University of California, Santa Barbara

Department of Statistics and Probability

Group Alpha: Predicting S&P 500 Trends

By

Ryan Bernstein

Robert Nicholas

Simrin Kacker

March 10, 2018

Dr. Sudeep Bapat- PSTAT 174 MW

Abstract

Our assignment was to analyze a time series of our choosing, in order to better understand how we can derive insight from univariate time series that helps us predict the future with a greater degree of accuracy. We chose to analyze the performance of the S&P 500 from 2010-2017. To do this, we utilized the software R, which allowed us to plot our data, transform it into a stationary series, fit it, and forecast how the S&P 500 will perform in 2018.

Our goal was to create a model that exhibited stationarity and normality. We also needed to make sure that it had a relatively low variance. By using Box-Cox, and differencing our data, we were able to normalize the data to the best of our ability. To get a better idea of what class our model should be, we analyzed the data's ACF and PACF, ran an Auto-Arima test, and then ran a Yule-Walker test. This allowed us to focus on two options: AR (4,0), and ARIMA (2,1). After gathering diagnostics, we came to the conclusion that AR (4,0) was the optimal choice. Next we forecasted the data, and noticed that our forecast had a confidence interval with a wide range. This led us to conclude that the information we have is not sufficient enough to allow us to accurately predict the future direction of the S&P 500.

Introduction

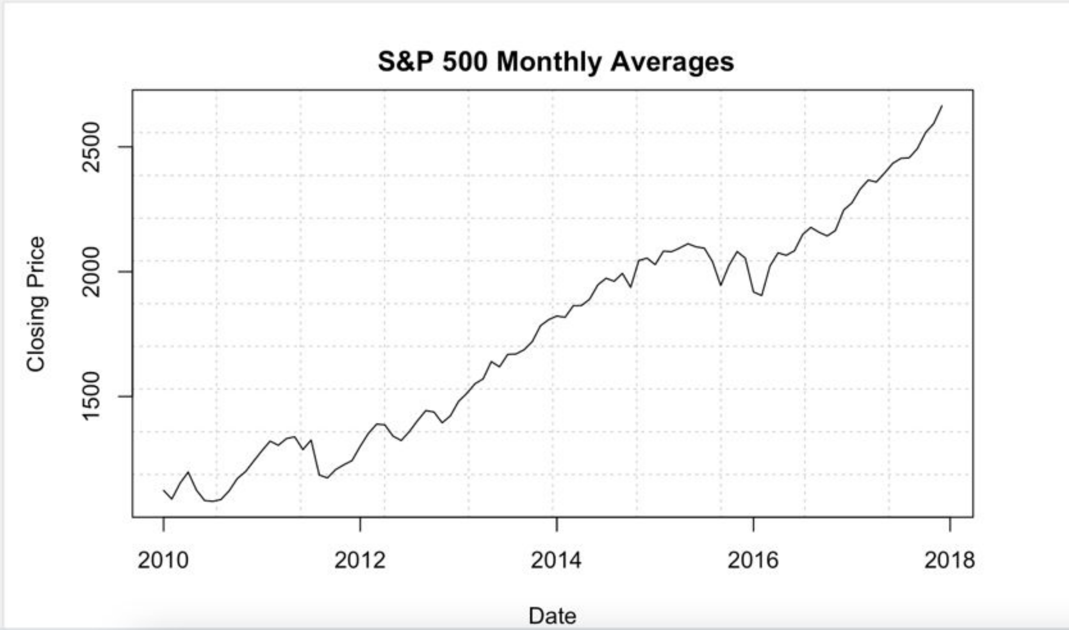
The state of the economy is always a question, a sense of unpredictability, a fear for economists everywhere. In order to better understand past trends of economic data as well as future predictions we analyzed the S&P 500 index and asked the question: *can the performance of the S&P 500 for the next year be forecasted?* The S&P 500 is a good indicator of the US market, as it indexes 505 large companies (Standard & Poor's 500 Index, 2017). Since the index captures 80 percent of the total market capital, we concluded that the S&P 500 would be a better index to predict overall market trends when compared to other indices such as the DOWJ (Amadeo, n.d.).

