partial autocorrelation of estimated residuals



After looking at the ACF and PACF graphs, potential candidates to fit our models are MA(1), MA(4), MA(5), and MA(7). Since the report examines the differenced time series, we have to use the combined model ARIMA (Autoregressive integrated moving average). By fitting the candidate models and comparing them, we can justify the choice of one model over the other models. Among four candidate models, it seems that the MA(7) is the best-fit candidate. The report also obtained the lowest AIC & BIC results when using MA(7) model. Moreover, as can be seen from the Ljung-Box statistic, the p-value is significant, indicating to reject the null hypothesis that there is autocorrelation within the lags. In other words, the lags can be used to determine the value of today.

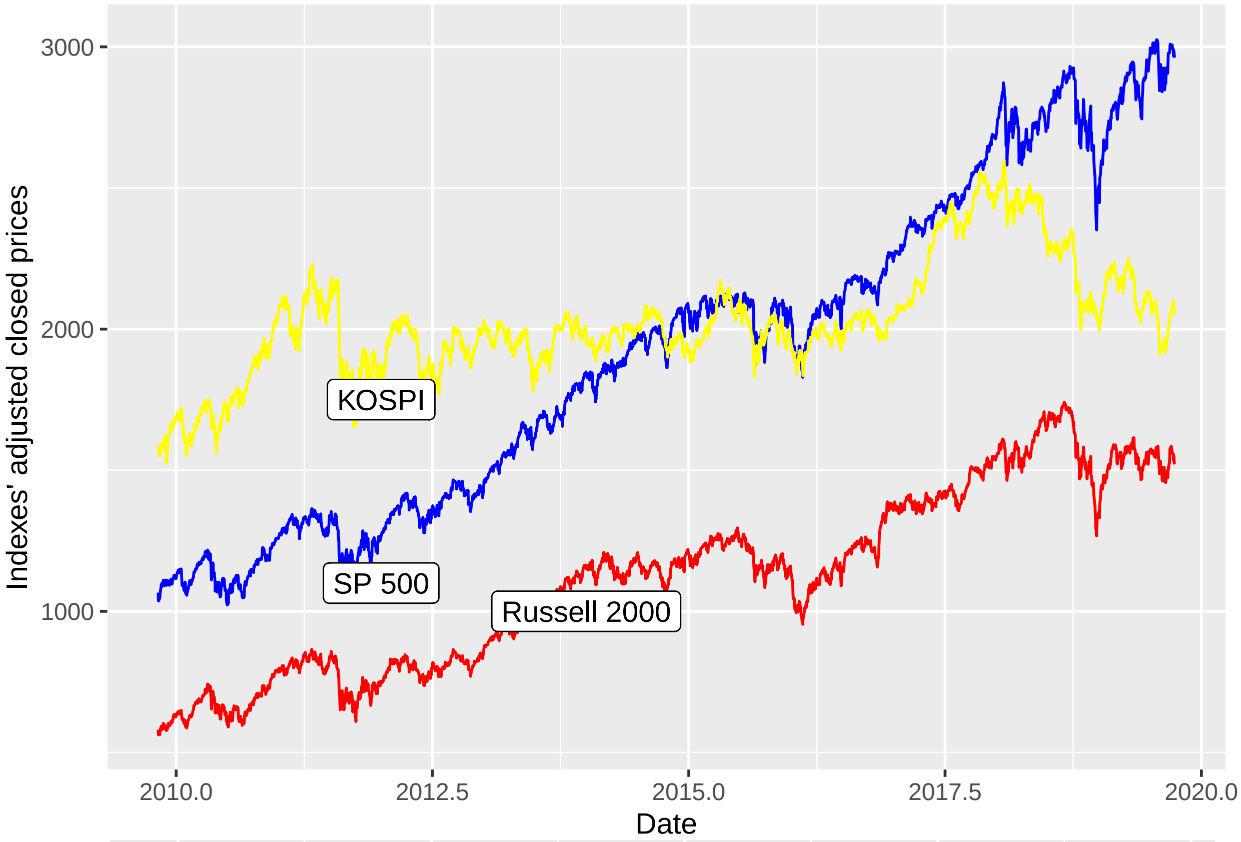


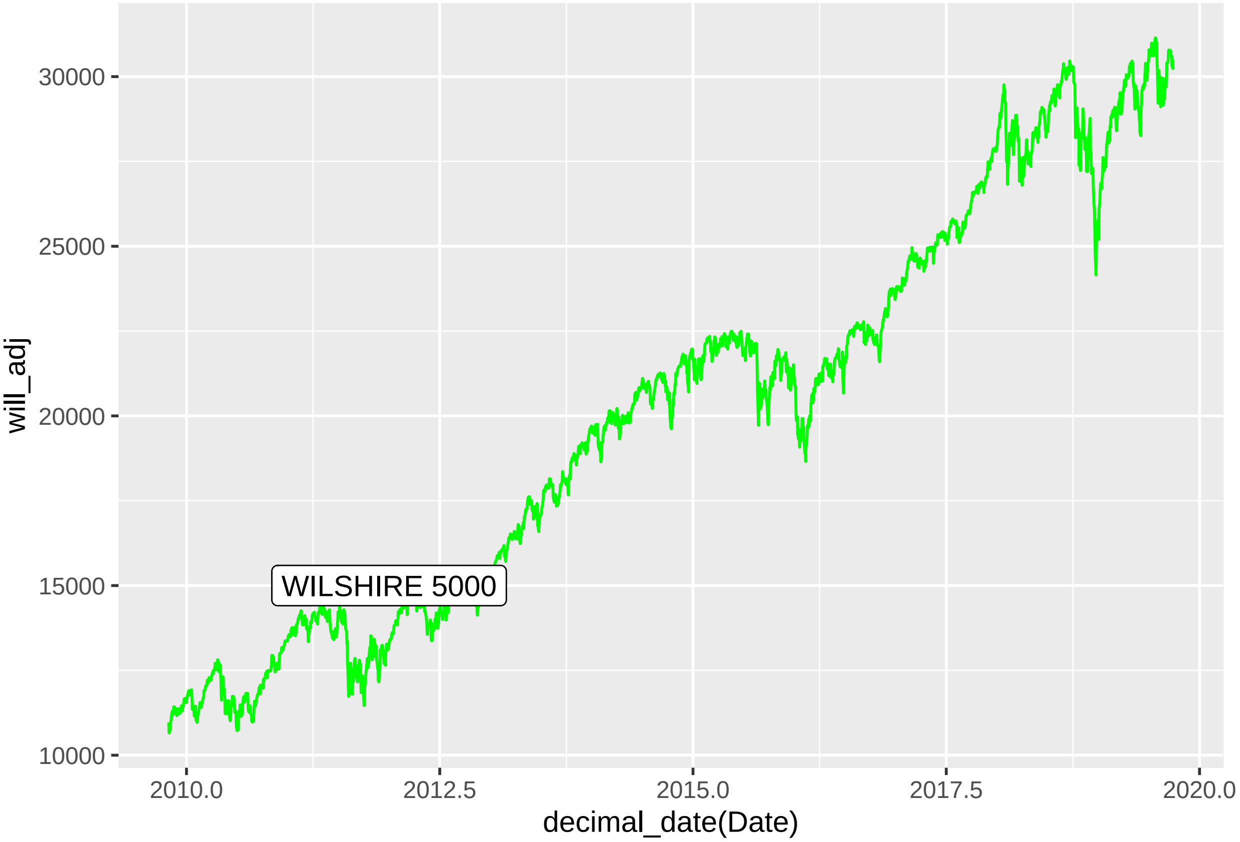
With the use of the MA(7) model, the report proceeds to predict the S&P 500 index in the following five days. Although there will be a drop in the predicting period, we observe an increasing trend overall.

