STA457 Final Project

Tian Yi Zhang

December 16 2021

1 Abstract

The unpredictability of the stock market means that no investment is ever guaranteed to be profitable. This, however, should not discourage one to seek out resources that could help them make more educated investment decisions. Hence, This report is dedicated to the analysis of trends and seasonal components in the New York stock exchange returns and arrive at a model that lets us predict future returns.

This report performs various diagnostics and arrives at a SARIMA$(2, 0, 2)$x$(0, 1, 1)_{12}$ model with parameter estimates of $\hat{\phi}_1 = 0.5029$, $\hat{\phi}_2 = -0.5239$, $\hat{\theta}_1 = -0.4049$, $\hat{\theta}_2 = 0.4364$ to forecast the future returns and conducts spectral analysis to identify three predominant periods of 0.064, 0.1785 and 0.0675 in the dataset. The parameter estimates show that stock returns that are 1 trading day apart are positively correlated and stock returns that are 2 trading days apart are negatively correlated. Major oscillations in daily exchange returns happens every 15.625 trading days, 5.602241 trading days, and every 14.81481 trading days.

Keywords: SARIMA, ACF, PACF, Seasonal, Black Monday, Trading Day, Returns, nyse, periodogram, statistically significant.

2 Introduction

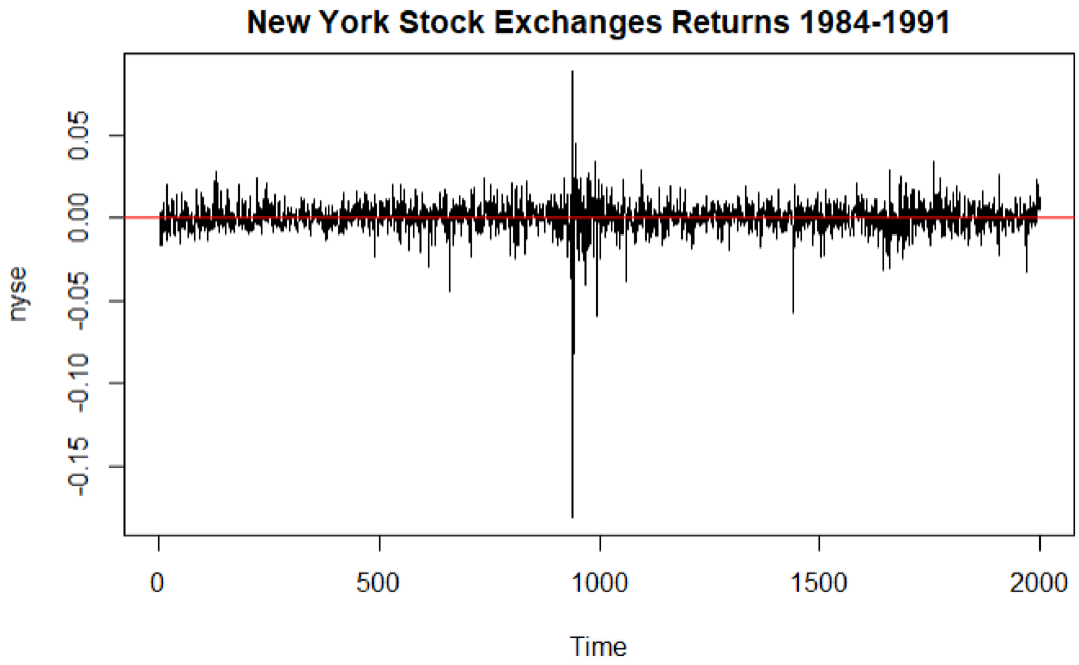
Internet sources commonly state that the historical average stock market return is about 8-12%. While this average is promising, there were [1] only six times where the returns were within this range between 1926 and 2014. Most of the times, the returns are either much lower or much higher; there is a high level of volatility at play in the stock market.

Although the above may lead one to believe that are no guarantees in the stock market, there are still ways to put the odds in favor of the investors by uncovering possible underlying periodical and seasonal components. This report aims to fulfill that goal by conducting an analysis on the *nyse* dataset and finding underlying trends, forecasting future values, and presenting the findings in hopes of helping future investors make more educated and beneficial decisions on the stock market.

This report analyzes the *nyse* dataset found in the [2] *astsa* package, which contains the daily returns of the 2000 trading days of the New York Stock Exchange from February 2, 1984 to December 31, 1991. For context, there are 253 tradings days per year, excluding weekends, holidays, etc. It's important to note that this dataset includes the stock market crash in October 19, 1987, which will appear as an outlier in the dataset.