We obtained 8 years worth of S&P 500 data from the Wall Street Journal. The data set included opening, closing, high, and low for every day between January 1st 2010 and January 1st 2018. We believed the best price point to truly understand the market was the closing price as it represented the index after a full day of trading and chose to build a time series model based on the closing price. Looking at the daily closing prices is not good for predicting macro-economic trends, so we aggregated the data by month.

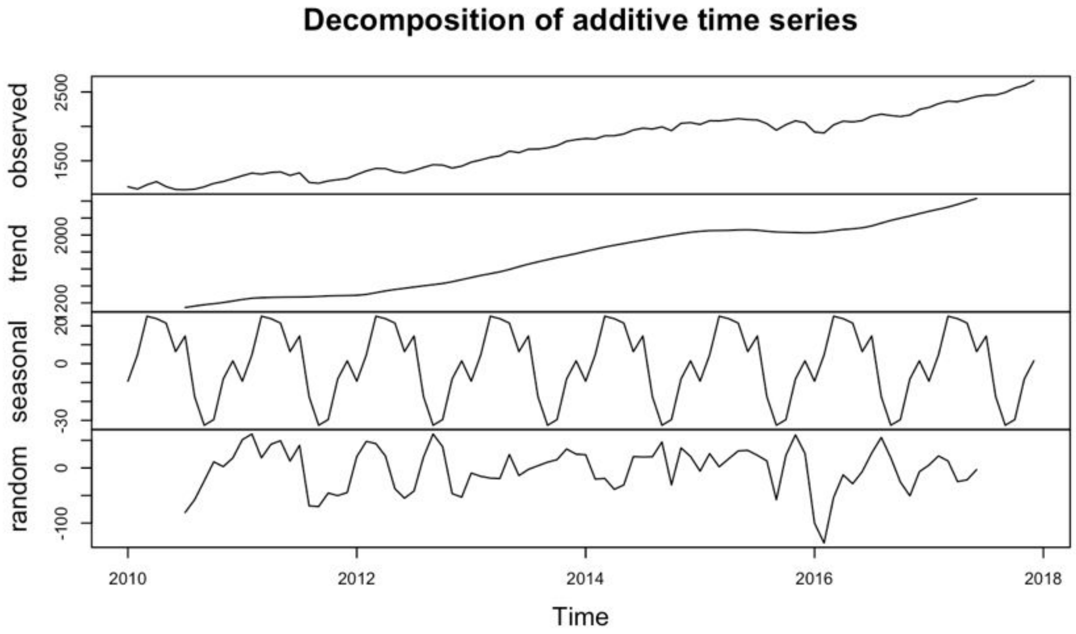
Using traditional time series techniques such as transforming the data to make it stationary, plotting the ACF and PACF, and model diagnostic checking we were able to come up with a multitude of models, but decided that an AR(4) model would best fit the data. However, this model, along with the others, gave us some issues as the residuals were not able to follow the normal distribution, as detailed later in the report. In addition the 95% confidence interval of the predictions proved to be rather large.

Plot and Analysis of Time Series

We will first plot our time series by listing the prices of the S&P 500 aggregated and averaged by month against time. We then can immediately see that there is a clear upward trend. This is of course something that is expected, most time series analysis consist of a systematic pattern as well as a random noise. If there was no pattern and only random noise then we would not be interested in forecasting the data to begin with.

