Time Series Methods In Forecasting NYSE Return

Abstract

Through time series analysis on the New York Stock Exchange (NYSE) data from astsa package, we predicted the daily returns of the NYSE . Three models were proposed under the investigation of ACF and PACF plots. ARIMA(0, 0, 5) is the model left to be used for forecast after filtering through diagnostic plots. The model was able to predict statistically significant results of 5 days' return, and the spectral analysis established no significance for the dominant peaks. Overall, more data would be beneficial for more predictions and helpful to build different models that suit multiple needs.

Keywords: ARIMA, time series, NYSE

Introduction

Many stock investors rely on technical analysis to help them make decisions, and a lot of the techniques originate from time series analysis such as autoregressive integrated moving average models (ARIMA). Unlike other forms of analysis that require multidimensional data, this technique only relies on one attribute of the data. In this report, our goal is to model the return of the NYSE through statistical methods and propose an ARIMA or SARIMA that can be used to forecast the returns in the future. The data that the report will be analyzing comes from the astsa package (https://cran.r-project.org/web/packages/astsa/). It collects the returns of the NYSE from February 2nd, 1984 to December 31, 1991; the time span is equivalent to 2000 trading days. The day that the stock market opens is called a trading day; there are 253 days that NYSE is open to trade every year. The data is recorded in percentage of return per day as a time series. Overall, we expect that NYSE returns are predictable using an ARIMA model.