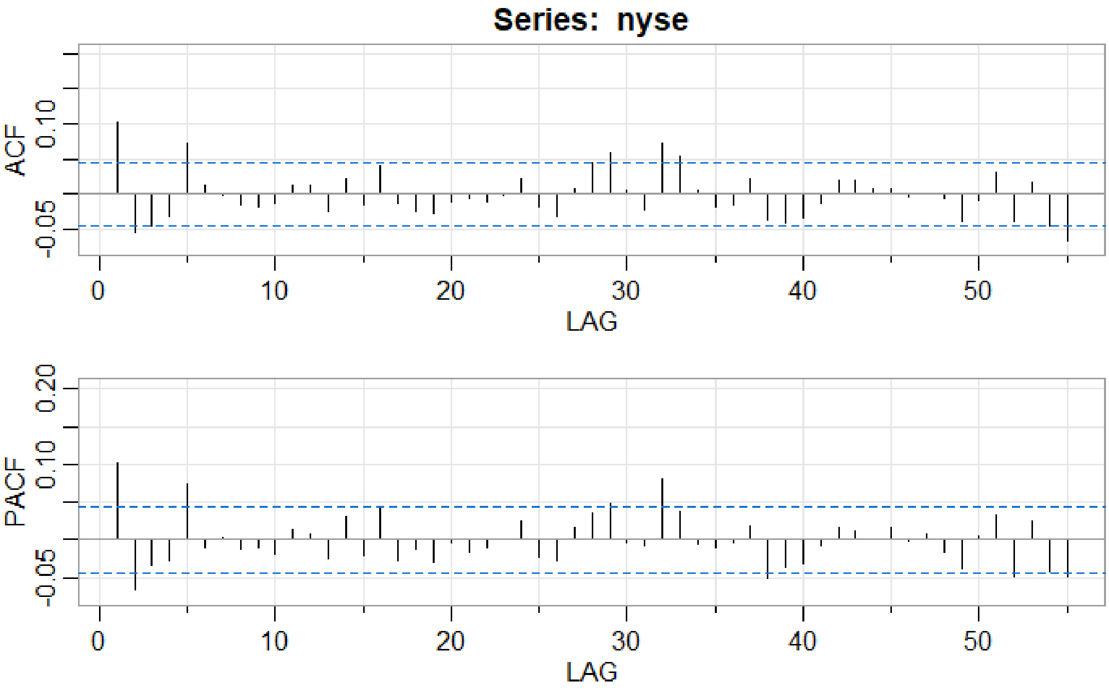
The Preliminary Analysis section will explore the time series plot and identify the dependence orders of potential ARIMA/SARIMA models. Diagnostics will be performed on the proposed models and a final model will be chosen in Model Selection and Diagnostics section. The report then forecasts the next ten time periods and presents their respective 95% confidence intervals in the Forecasting section. The Spectral Analysis section identifies first three predominant frequencies in the dataset. Finally, the report concludes the findings and comments on them in Conclusion and Limitations Sections.