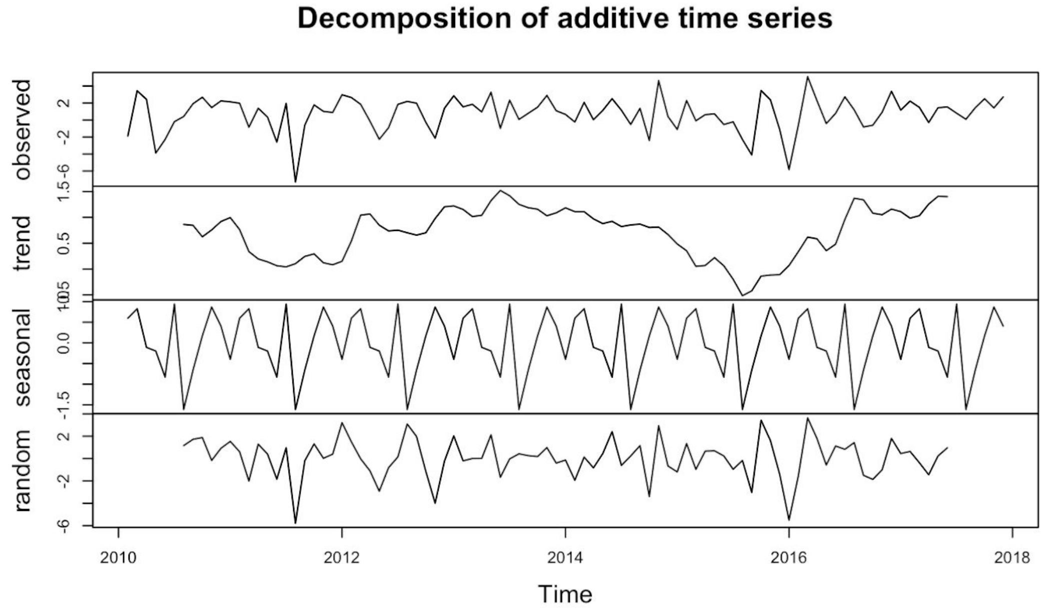


Furthermore by decomposing our time series (as seen below) we notice there is a small to near non-existent seasonal component (notice how small the values of the seasonal component are). Why is this important? Well before we can make accurate predictions of our model we must make transformations to the data in order to make sure the time series is stationary. In its current untransformed state it displays clear signs of non-stationarity, namely the fact that it is monotonously increasing and it's mean, variance, autocorrelation are not constant through time.

