STA457 Final Proejct

T.Z. 6607

14/12/2021

Abstract

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.

Introduction

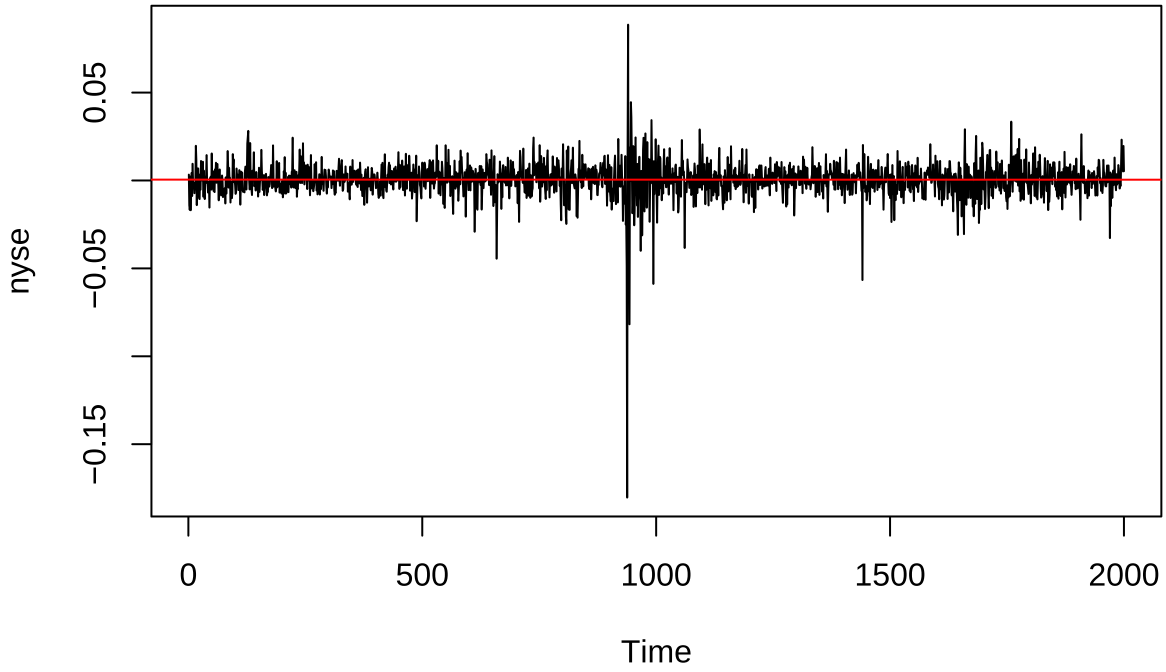
Internet sources commonly state that the historical average stock market return is about 8-12%. While this average is promising, there were only six times where the returns were within this range between 1926 and 2014. Most of the times, the returns are either much lower or much higher; there is a high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

This report analyzes the *nyse* dataset found in the *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.