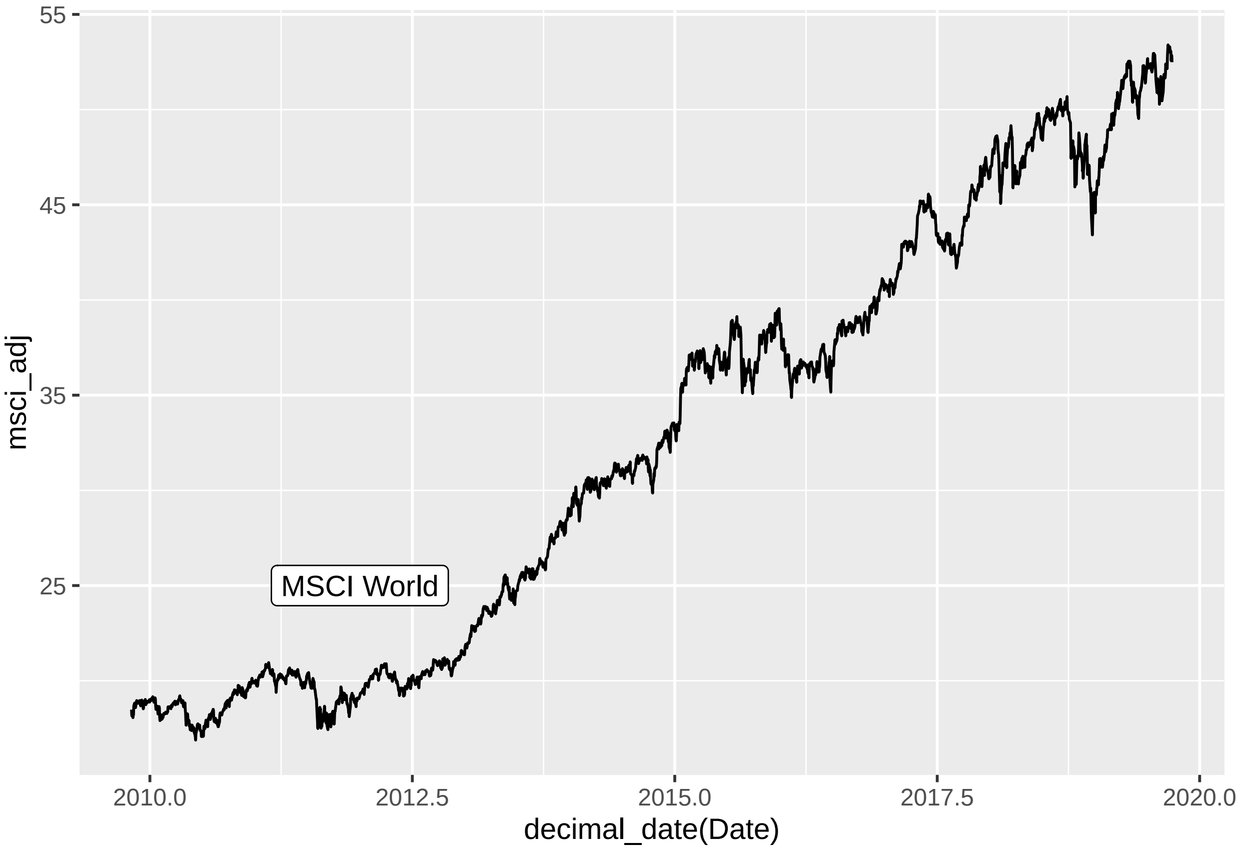


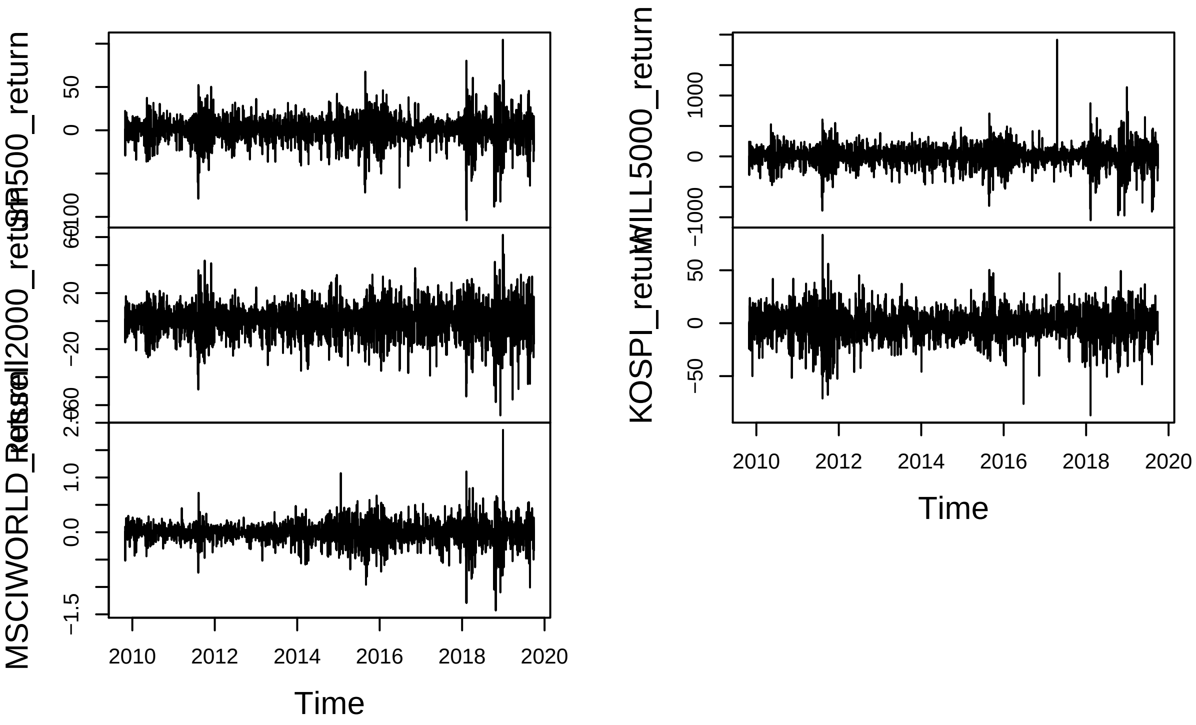
VAR Model

After using ARIMA model for prediction, the report then aims to perform multivariate time series analysis with the use of VAR model. For this model, we use the return values of each index by subtracting the close value to the open value. The data we used were from 2010 to 2019. Looking from the visualizations below, we see that most of indices have similar patterns except for KOSPI (Korean index).


Indexes' adjusted closed prices
3000
2000
1000
KOSPI
SP 500
Russell 2000
2010.0
2012.5
2015.0
2017.5
2020.0
Date


will_adj
30000
25000
20000
15000
10000
WILSHIRE 5000
2010.0
2012.5
2015.0
2017.5
2020.0
decimal_date(Date)