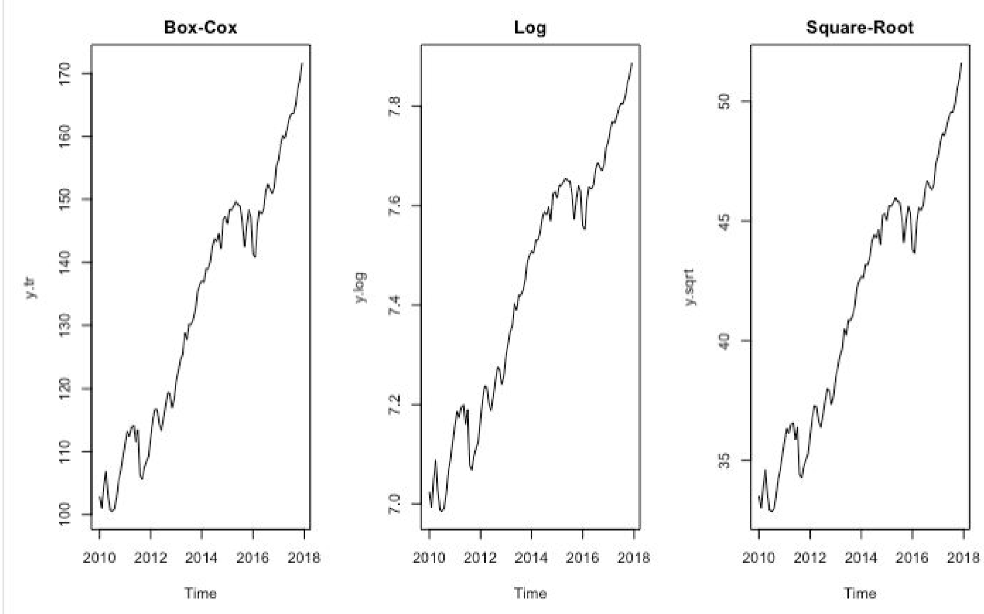


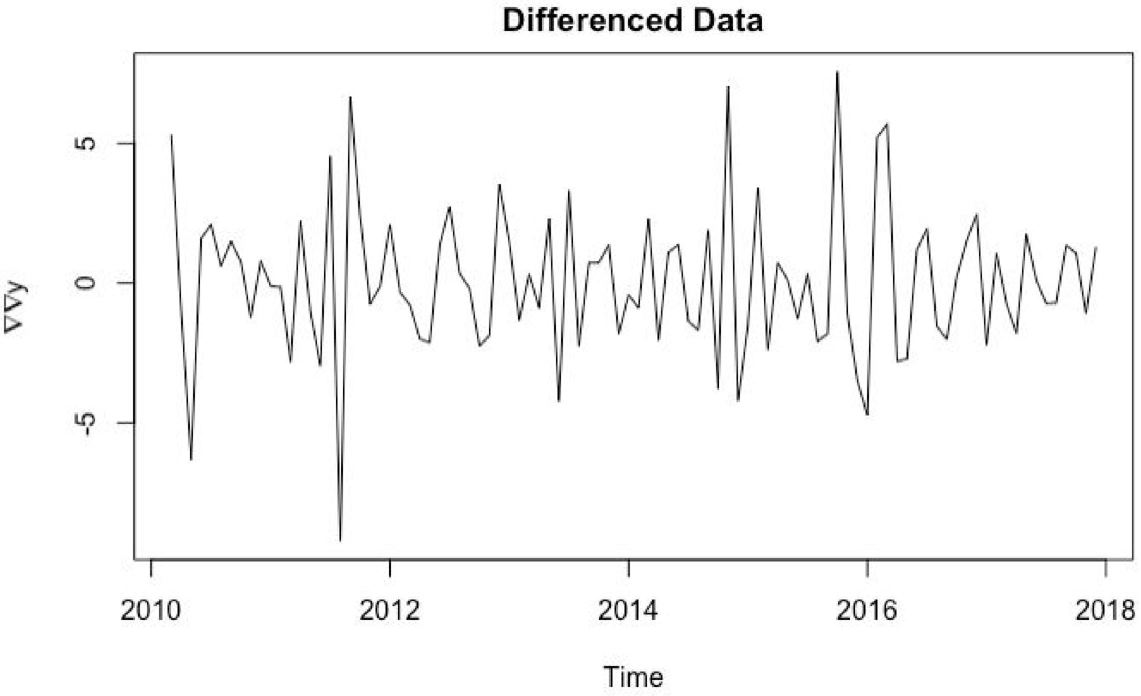
Transformation

In order to stabilize the variance of the data, we applied a square-root, log and box-cox transformation, as can be seen in figure 1. The box-cox transformation was the best fit. Upon first glance of the data it appeared to show no indication of seasonality, but to make sure, we first differenced the data by 12 to get rid of any hints of seasonality, then we differenced it again by 1 to get rid of the upward trend. We ran an Augmented Dickey Fuller test which revealed that this differenced data set was not stationary. However there was an apparent trend within the data set that would prevent us from successfully predicting prices. Therefore, we differenced the data once and checked the mean, which was not close to zero. After the first difference, it was clear we did remove a type of seasonal component. As seen in the below graph.



It is clear that this removed seasonality as the seasonality parameters went from a range of 50 to a range of 3. Since the seasonality was gone, we then differenced the data one more time to get rid of the trend. The variance increased, however the mean was now close to zero. Differencing the data twice to get rid of trend and seasonal component created a stationary data set. The final differenced data is pictured below. The twice differenced, box-cox transformed data passed the Augmented Dickey Fuller test and therefore we had a stationary data set.