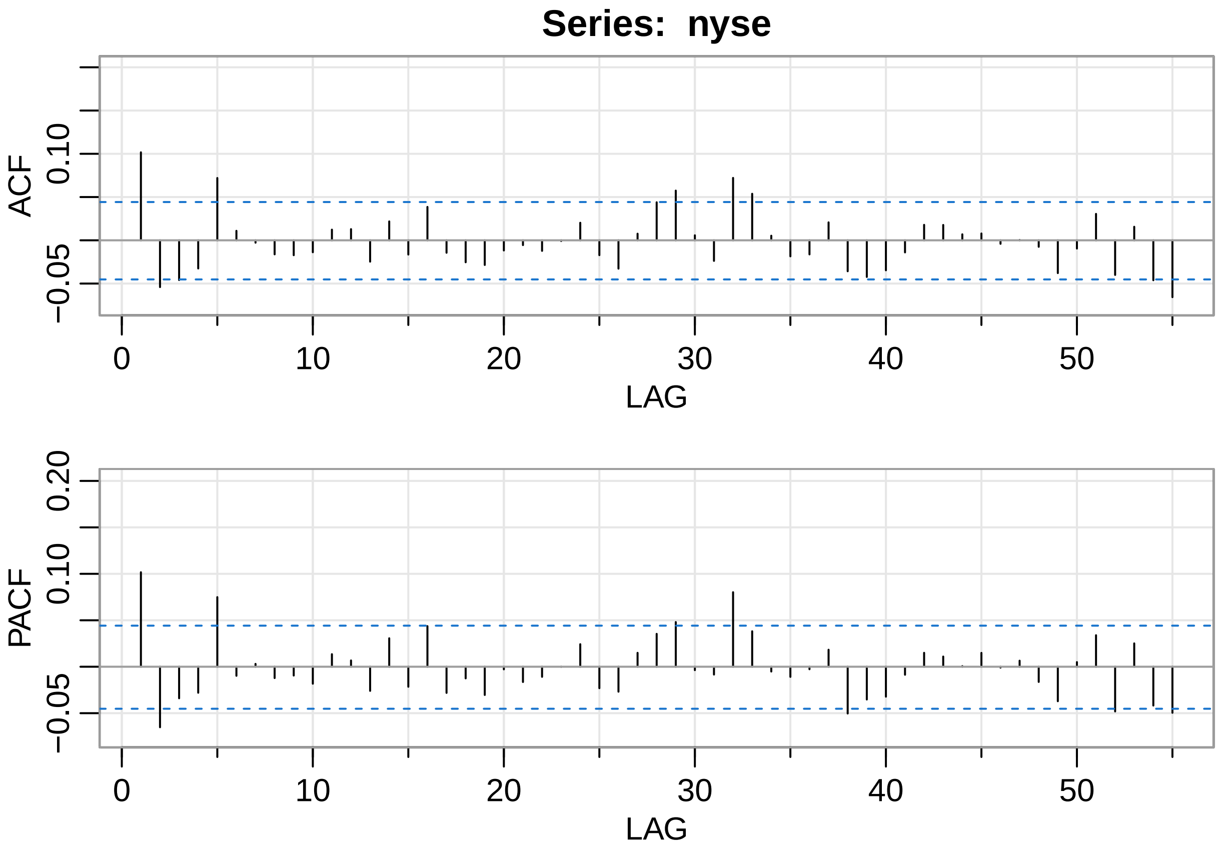
The following analysis in this report is dedicated to finding underlying trends, forecasting future values, and presenting the findings in hopes of helping future investors make more educated and beneficial decisions on the stock market.

Statistical Methods



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 seems to indicate that the mean value of the dataset is 0 and does not depend on time.

Results from the Dickey-Fuller test supports this observation.