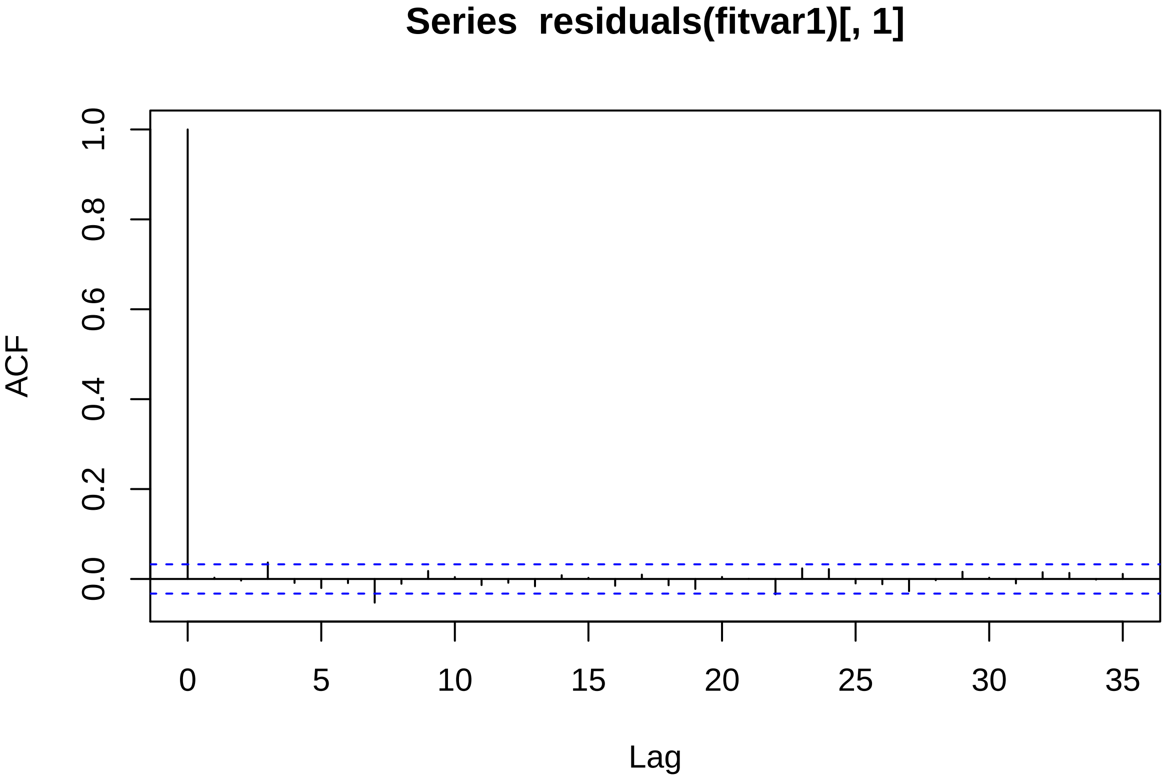


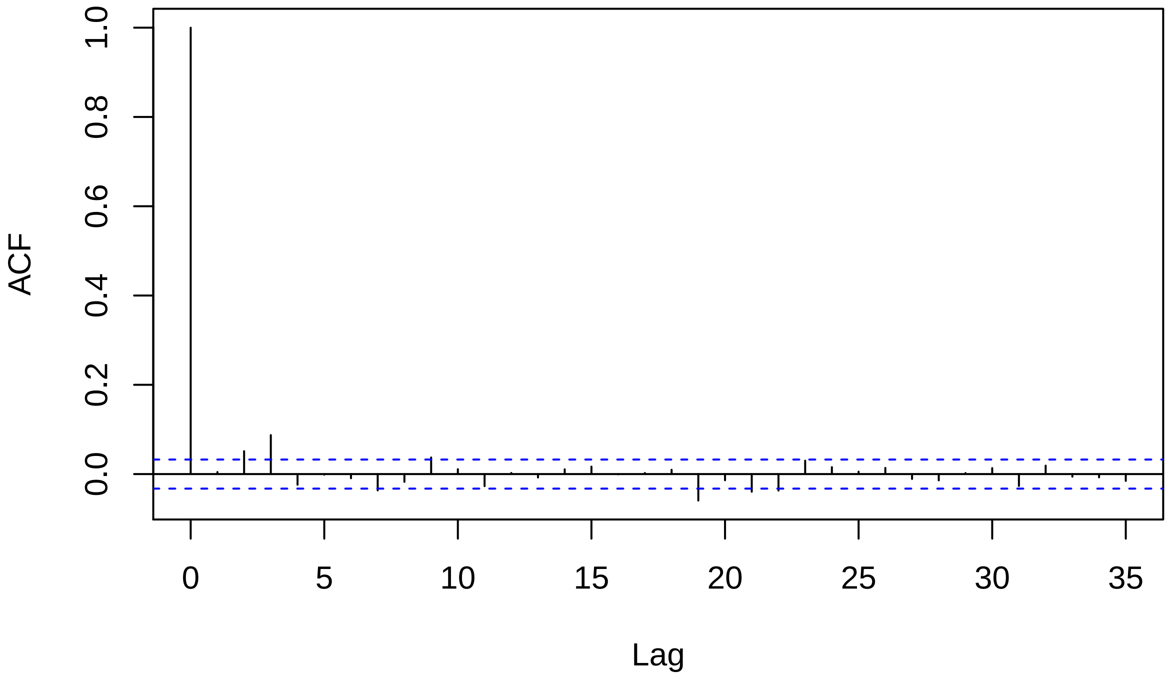

Times series of returns on different stock indices

From our VAR select function, we realize that the best models to use are either VAR(3) or VAR(1) models. Although AIC criteria are most widely used to evaluate the best model to use because we are trying to predict stock markets, which can be very variable from day to day, lag 1 VAR models make more sense. They provide us with a simpler model.

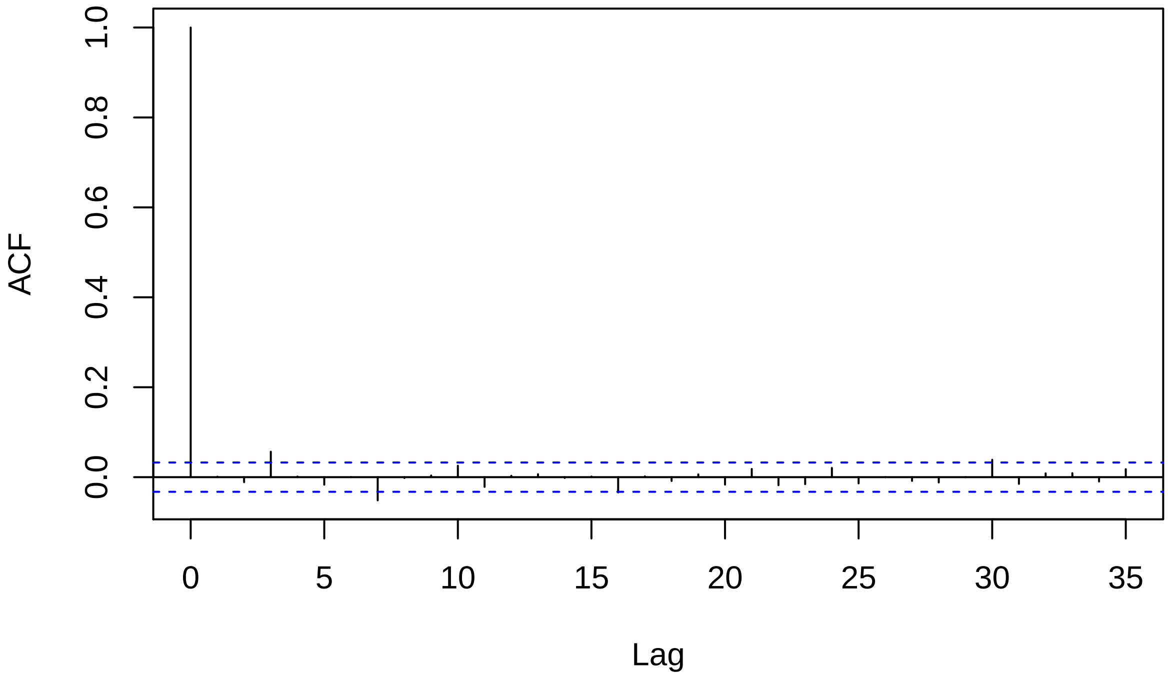
From the summary statistics of our VAR estimation results, we see that to predict the SP500, the most significant lag values come from the Russell 2000 index and the Will 5000 index, while one of the least significant lag values come from the KOSPI index. This makes sense as the Russell 2000 index and the Willshire 5000 index are both US Market benchmark indices, and they should relate the most with the SP500, which is another composite US index. Surprisingly in this study, we find that lagged values SP500 return is the least predictive lag. We would expect that the SP500 return lagged values to be the most significant lag value. This is prevalent across the indices as the SP 500 lagged return seems to be the least significant predictor across indices. All other indices have their own lagged values to be a significant predictor of current return. We also see that the best predictor for indices outside of the US, the KOSPI and the MSCI world the lagged values of each other are very significant to predict the current return. This suggests that the returns of some world indices are also highly correlated with each other. It makes sense as some companies in the MSCI world index are in the overall KOSPI Index, such as Samsung.

Finally, if we look at the correlation matrix of the residuals between the different index returns, the values make empirical sense as the SP500 should be most correlated with the Russell 2000 and the Willshire 5000 composite indices. However, the MSCI World and the KOSPI index are not correlated as we would expect them to be even though we saw that its lagged values were significant in predicting each other.