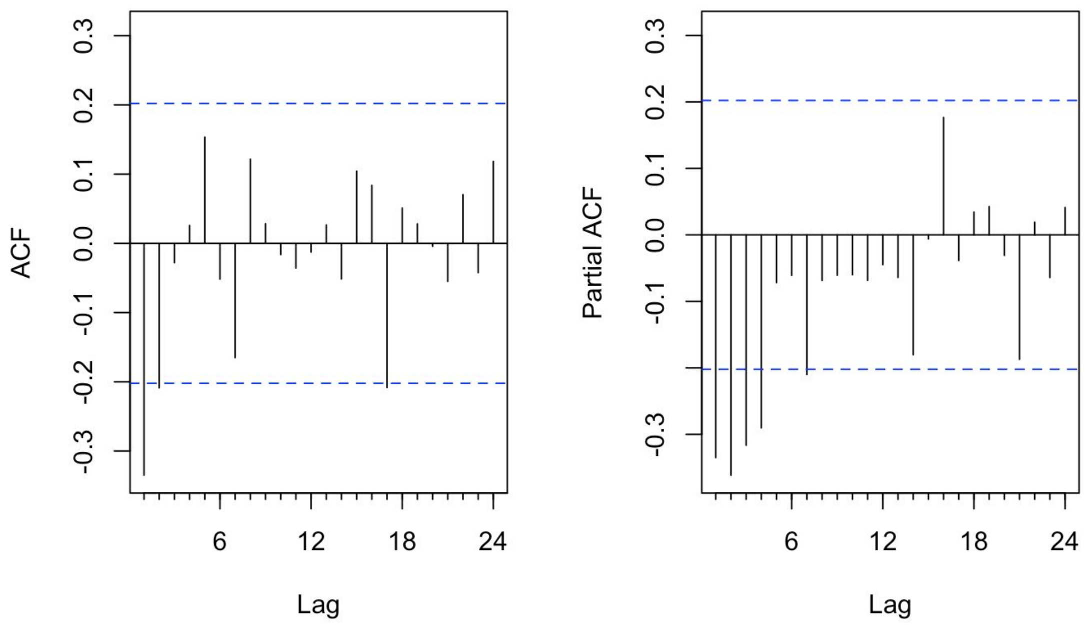


Plot and Analysis of ACF and PACF

Before the estimation can begin, we need to decide on (identify) the specific number and type of ARIMA parameters to be estimated. The major tools used in the identification phase are plots of the series, correlograms of autocorrelation (ACF), and partial autocorrelation (PACF). The decision is not straightforward and in less typical cases requires not only experience but also a good deal of experimentation with alternative models (as well as the technical parameters of ARIMA). However, a majority of empirical time series patterns can be sufficiently approximated using typical rules of thumb such as:

- an exponential decay in the ACF coupled with a spike at lag 1 in the PACF where no other lags correlate represents AR(1); the reverse of this would suggest MA(1)
- a sine-wave shape pattern of a set of exponential decays in the ACF as well as spikes up to lags k with no other lags correlating suggests AR(k); the reverse of this would suggest MA(k)
- lastly an exponential decay in both the ACF and PACF would suggest an ARMA(p,q)

Now, when we analyze the Autocorrelation and Partial-Autocorrelation functions of our transformed data we can see that there is a strong indication of an AR(4) model. There is a 'sine-wave' shape pattern on the ACF and there are 4 spikes on the PACF with almost no correlation with any other lags (aside from 7 and notable upticks at lag 14 and 16). If we ignore the significance at lag 7 then there is clearly a representation of an AR(4) model. Lastly none of the upticks are spaced exactly 12 lags apart once again reinforcing that the seasonal component is successfully removed from the transformed data.