Methods

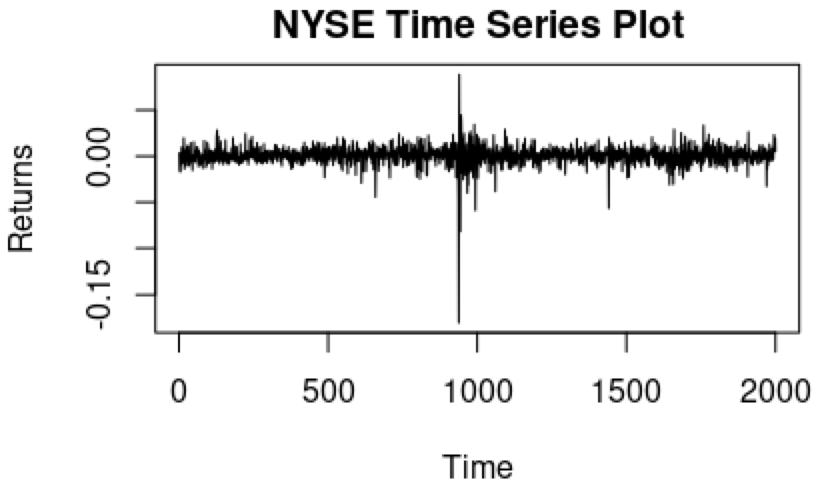


Figure 1

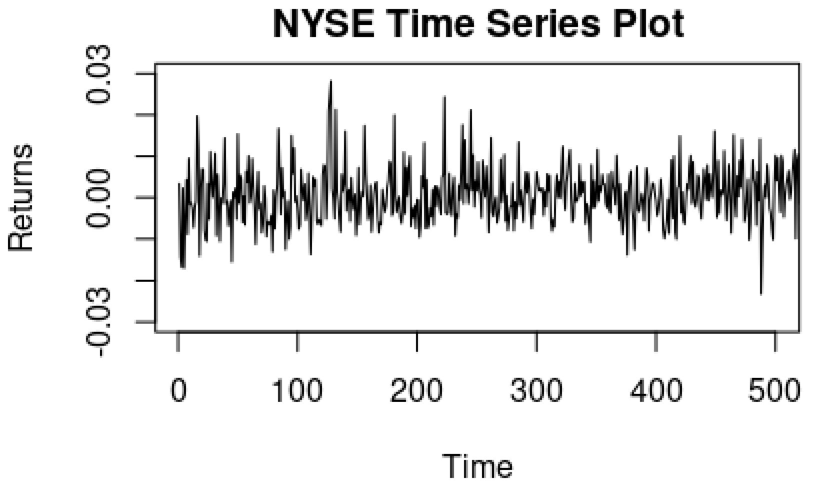


Figure 2

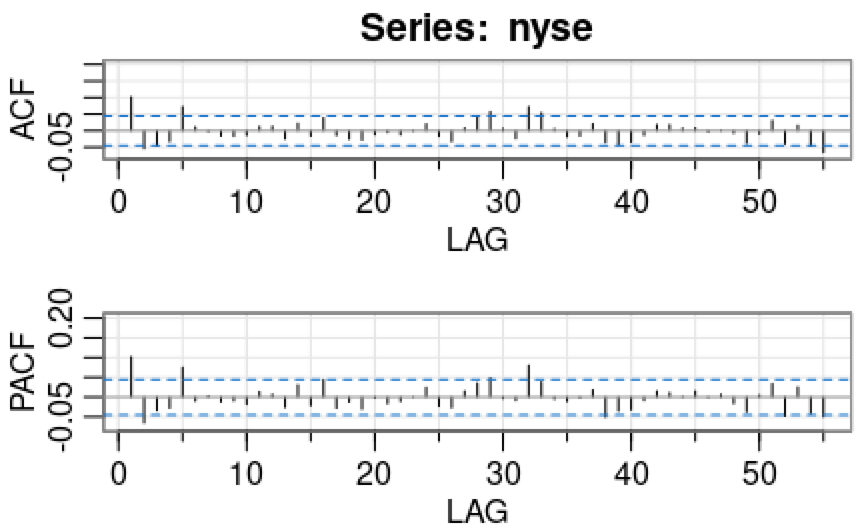
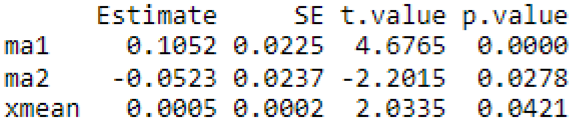


Figure 3

Figure 1 shows a time series plot of the astsa's NYSE return data versus time in days since February 2nd, 1984. According to Figure 1, the mean of the process is centered at zero. Figure 2 zooms in on day 1 to day 500 to further examine the data, and the plot shows that NYSE does not have a seasonality at a closer look. Also, no patterns are found in both figures except the obvious outliers at around day 900, thus we conclude that the process is stationary. The ACF plot from figure 3 further suggests that differencing is not needed as the autocorrelation function decays fast enough. Moreover, according to figure 3, both ACF and PACF indicate cutting offs at lag 2 and lag 5. Hence, the possible models are ARIMA(0, 0, 2), ARIMA(2, 0, 0), and ARIMA(0, 0, 5).

Results

Table 1: Parameter Estimations for ARIMA(0, 0, 2)

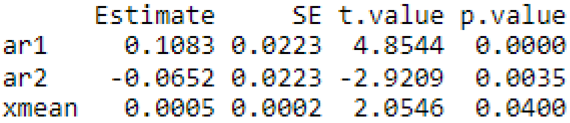


```
      Estimate     SE t.value p.value
ma1     0.1052 0.0225  4.6765  0.0000
ma2    -0.0523 0.0237 -2.2015  0.0278
xmean   0.0005 0.0002  2.0335  0.0421
```

ARIMA(0, 0, 2): $x_t = \mu_0 + w_t + \theta_1 \cdot w_{t-1} + \theta_2 \cdot w_{t-2}$

Table 1 contains the parameter estimations of ARIMA(0, 0, 2). The moving average coefficients, MA1 and MA2, of ARIMA(0, 0, 2) indicate the return of NYSE would change by 10.52% of the error term one day ago, and -5.23% of the error term 2 days ago, depending on the error on that day. Xmean is the constant of this model, which means there is always a 0.05% return at any time. According to table 1, MA1, MA2 and the constant are all statistically significant because their p-values are below significance level 0.05.