3 Preliminary Analysis



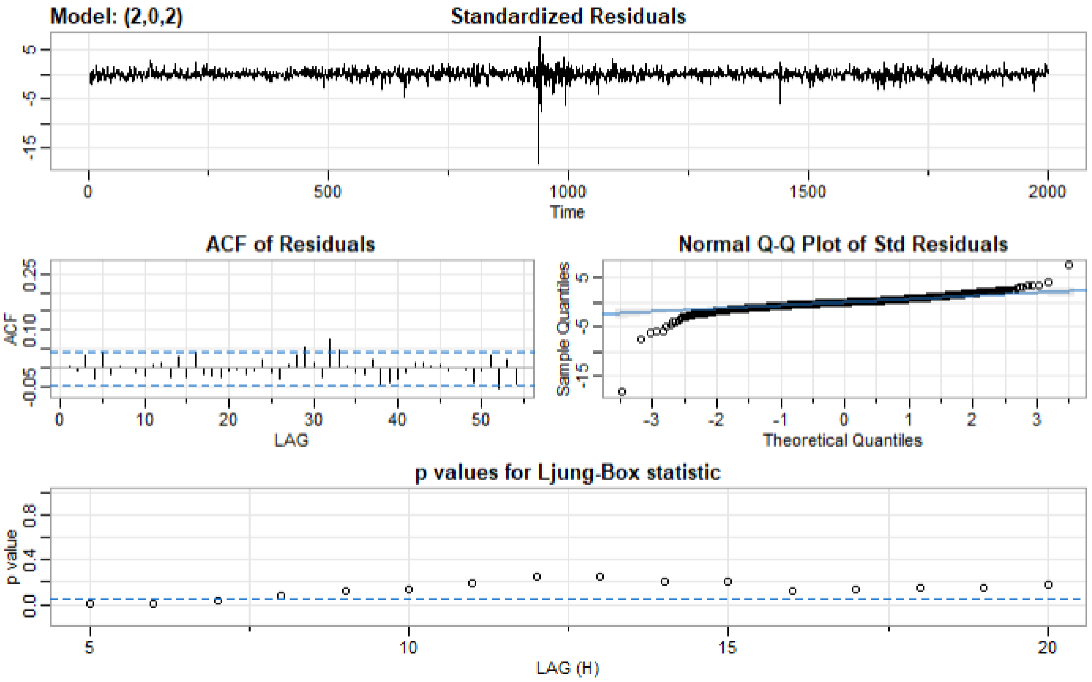
By looking at the dataset, we can see that the data is approximately stationary, with a few outliers. Additional research indicates that the outliers may be the data for the stock market crash in October 19, 1987, known as Black Monday. The red line represents the average and seems to indicate that the mean value of the dataset is 0 and does not depend on time. Results from a Dickey-Fuller test supports this observation and furthermore rejects the null hypothesis that the autocovariance function depends on the time points themselves. Hence there will be no transformations or differencing needed to convert the dataset into a stationary process.

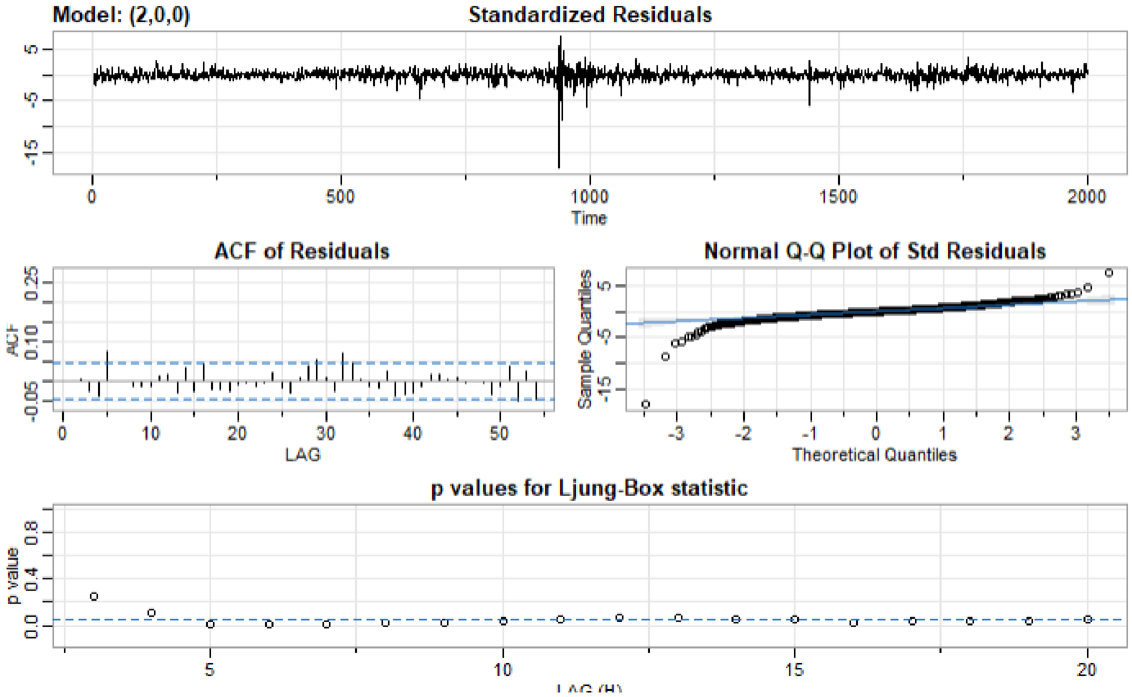
4 Model Selection and Diagnostics



The sample autocovariance seems to decay to zero quite quickly, so there is most likely no need for differencing.