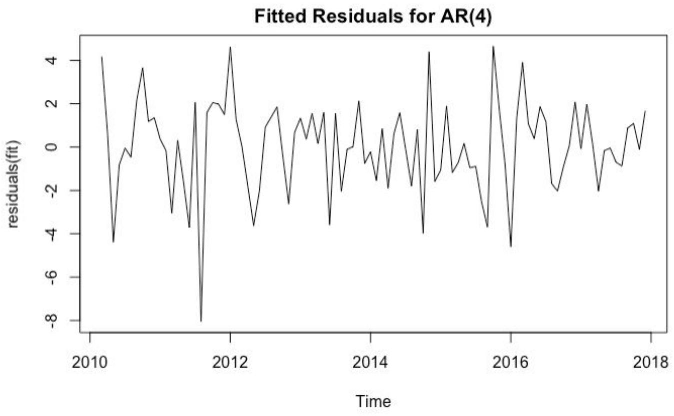
Fitting the Model

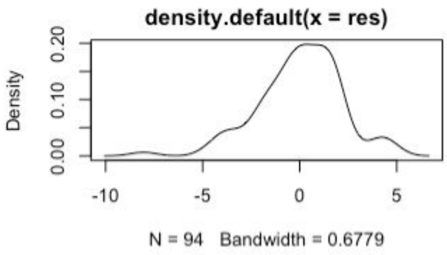
The ACF and PACF suggested an AR model, therefore we started to look at three different ways to analyze the model in RStudio - including using the autoarima function, yule-walker, and looking at different model's AIC's. After looking at all three models, we decided the best model to follow would be AR(4) since it is the simplest model that also has a low AIC. In addition all three methods indicated that an AR(4) process would be accurately. While the residuals never pass the Shapiro-Wilk test, the model appears to follow the best normal distribution of all the ones we tested. The reason the model did not pass the Shapiro-Wilk test is probably due to the fact that the actual model includes another term dependent on t, such as a +t, or +t^2, that we are unable to gleam from the data we have available since we are performing a univariate analysis. In addition, our research concludes that it is extremely difficult to have residuals pass the Shapiro-Wilk test with large data sets. The following analyzes all three models we looked at.

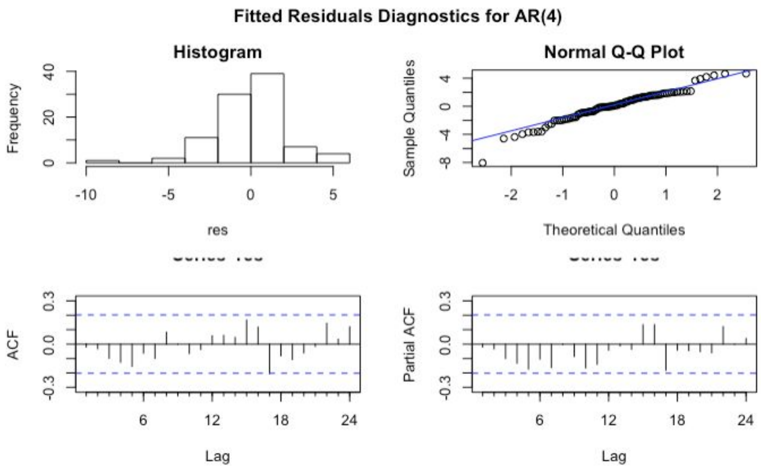
Auto.Arima Function. - First we looked at the auto arima function which suggested an ARMA(4,0). The model produced by the auto.arima function indicates:

$$X_t -+.68X_{t-1} + .67X_{t-2} + .51X_{t-3} + .3X_{t-4} = Z_t$$

We then preformed model diagnostic checks such as the Ljung Box Test, the Box-Pierce test, and the Shapiro-WIlk test. The model passed both the Ljung and Box-Pierce test, however it did not pass the Shapiro test. The small p-value produced by the Shapiro test indicates that the residuals of the fitted model do not follow the normal distribution. The following figure displays the results of the residuals.







Yule-Walker. - Since the Auto Arima function showed it was an AR process, we wanted confirm the AR(4) by running a Yule-Walker test on the data. This produced an AR(4) model, which includes less terms that our first model, since there is no indication of the MA process. Yule-Walker also produced an AR(4) model with the following coefficients:

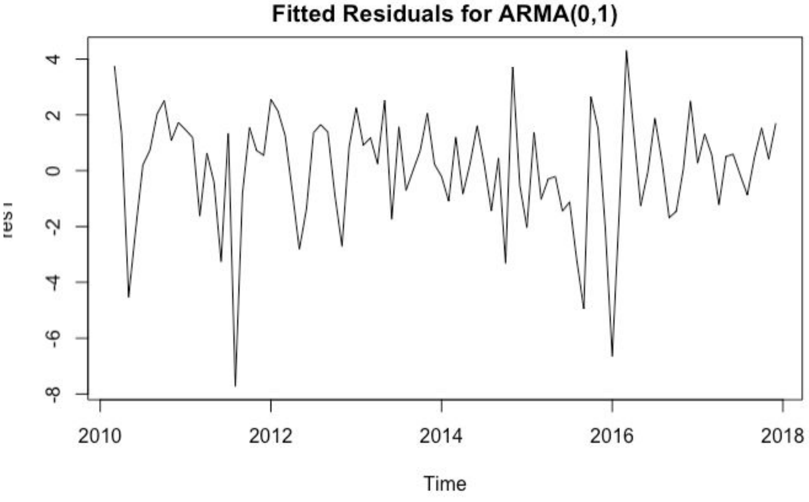
$$X_t +.66X_{t-1} + .65X_{t-2} + .48X_{t-3} +.29X_{t-4} = Z_t$$