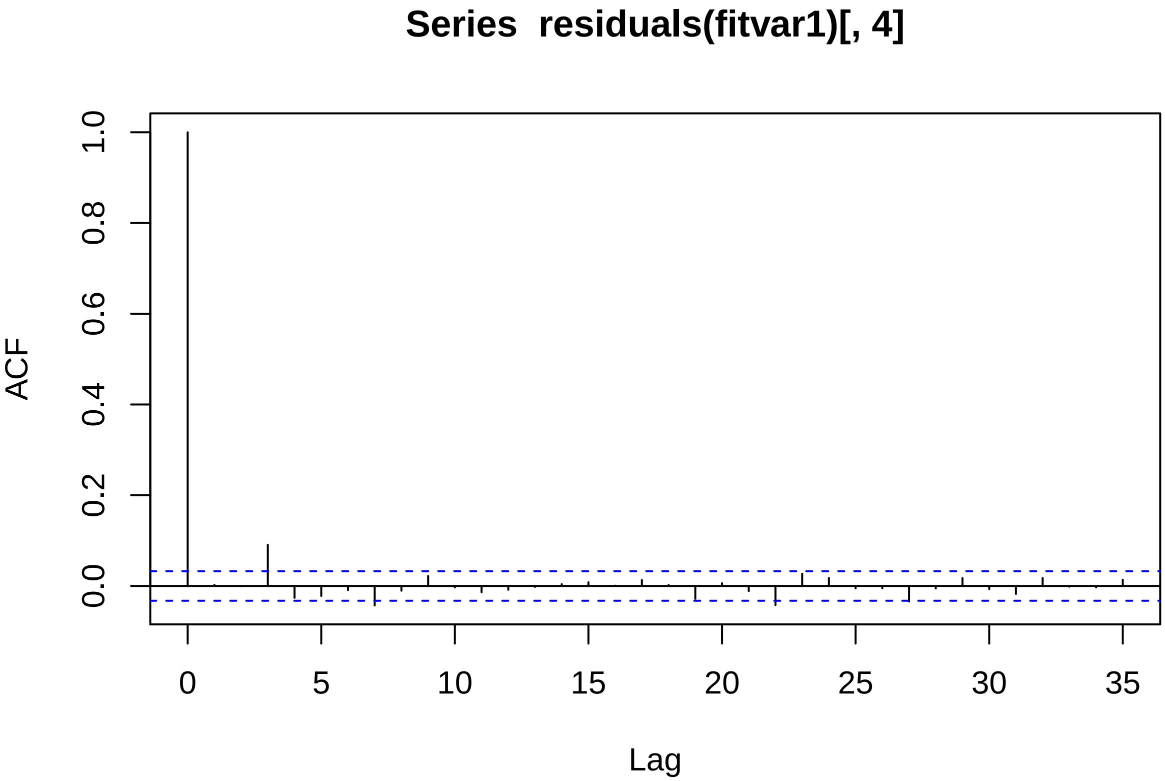


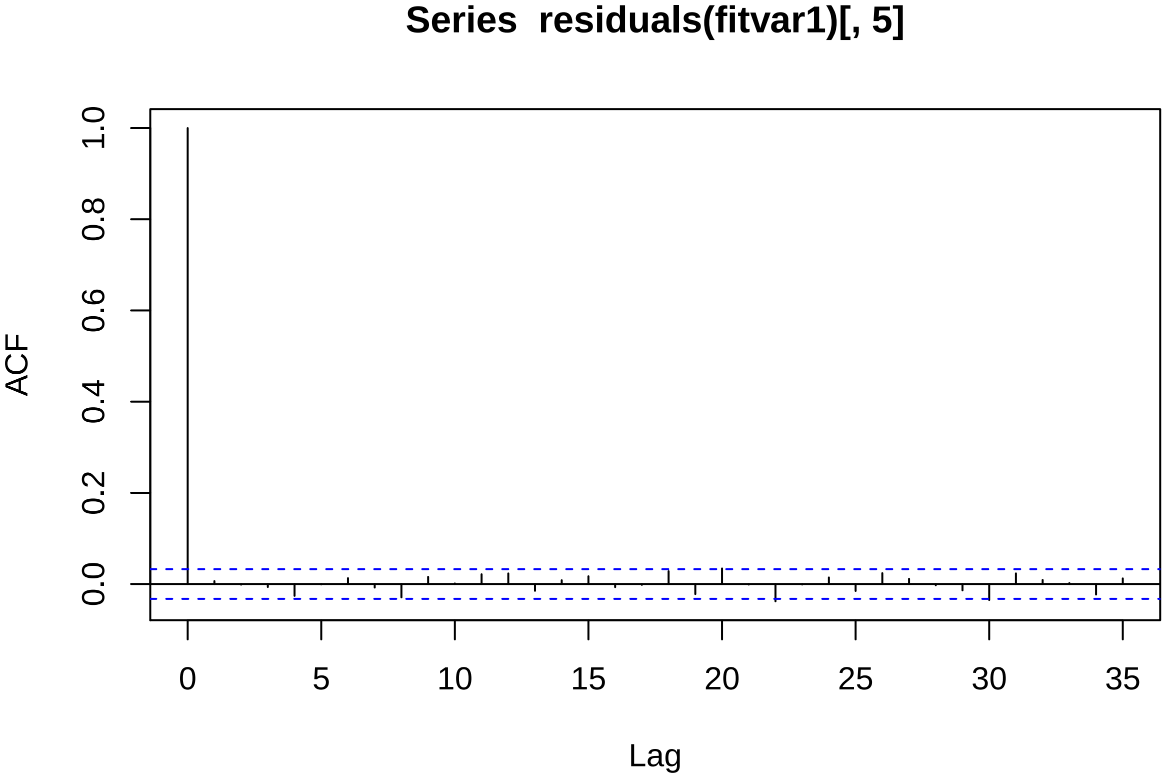


Series residuals(fitvar1)[, 2]



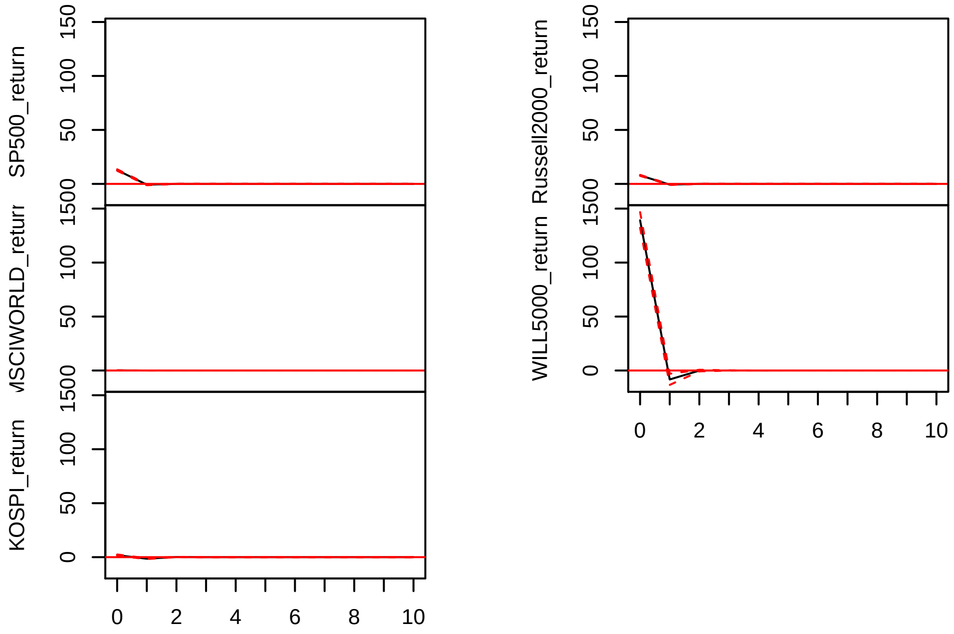
Series residuals(fitvar1)[, 3]