Inspecting the ACF and PACF, we could argue that both the ACF and PACF tail off. Alternatively, we could argue that the ACF tails off and the PACF cuts off at 2. The first model is ARMA(2,0,2), the second model is ARIMA(2,0,0)





ARIMA(2,0,2): AIC: -6.412576, AICc: -6.412561, BIC: -6.395773

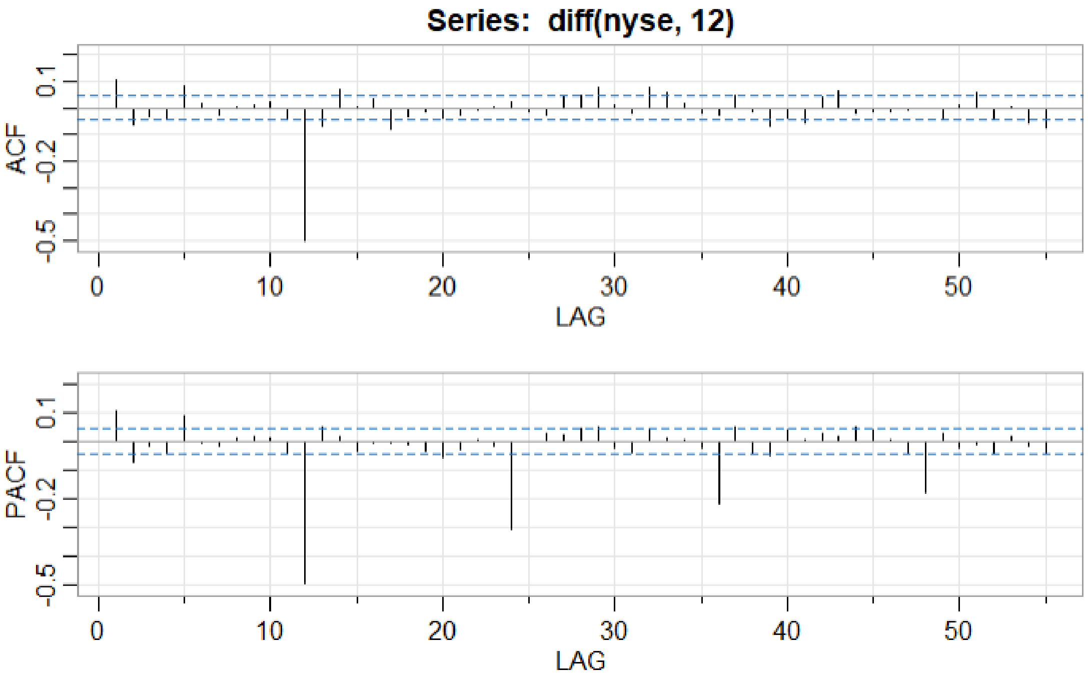
ARIMA(2,0,0): AIC: -6.411386, AICc: -6.41138, BIC: -6.400184

We see no obvious pattern in standardized residuals. Few outliers are exceeding five standard deviations from the mean, however the origin of the outliers seems to be the 1987 stock market crash. ACF Residuals plot show a significant spike at lag 32 in both cases, but not quite enough to be significant at 5% level. This indicates that there is no apparent departure from the randomness assumption of the model. The residual normal QQ plot show that the assumption of normality is reasonable except for the few outliers, represented as outliers at the tails of the QQ plot. The p-values for Ljung-Box statistics are above the reasonable significant level for most lags for the ARIMA(2,0,2) model, but are less so for the ARIMA(2,0,0) model. This shows that we should not reject the null hypothesis that the residuals are independent for the ARIMA(2,0,2) model.

Not all model parameters are statistically significant for the ARIMA(2,0,0) model; the ma1 parameter estimate has a t-test p-value greater than $\alpha = 0.05$, which shows that adding an extra ma parameter does not significantly change the result. For the ARIMA(2,0,2) model, all parameter estimates have t-test p-values below $\alpha = 0.05$ threshold, meaning that all model parameters are statistically significant. The AIC, AICc and BIC are nearly identical for both models.

Based on the above, we are choosing the ARIMA(2,0,2) model between these two models.