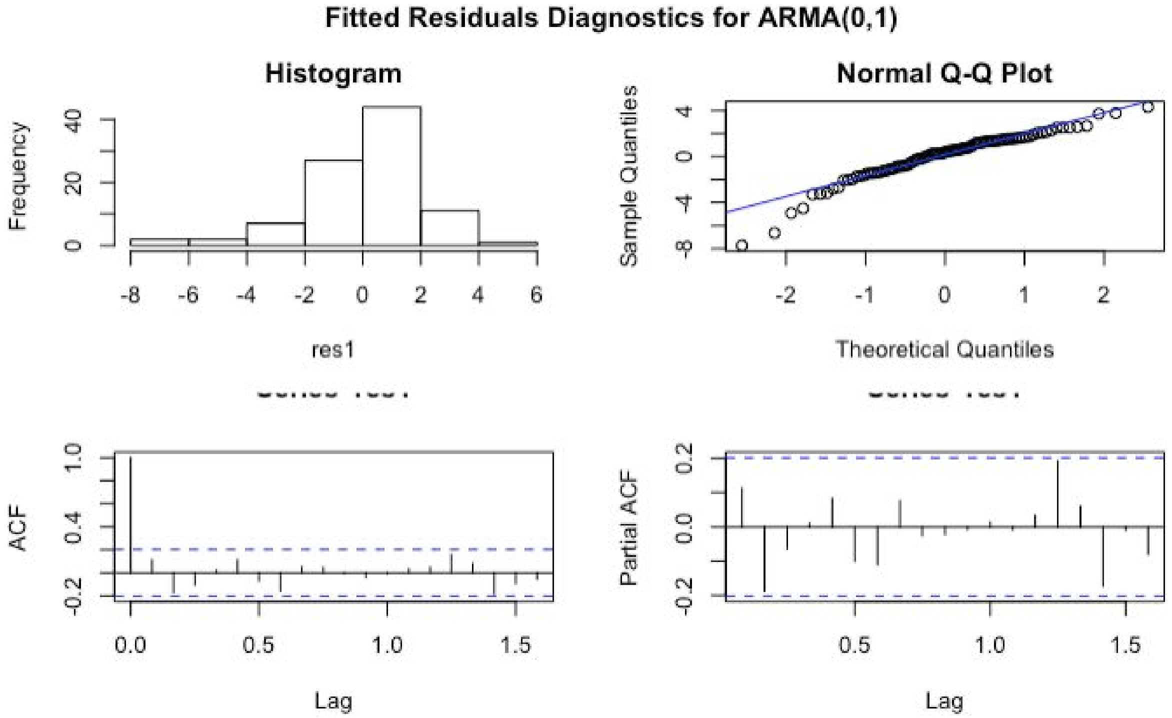
We then performed diagnostic checking on this model as well. Again, we saw that the residuals do not follow the normal distribution according to the Shapiro-Wilk test. However, it did pass the Box-Pierce test and the Box-Ljung test. While the coefficients differed with the Yule-Walker method, it did confirm that an AR(4) process would be the best to model the data.

Maximum Likelihood Estimation. - We then fit different ARMA models using maximum likelihood estimation and compared the model fits by using the AICC. The lowest AICC was 410.4227 and that was for an ARMA(0,1), the second lowest was for an AR(4) model which had an AIC of 410.9. We performed additional model diagnostic of the ARMA(0,1) model, which looks as follows:

$$X_t = Z_t \;\; + 1Z_{t-1}$$

We checked the Box-Pierce, Box-Ljung, and Shapiro-Wilk test on the new model. The residuals for the ARMA(0,1) passed both the Box-Pierce and Box-Ljung, but failed the Shapiro-Wilk test. The p-value for the Shapiro-Wilk test was extremely small and the data does not appear to follow a normal distribution at all. The following images represent the model, as we can see the ACF and PACF all lie within the 95% confidence interval. The fitted residuals seem to have a larger variance when compared to the AR(4) model.