Table 2: Parameter Estimations for ARIMA(2, 0, 0)

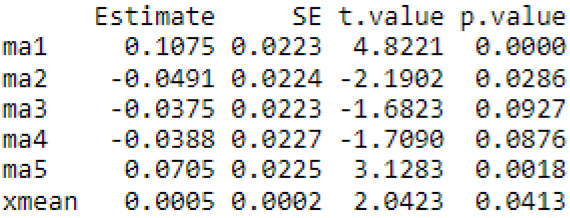


```
      Estimate     SE t.value p.value
ar1     0.1083 0.0223  4.8544  0.0000
ar2    -0.0652 0.0223 -2.9209  0.0035
xmean   0.0005 0.0002  2.0546  0.0400
```

ARIMA(2, 0, 0): $x_t = \mu_0 + \Phi_1 \cdot x_{t-1} + \Phi_2 \cdot x_{t-2}$

Table 2 contains the parameter estimations of ARIMA(2, 0, 0). The autoregression coefficients, AR1 and AR2 , of ARIMA(0, 0, 2) indicate the return of NYSE would fluctuate by 10.83% of the return one day ago, and -6.52% of the return two days ago, depending on the return of that day. Also, there is a constant 0.05% return at any time. According to table 1, AR1, AR2 and constant are all statistically significant because their p-values are below significance level 0.05.

Table 3: Parameter Estimations for ARIMA(0, 0, 5)



```
      Estimate     SE t.value p.value
ma1     0.1075 0.0223  4.8221  0.0000
ma2    -0.0491 0.0224 -2.1902  0.0286
ma3    -0.0375 0.0223 -1.6823  0.0927
ma4    -0.0388 0.0227 -1.7090  0.0876
ma5     0.0705 0.0225  3.1283  0.0018
xmean   0.0005 0.0002  2.0423  0.0413
```

ARIMA(0, 0, 5): $x_t = \mu_0 + w_t + \theta_1 \cdot w_{t-1} + \theta_2 \cdot w_{t-2} + \theta_5 \cdot w_{t-5}$

Table 3 contains the parameter estimations of ARIMA(0, 0, 5). According to the table, only MA1, MA2, MA5 and constant are statistically significant because their p-values are below significance level 0.05. The moving average coefficients, MA1, MA2 and MA5 indicate the return of NYSE would change by 10.75% the error term a day ago, -4.91% of the error term two days ago, 7.05% of the error term three days ago. Also, there is a constant 0.05% return at any time.