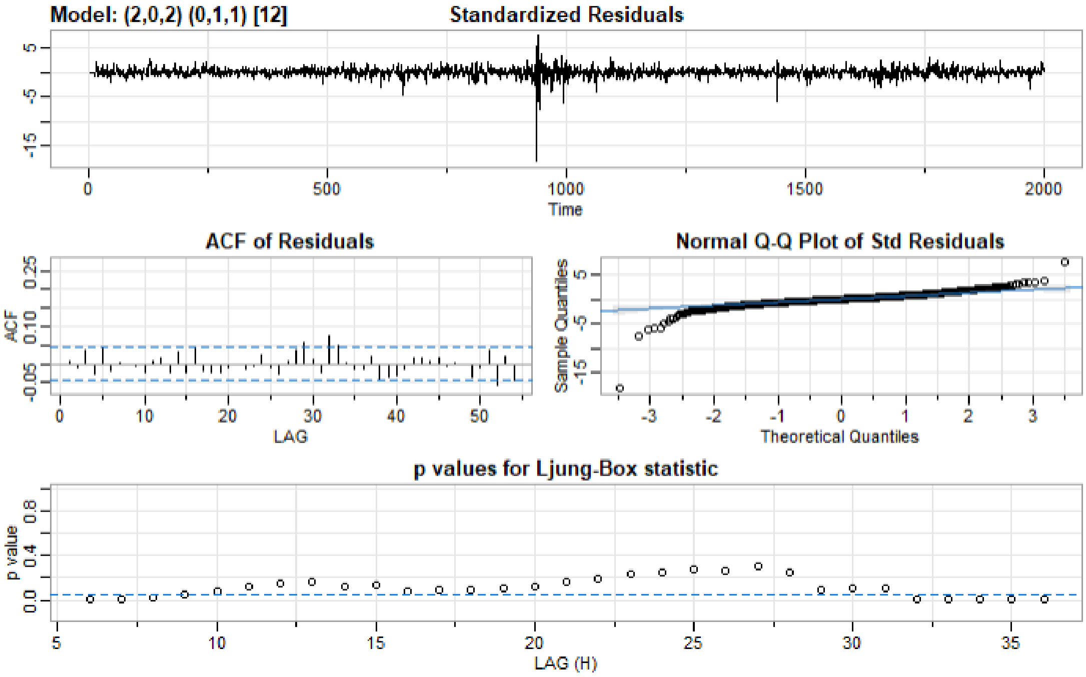
Furthermore, because we are dealing with stock exchange returns, one would expect a seasonal component to be present in the dataset, this is visible from the fluctuations in the dataset plot. We propose a revision to the ARIMA(2,0,2) model by including a seasonal component with $s = 12$:



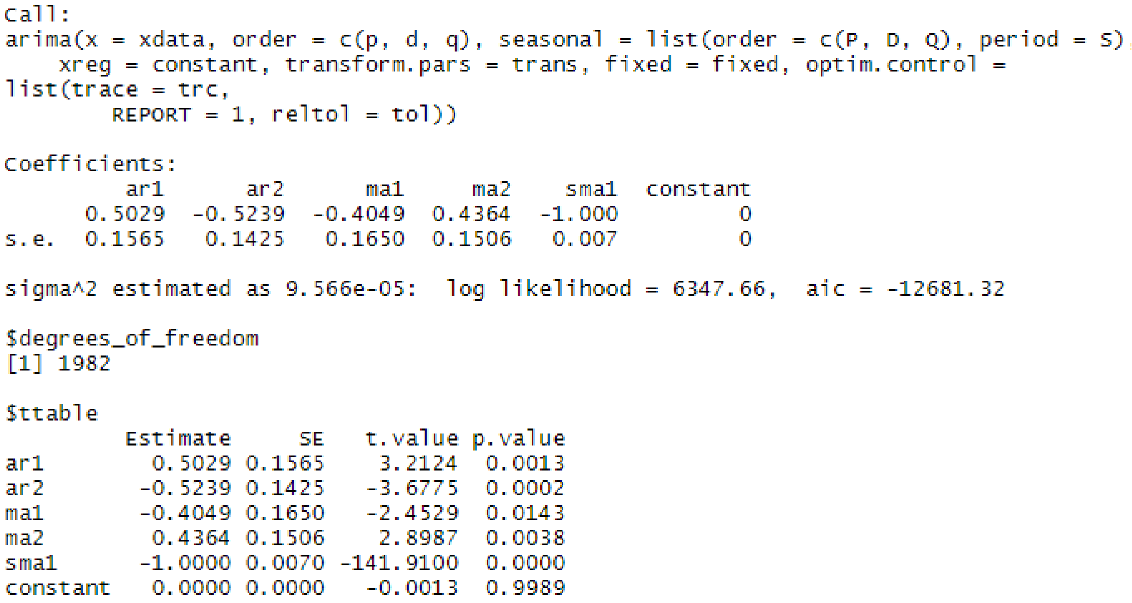
ACF cuts off at lag = 1s (s=12). PACF cuts off at lag = 1s, 2s, 3s, 4s (s=12).