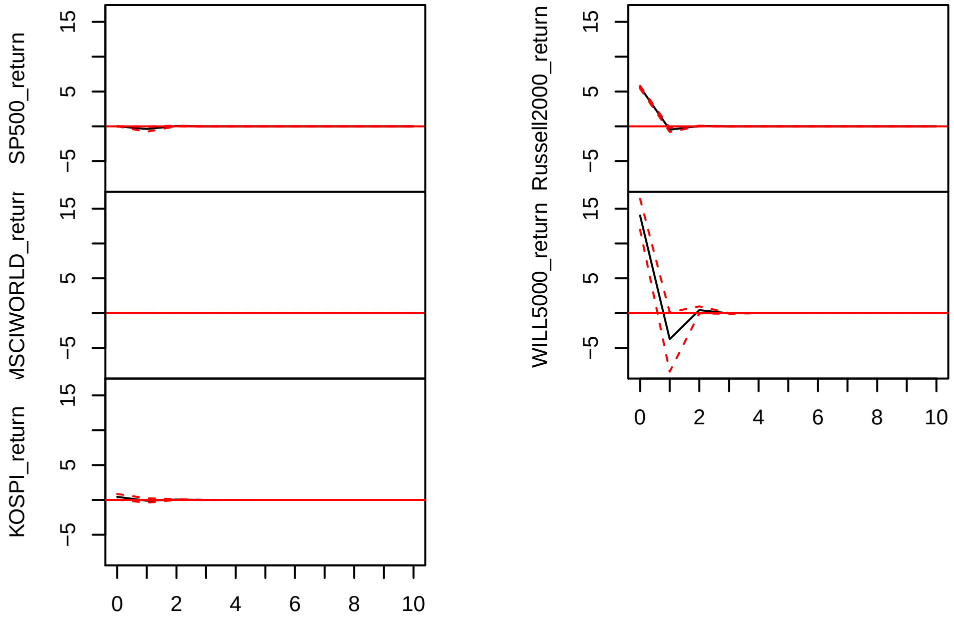
Furthermore, as we can see from the residual plots of the indices, we can see that most of the values lie in between the confidence intervals, which suggests that the errors in our VAR(1) model close to a stationary process. If we did not have stationary errors we would need to remove trend and seasonality components in our residuals further so that we could make reasonable predictions of the returns.