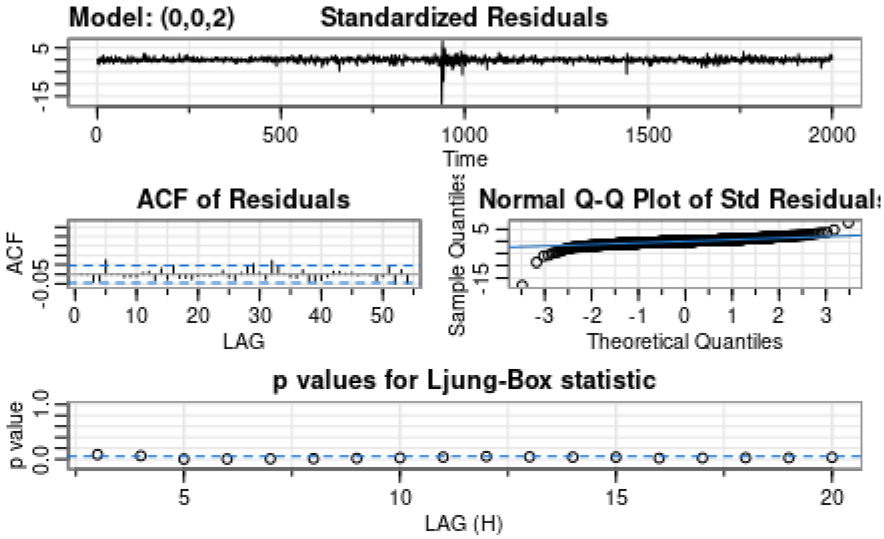


Figure 4

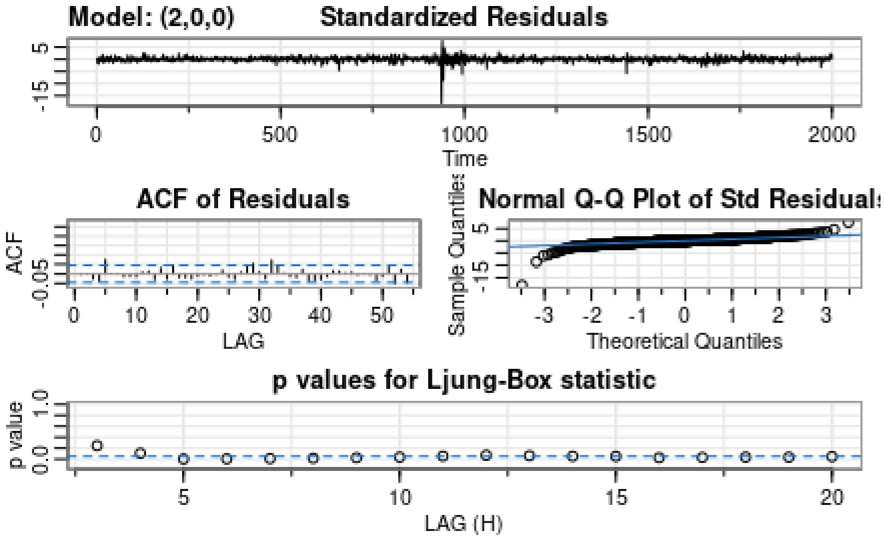


Figure 5

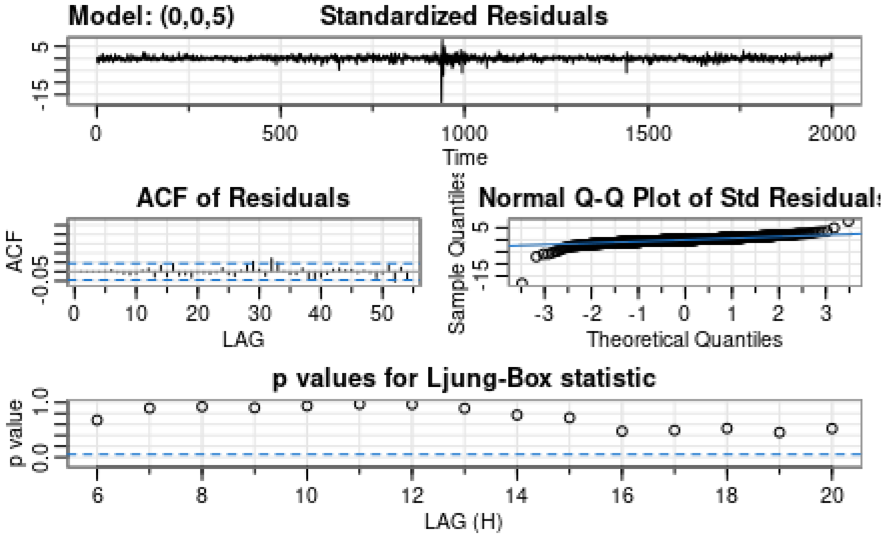


Figure 6

By inspecting the proposed models' diagnostic plots. The standardized residuals of all three models do not contain any visible patterns. Outliers are rare for all three models. Although spike at lag 32 is common for all three models, spike at lag 5 is only obvious in the ACF residuals of ARIMA(0, 0, 2) and ARIMA(2, 0, 0). Furthermore, the assumption of normality is evident in the residual's normal Q-Q plots of all three models. It is noteworthy that they all have common outliers on the tails of the Q-Q plots, and some extreme outliers are very far from normality. This is most likely the outlier that is mentioned in the beginning of the report. Last of all, both ARIMA(0, 0, 2) and ARIMA(2, 0, 0) reject the null hypothesis that residuals are independent, because most of their p-values for Ljung-Box are below or on the edge of the significant level. On the other hand, the Ljung-Box graph of ARIMA(0, 0, 5) demonstrates reasonable p-values that do not reject the independence of the residuals. Ultimately, ARIMA(0, 0 ,5) is the only model that satisfies all assumptions: mean zero, constant variance, and normally independently identically distributed residuals. Therefore, ARIMA(0, 0, 5) is the best choice for the forecast.