This suggests SARIMA$(2, 0, 2)$x$(0, 1, 1)_{12}$, where the the ACF not tailing off could be a function of the sample auto covariance.

We perform diagnostics for this new model:



We see no obvious pattern in standardized residuals. Few outliers are exceeding five standard deviations from the mean, however the origin of the outliers seems to be the 1987 stock market crash. ACF Residuals plot show a relatively high spike at lag 32, but not significantly high enough at the 5% level. This indicates that there is no apparent departure from the randomness assumption of the model. The residual normal QQ plot show that the assumption of normality is reasonable except for the few outliers, represented as outliers at the tails of the QQ plot. The p-values for Ljung-Box statistics are above the reasonable significant level for most lags for the model.


```
Call:
arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D, Q), period = S),
    xreg = constant, transform.pars = trans, fixed = fixed, optim.control =
list(trace = trc,
        REPORT = 1, reltol = tol))

Coefficients:
         ar1      ar2      ma1     ma2    sma1  constant
      0.5029  -0.5239  -0.4049  0.4364  -1.000         0
s.e.  0.1565   0.1425   0.1650  0.1506   0.007         0

sigma^2 estimated as 9.566e-05:  log likelihood = 6347.66,  aic = -12681.32

$degrees_of_freedom
[1] 1982

$ttable
         Estimate     SE   t.value p.value
ar1        0.5029 0.1565    3.2124  0.0013
ar2       -0.5239 0.1425   -3.6775  0.0002
ma1       -0.4049 0.1650   -2.4529  0.0143
ma2        0.4364 0.1506    2.8987  0.0038
sma1      -1.0000 0.0070 -141.9100  0.0000
constant   0.0000 0.0000   -0.0013  0.9989
```

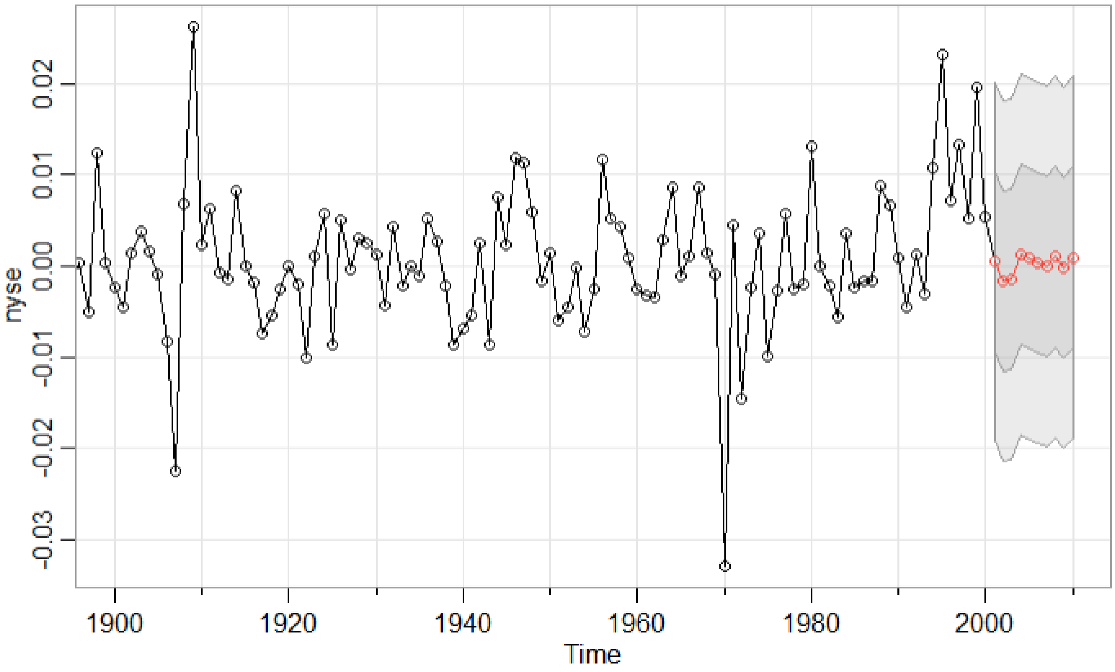
The parameter estimates are $\hat{\phi}_1 = 0.5029$, $\hat{\phi}_2 = -0.5239$, $\hat{\theta}_1 = -0.4049$, $\hat{\theta}_2 = 0.4364$. The ϕ estimates show that stock returns with time lag $h = 1$ apart are positively correlated by factor of 0.5029 and stock returns with time lag $h = 2$ apart are negatively correlated by a factor of 0.5239 The θ estimates show that a random shock on an arbitrary stock exchange return is estimated to be equal to the previous shock multiplied by -0.4049 plus the second previous shock multiplied by 0.4364.