Orthogonal Impulse Response from SP500_return



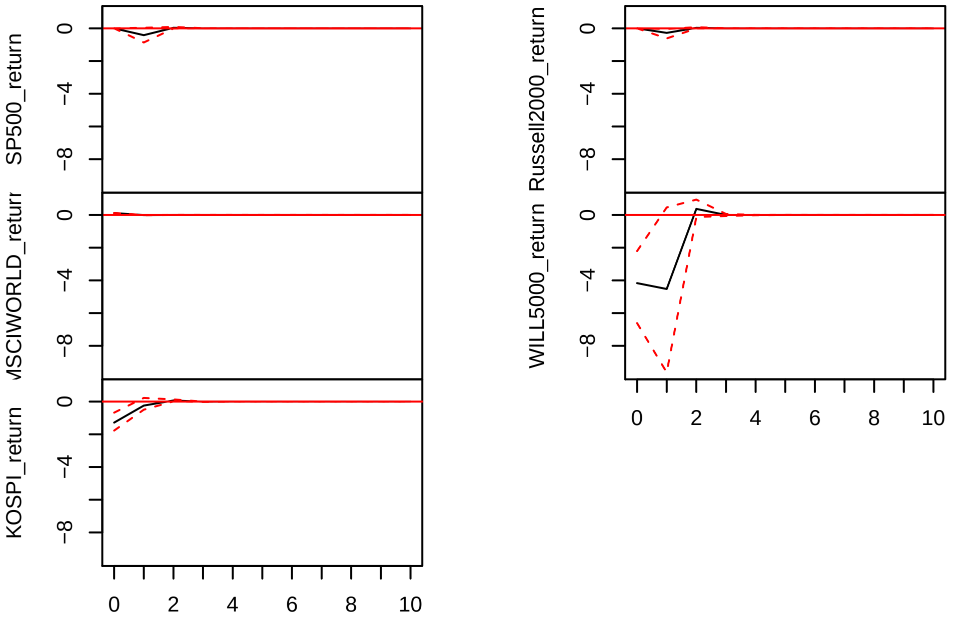
95 % Bootstrap CI, 100 runs

Orthogonal Impulse Response from Russell2000_return



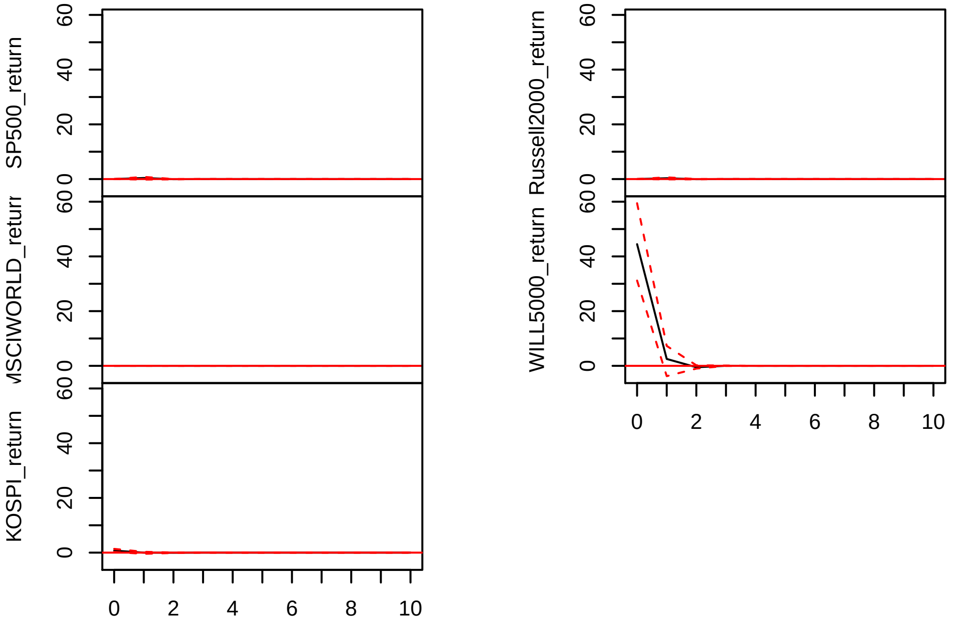
95 % Bootstrap CI, 100 runs

Orthogonal Impulse Response from MSCIWORLD_return

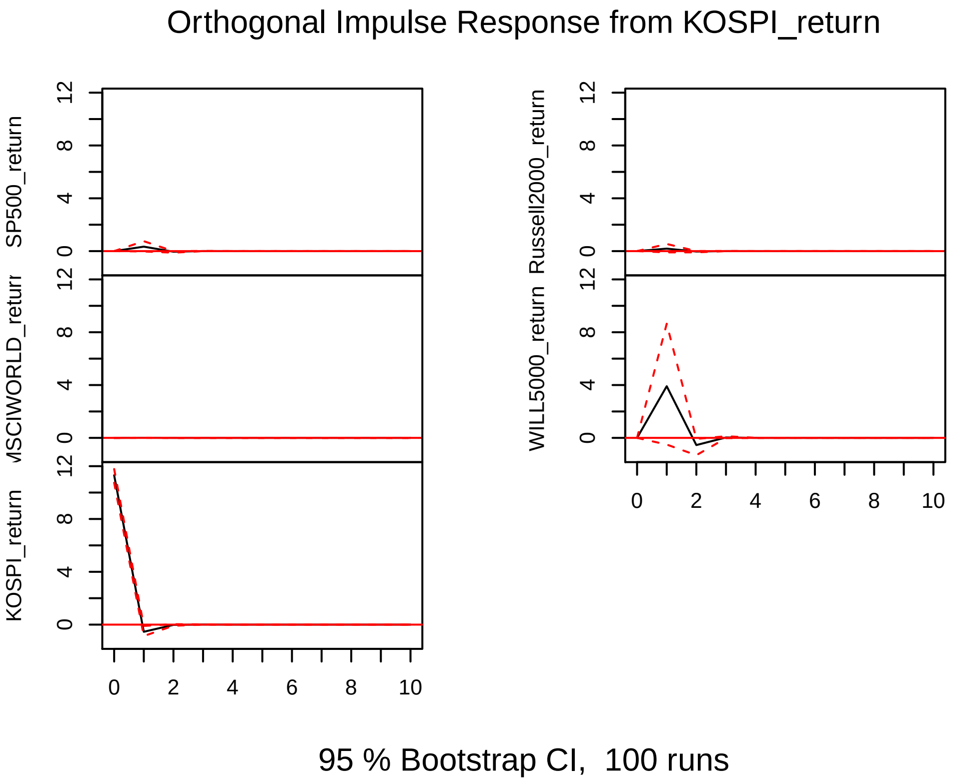


95 % Bootstrap CI, 100 runs

Orthogonal Impulse Response from WILL5000_return



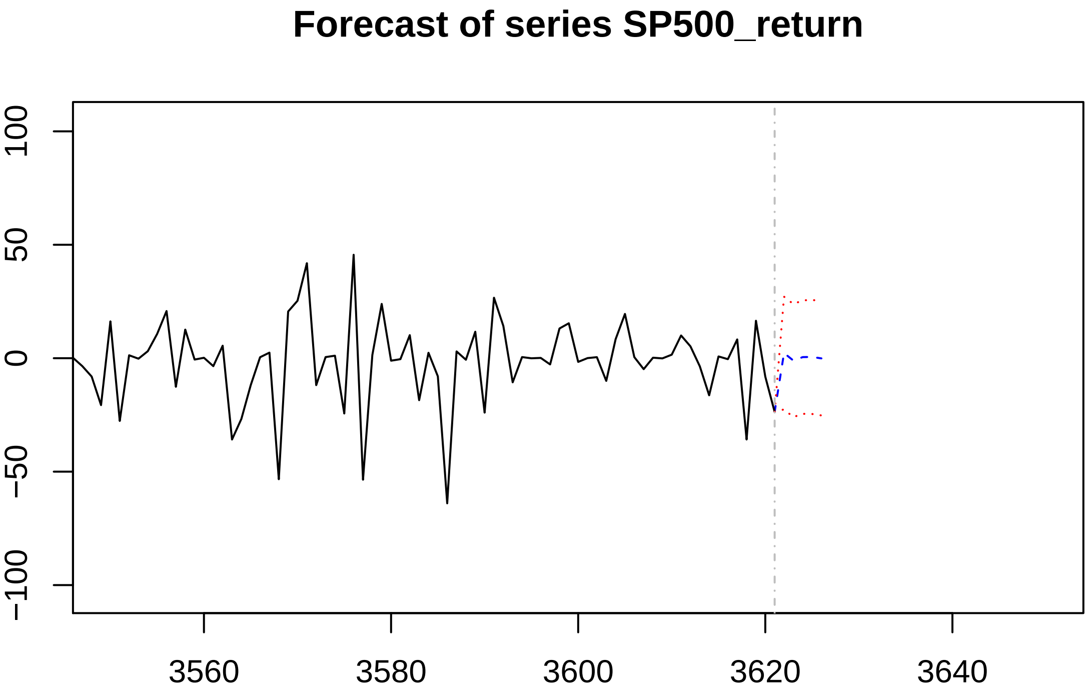
95 % Bootstrap CI, 100 runs



Impulse response functions are the reaction of the system of indices in our VAR model in response to a one-unit increase in one of the indices. In other words, these graphs give us information on how different indices would respond to a one-unit increase in one of the indices.