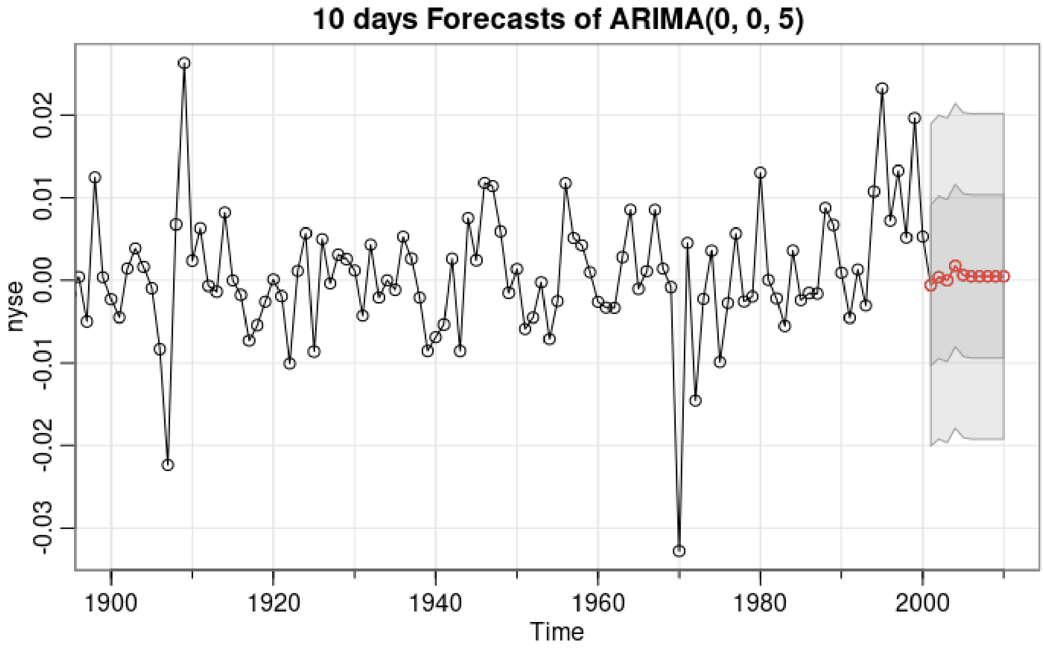
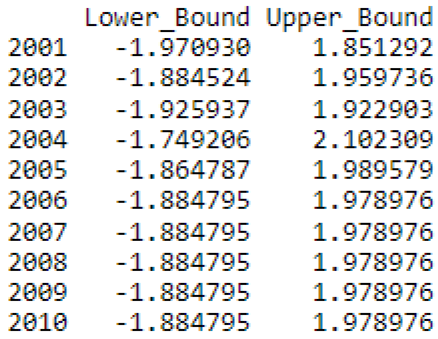


Figure 7

Table 4: 95% Confidence Intervals of 10 Days Prediction



```
     Lower_Bound Upper_Bound
2001   -1.970930    1.851292
2002   -1.884524    1.959736
2003   -1.925937    1.922903
2004   -1.749206    2.102309
2005   -1.864787    1.989579
2006   -1.884795    1.978976
2007   -1.884795    1.978976
2008   -1.884795    1.978976
2009   -1.884795    1.978976
2010   -1.884795    1.978976
```

Figure 7 illustrates the prediction of NYSE returns of the next 10 days. The grey area in the graph is the confidence interval of the forecasts, and the red points indicate the median of the prediction intervals. For example, table 4 indicates the model is 95% confident that the return of NYSE is approximately between -1.97% and 1.85% when time(in days) is 2001. However, the intervals from day 2005 to day 2010 remain the same. This is because when time exceeds 2006, ARIMA(0, 0, 5) fully relies on the error terms calculated from our own model instead of the values from the original data. Therefore, only predictions from day 2001 to 2005 are useful.