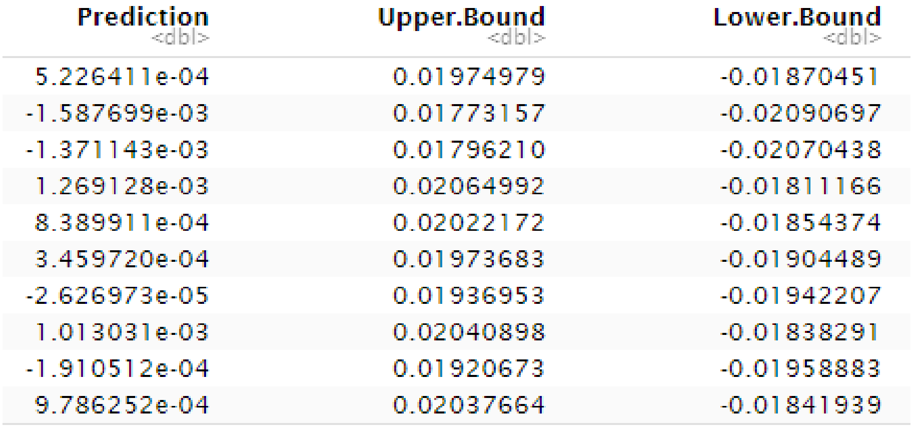
All parameter estimates have p-values below $\alpha = 0.05$ threshold, as showcased in the R output above, meaning that all model parameters are statistically significant. SARIMA$(2, 0, 2)$x$(0, 1, 1)_{12}$ also has less AIC, AICc, BIC than ARIMA(2,0,2) model.

Estimate for the constant term is not significant at $\alpha = 0.05$, this could imply that there is no apparent drift in the differenced *nyse*, as denoted on slide 50 in lecture 9. Therefore, we will add the command no.constant=TRUE in R, removing the constant term.

5 Forecasting

The plot below showcases forecast into the next ten-time periods (represented by the red points) and their respective 95% confidence intervals.

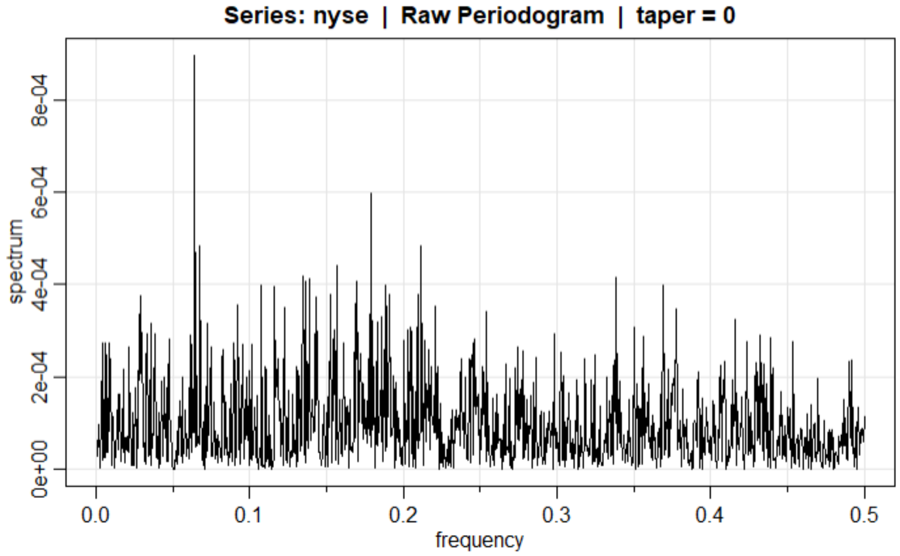




Prediction	Upper.Bound	Lower.Bound
<dbl>	<dbl>	<dbl>
5.226411e-04	0.01974979	-0.01870451
-1.587699e-03	0.01773157	-0.02090697
-1.371143e-03	0.01796210	-0.02070438
1.269128e-03	0.02064992	-0.01811166
8.389911e-04	0.02022172	-0.01854374
3.459720e-04	0.01973683	-0.01904489
-2.626973e-05	0.01936953	-0.01942207
1.013031e-03	0.02040898	-0.01838291
-1.910512e-04	0.01920673	-0.01958883
9.786252e-04	0.02037664	-0.01841939