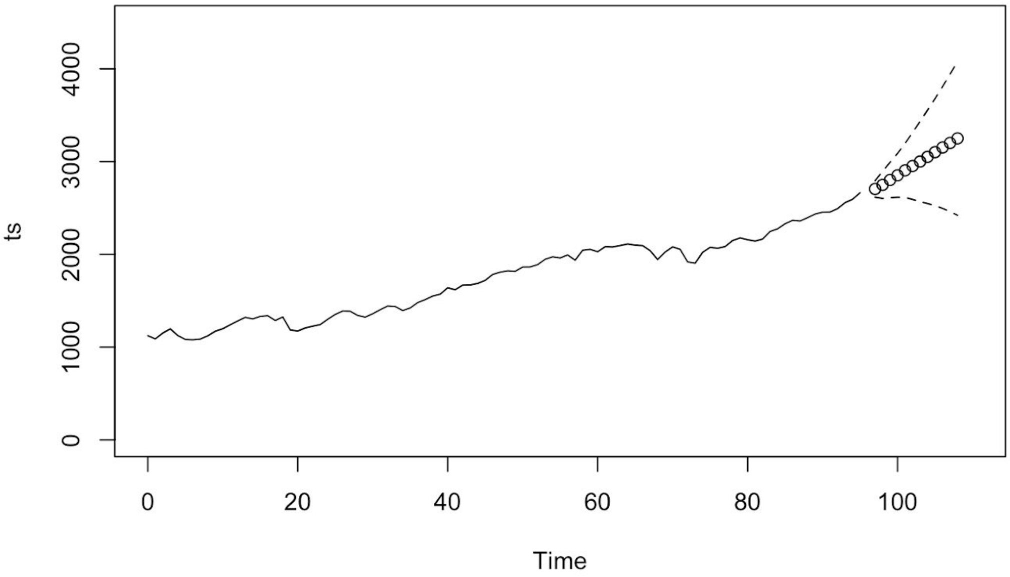

Fitted Residuals Diagnostics for ARMA(0,1)

Histogram

Normal Q-Q Plot

Forecast

Ultimately there were no grand sweeping revelations from our analysis but what is definitely apparent is that there is tremendous upside to spreading your investments across the S&P 500. Due to the unpredictability of innumerable outside forces acting upon the market you can never discount the likelihood of a market correction or small dip in the stock market. While our predictors are indicative of where the market will likely go, the dotted lines are just as important to study since they represent the 95% confidence interval of the S&P 500's forecasted performances over the next 12 months (i.e. while we cannot safely bet that our predictors will be spot on, we can safely bet that the S&P performance will lie within our two confidence bounds).



Conclusion

Our analysis of this time series allowed us to confirm many of the suspicions that we had about the nature of the stock market. It is of the utmost importance that the S&P 500 is not easy to predict because if it were, every statistician in the world would quit their jobs and become full-time investors. So it was not surprising when our forecast could at the very best give us a picture of the best likely scenario and worst likely scenario; we can assume that the actual performance of the S&P 500 will perform somewhere between those extremes but any more precise predictions then the ones we formed would clearly require multivariate analysis. . Additionally, we expected the data to be void of seasonality. If the stock market were to do better than average every January, for example, people would buy stocks on January 1st, and sell stocks on January 31st.

However, we did expect to be able to create a model where the residuals follow a normal distribution. When we were unable to do this, we realized that the model must be dependent on time in a way that we could not foresee prior to undertaking this project.

Professor Bapat was an instrumental resource for us whenever we had any questions about how to proceed, and his lecture slides proved to be very useful. It was a unique pleasure to learn from him. We also wanted to thank TA's Patricia Ning and Zhipu Zhou for guiding us through our laboratory sections, which helped us become more familiar with R.

Our final model:

$$X_t -+.68X_{t-1} + .67X_{t-2} + .51X_{t-3} + .3X_{t-4} = Z_t$$

Sources

Amadeo, Kimberly. (n.d). The S&P 500 and How It Works. Retrieved from https://www.thebalance.com/what-is-the-sandp-500-3305888

Standard & Poor's 500 Index - S&P 500. (2017, November 15). Retrieved from https://www.investopedia.com/terms/s/sp500.asp

S&P 500 Index. (n.d). Retrieved from http://quotes.wsj.com/index/SPX/historical-prices.