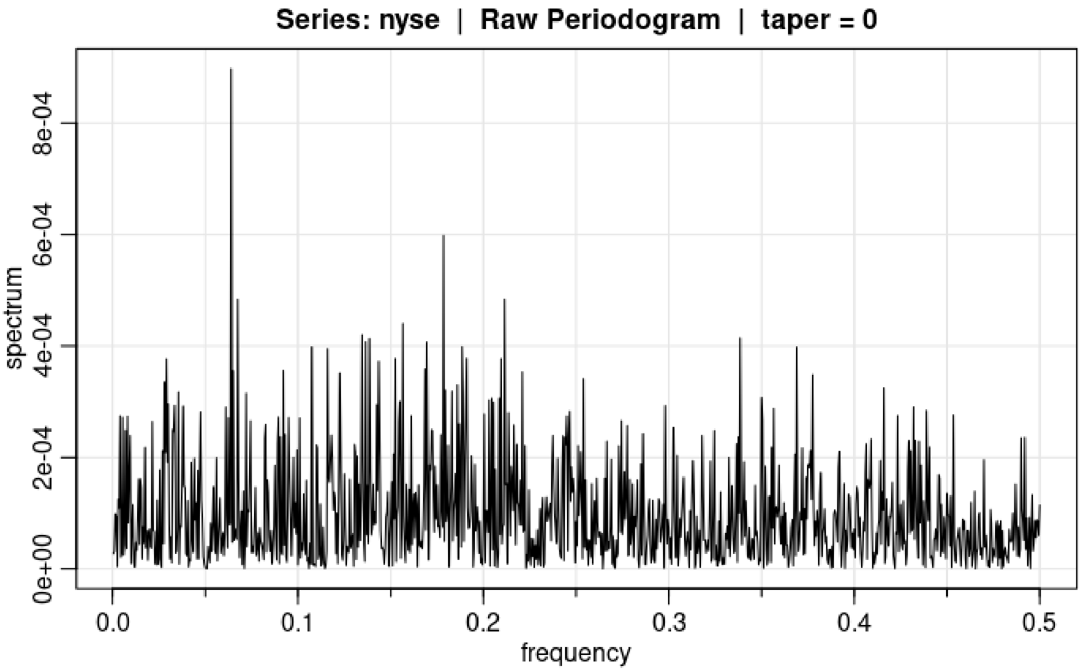
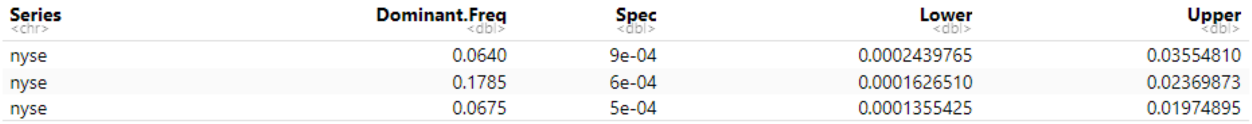


Figure 8

Table 4: Spectral Analysis 95% Confidence Intervals



Series	Dominant.Freq	Spec	Lower	Upper
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
nyse	0.0640	9e-04	0.0002439765	0.03554810
nyse	0.1785	6e-04	0.0001626510	0.02369873
nyse	0.0675	5e-04	0.0001355425	0.01974895

We found the first three dominant frequencies are 0.0640, 0.1785, 0.0675, and the cycles are occuring at 15.625, 5.602241, 14.81481. According to the 95% confidence intervals in table 4, the first peak is not statistically significant because the periodogram ordinate is 0.0009, which is bounded by the intervals of the second and third peak. The second peak is not statistically significant because the periodogram ordinate is 0.0006, which is bounded by the intervals of the second and third peak. Last but not least, the third peak is not statistically significant because the periodogram ordinate is 0.0005, which is bounded by the intervals of the second and third peak.

Discussion

In conclusion, the analysis on the NYSE data of the astsa package produced an ARIMA(0, 0, 5) model that allows us to give a reasonable forecast for the five following days. For instance, we estimated the return would be around -0.006% (prediction interval: -1.97%, 1.85%) on the first trading day of 1992. This model enables NYSE investors to plan their trade ahead, and offers them a good chance to earn equities for their portfolio. Nonetheless, this model may perform poorly in the long run, since the model recursive employs the previous predictions to model the new return when lag increases. As a result of this, forecasts will become insignificant and redundant, which is manifested in figure 7; prediction intervals plateaued after day 2005. For traders who seek long term investment strategies, a larger data with seasonality and different period options such as weeks, months and years would be helpful. Otherwise, this model is still an adequate option for day traders who can update their data everyday.