The fluctuations in the predictions do not seem to be as drastic as the prior points in the dataset; the seasonal trend may not be very well incorporated in the final model. Additionally, the confidence intervals are also very wide, represented as the grey regions on the plot. These are signs that the model has significant room for improvement.

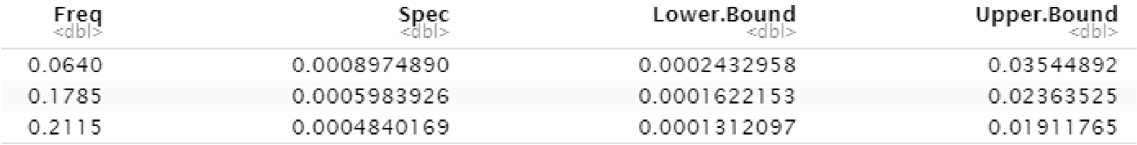
6 Spectral Analysis



The periodogram of the dataset shows us the approximate indices of the top three dominant frequencies, which are indices 128, 357, and 423 respectively.

Most dominant period has a frequency of 0.064, which has a cycle of 15.625 daily exchange returns. Second most dominant period has a frequency of 0.1785, which has a cycle of 5.602241 daily exchange returns. Third most dominant period has a frequency of 0.0675, which has a cycle of 14.81481 daily exchange returns.

Below is the table for the 95% confidence intervals of the dominant periods.



Freq <dbl>	Spec <dbl>	Lower.Bound <dbl>	Upper.Bound <dbl>
0.0640	0.0008974890	0.0002432958	0.03544892
0.1785	0.0005983926	0.0001622153	0.02363525
0.2115	0.0004840169	0.0001312097	0.01911765

We cannot establish the significance of the first peak since the periodogram ordinate is 0.0008974890, which lies in the confidence intervals of the second and third peak. We cannot establish the significance of the second peak since the periodogram ordinate is 0.0005983926, which lies in the confidence interval of the first and third peak. We cannot establish the significance of the third peak since the periodogram ordinate is 0.0004840169, which lies in the confidence interval of the second peak.

7 Conclusion

Our spectral analysis helps us conclude that the three biggest oscillations in daily exchange returns happens every 15.625 trading days, 5.602241 trading days, and every 14.81481 trading days.

The parameter estimates show that stock returns that are 1 trading day apart are positively correlated by factor of 0.5029 and stock returns that are 2 trading days apart are negatively correlated by a factor of 0.5239. The short-term fluctuations in stock exchange returns are based on the fluctuation from the mean in the previous trading day by a factor of -0.4049 plus the fluctuation from mean in the second previous trading day by a factor of 0.4364.

8 Limitations

The model predictions may have been affected negatively by the outliers detected at the tails of the QQ plot. Although the final model incorporated a seasonal trend with s=12, there is room for debate about the existence of a better fitting seasonal trend: The dataset is not taken monthly but rather daily (only the trading days), so it is possible that there is instead a quarterly or yearly seasonal component in the dataset.

The ACF and PACF plots of the nyse dataset suggests the possibility of more than two or three models. For example, one could consider that the sample ACF decays too slow as h increases and consider a model that differences the data once. (d=1). It is entirely possible that there are far more fitting models for this dataset, as evident in the forecasting section of this report, where the predictions fluctuate relatively less compared to the dataset and have very wide confidence intervals. Therefore, the first step one would take to further this report is to consider integrated models.

9 References

1. James F. Royal. What is the average stock market return? Retrieved December 16, 2021 from https://www.nerdwallet.com/article/investing/average-stock-market-return

2. @Manual, title = astsa: Applied Statistical Time Series Analysis, author = David Stoffer, year = 2021, note = R package version 1.14, url = https://CRAN.R-project.org/package=astsa,