The first set of graphs explains the impulse response to a 1 unit increase in the SP500. Here we see that most indices do not see a response. For example, the MSCI World index and the KOSPI index return see no response from a change in the SP500. However, the Willshire 5000 index sees a hugely positive response. This might be explained by the fact that the Willshire index is a more broad index of the US stock market, an increase in the SP500 stocks has spillover effects on the remaining US stocks.

From what we can see from all the graphs, overall, the Willshire 5000 index returns are profoundly affected by all the other indices in the study. This may be because the Willshire index contains more stocks than any other index in the study, and thus market shocks around the world have an impact on some stocks in the index, which can affect returns in the Willshire 5000. Also, it looks like a positive shock in the MSCI world has a negative effect on the Willshire 5000 index, SP 500, and Russell 2000 returns as investors may move their assets from US-based stocks to other stocks around the world. This makes sense as investors might see higher possible returns outside of the US when the MSCI world index increases.



The plot above shows 5-day-predictions into the future of the S&P 500 in the model. It is not very easy to look further ahead in our model, as we would not be able to get accurate predictions. Overall, VAR models and financial models, in general, can be useful in predicting short-timed fluctuations in the stock market returns.

Limitations

When dealing with financial data, we come with a joint restriction for time series estimations. Given that markets are not always open, there is a discontinuity in the data that we have to deal with before we can start our time series analysis. This comes with significant challenges, what do we do with those days that the stock market was closed? In our project, many different techniques were used to try and solve the discontinuity, but we came with challenges. First, we tried filling the value of the SP500 using the value of the previous week. Intuitively, it makes sense to do this as the value of the SP500 remains constant when the stock market is closed. However, it comes with many issues when looking at time-series data, as we end up having a perfect correlation between the last day, the stock market was open and the next close days. This is problematic when doing our analysis because it can make our errors correlated with each other.

Given that the whole purpose of using SARIMA, ARIMA, AR, and MA models is to model random independent noise, filling the values of missing data using previous values can hinder the modeling process. To solve this problem, we used the Kalman filter to find predictions for the missing values, as explained above.

Another issue that we have with financial data is that during upturns, the variability of the values in the S&P 500 tends to be small, as investors are cautious of the future. Nevertheless, during downturns, we see large variability in the values in the S&P 500, mainly due to sentiment and desperation. The differences in variability during financial crises and growth periods give us a hard time to be able to capture both trends and seasonality in the data effectively.

Financial data is also highly volatile and unpredictable. Stock prices data is non-stationary, and there are unpredictable circumstances such as the great recession during the late 2000s that make it very hard for us to model the data appropriately. Hence we decided to use the differencing method to do our best to remove the trend and seasonality. During our model selection process to capture the trend, we used the differencing method and finally fit an MA(7), model.

One of the limitations of a VAR model is that the standard methods of statistical inference (such as computing standard errors for impulse responses) may give misleading results if some of the variables are highly persistent. However, we have decided to use the VAR model because it is one of the models that can impose a multidirectional relationship to examine the relationship between each index.

Our final predictions from the VAR model is only applicable to capture small future values because we would not be able to get accurate forecasts from as we cannot predict the actual magnitude and direction of the shocks on the indices.

Conclusion

Forecasting the health of the US Markets is very important for foreign investors and economists. It helps many stakeholders in the economy make crucial decisions regarding the changing landscapes of the economy. Our research project attempts to predict the future value of the S&P 500 using ARIMA models. The MA(7) model was used to estimate the errors of our time series. After differencing our data to adjust for non-stationarity, we were able to use these estimates to predict future values of the S&P 500.

We selected the models that best fit the residuals using the estimating approach and rigorous visualization analysis. Using our predictions from the MA(7) model, we can see that there is a slight drop in the predictions but an increasing trend overall for the next five days.

However, due to the nature of financial data, our results must be taken with a grain of salt as the S&P 500 is highly unpredictable. As a result, we expect our predictions to be highly unlikely given the limitations described above and the unpredictability of financial markets.

We also used the VAR(1) model on stock returns to analyze the multivariate relationships between the S&P 500, Russell 2000, MSCI World, Willshire 5000, and the KOSPI. Overall, from the impulse response functions, we see that the MSCI World Index and KOSPI are not affected by a shock in the S&P 500, while the Willshire 5000 index is greatly affected by the S&P 500. The Russell 2000 is affected by a small margin by a shock in the S&P 500. Our predictions from the VAR model shows that there is an increasing trend in the S&P 500 return.

Acknowledgements

We would like to thank our teacher Prof. Brianna Heggeseth and the MSCS department of Macalester College.

References

Pang and Li (2018)

Eulogio, Raul. 2018. "Performing a Time-Series Analysis on the S&P 500 Stock Index." *Oracle Data Science*. https://blogs.oracle.com/datascience/performing-a-time-series-analysis-on-the-sandp-500-stock-index.

Eun, Cheol S, and Sangdal Shim. 1989. "International Transmission of Stock Market Movements." *Journal of Financial and Quantitative Analysis* 24 (2). Cambridge University Press: 241–56.

Ionides, Edward. 2018. "Time Series Analysis for Log Returns of S&P500." https://ionides.github.io/531w18/midterm_project/project38/Midterm_proj.html.

Khashei, Mehdi, Mehdi Bijari, and Gholam Ali Raissi Ardali. 2009. "Improvement of Auto-Regressive Integrated Moving Average Models Using Fuzzy Logic and Artificial Neural Networks (Anns)." *Neurocomputing* 72 (4-6). Elsevier: 956–67.

Mukherjee, Paramita, and Suchismita Bose. 2008. "Does the Stock Market in India Move with Asia?: A Multivariate Cointegration-Vector Autoregression Approach." *Emerging Markets Finance and Trade* 44 (5). Taylor & Francis: 5–22.

Pang, Rock, and Leo Li. 2018. "A Brief Analysis of the Vector Autoregressive Model and Its Applications."

Yang, Jian, James W Kolari, and Insik Min. 2003. "Stock Market Integration and Financial Crises: The Case of Asia." *Applied Financial Economics* 13 (7). Taylor & Francis: 477–86.

Yoon, Youngohc, and George Swales. 1991. “Predicting Stock Price Performance: A Neural Network Approach.” In *Proceedings of the Twenty-Fourth Annual Hawaii International Conference on System Sciences*, 4:156–62. IEEE.