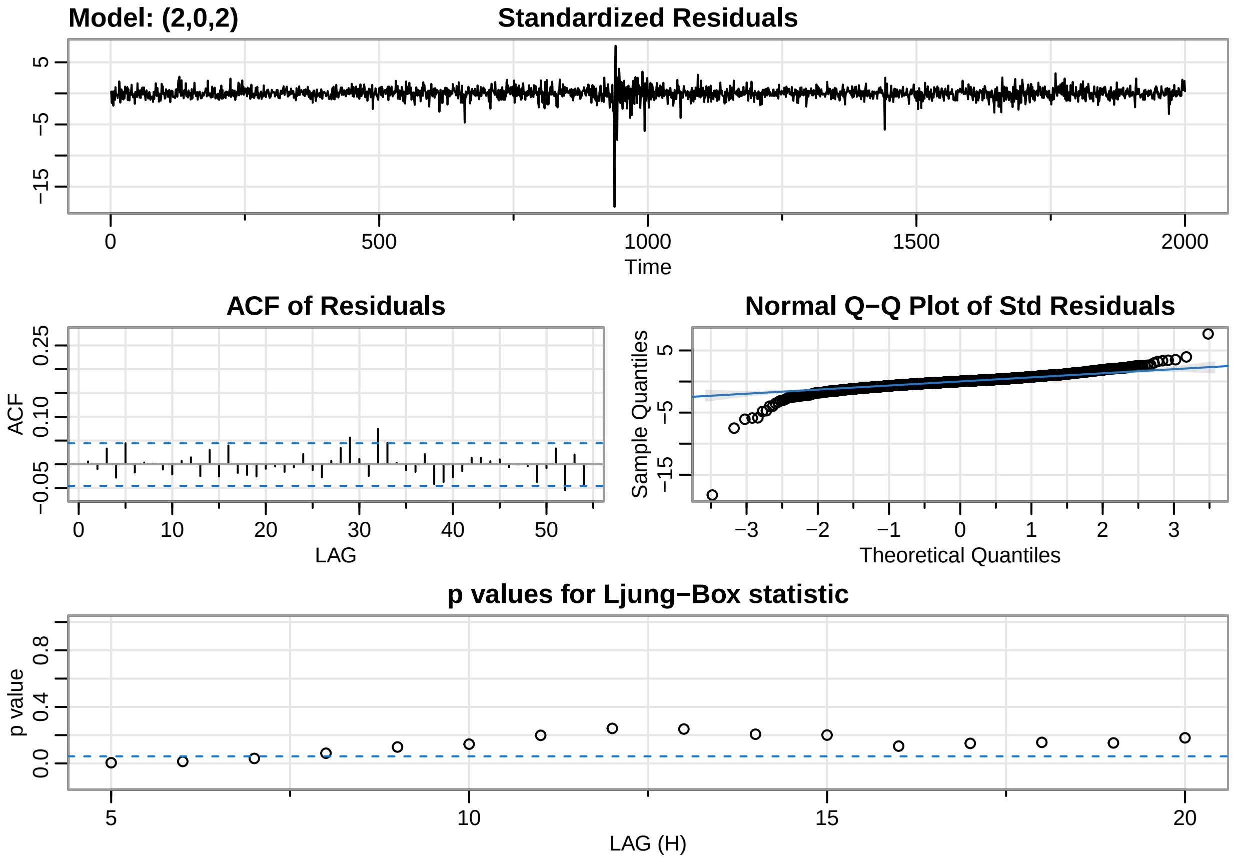


```
##      [,1]  [,2]  [,3]  [,4] [,5]  [,6] [,7]  [,8]  [,9] [,10] [,11] [,12] [,13]
## ACF   0.1 -0.05 -0.05 -0.03 0.07  0.01    0 -0.02 -0.02 -0.01  0.01  0.01 -0.02
## PACF  0.1 -0.07 -0.03 -0.03 0.07 -0.01    0 -0.01 -0.01 -0.02  0.01  0.01 -0.03
##      [,14] [,15] [,16] [,17] [,18] [,19] [,20] [,21] [,22] [,23] [,24] [,25]
## ACF   0.02 -0.02  0.04 -0.01 -0.03 -0.03 -0.01 -0.01 -0.01     0  0.02 -0.02
## PACF  0.03 -0.02  0.04 -0.03 -0.01 -0.03  0.00 -0.02 -0.01     0  0.02 -0.02
##      [,26] [,27] [,28] [,29] [,30] [,31] [,32] [,33] [,34] [,35] [,36] [,37]
## ACF  -0.03  0.01  0.04  0.06  0.01 -0.02  0.07  0.05  0.01 -0.02 -0.02  0.02
## PACF -0.03  0.01  0.04  0.05  0.00 -0.01  0.08  0.04 -0.01 -0.01  0.00  0.02
##      [,38] [,39] [,40] [,41] [,42] [,43] [,44] [,45] [,46] [,47] [,48] [,49]
## ACF  -0.04 -0.04 -0.03 -0.01  0.02  0.02  0.01  0.01     0  0.00 -0.01 -0.04
## PACF -0.05 -0.04 -0.03 -0.01  0.01  0.01  0.00  0.01     0  0.01 -0.02 -0.04
##      [,50] [,51] [,52] [,53] [,54] [,55]
## ACF  -0.01  0.03 -0.04  0.02 -0.05 -0.07
## PACF  0.01  0.03 -0.05  0.03 -0.04 -0.05
```

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)

Results

```r
sarima(nyse,2,0,2, details=FALSE)
#fit the ARIMA(2,0,2) model
ARIMA1 <- sarima(nyse, 2, 0, 2)
```



AIC: -6.412576, AICc: -6.412561, BIC: -6.395773

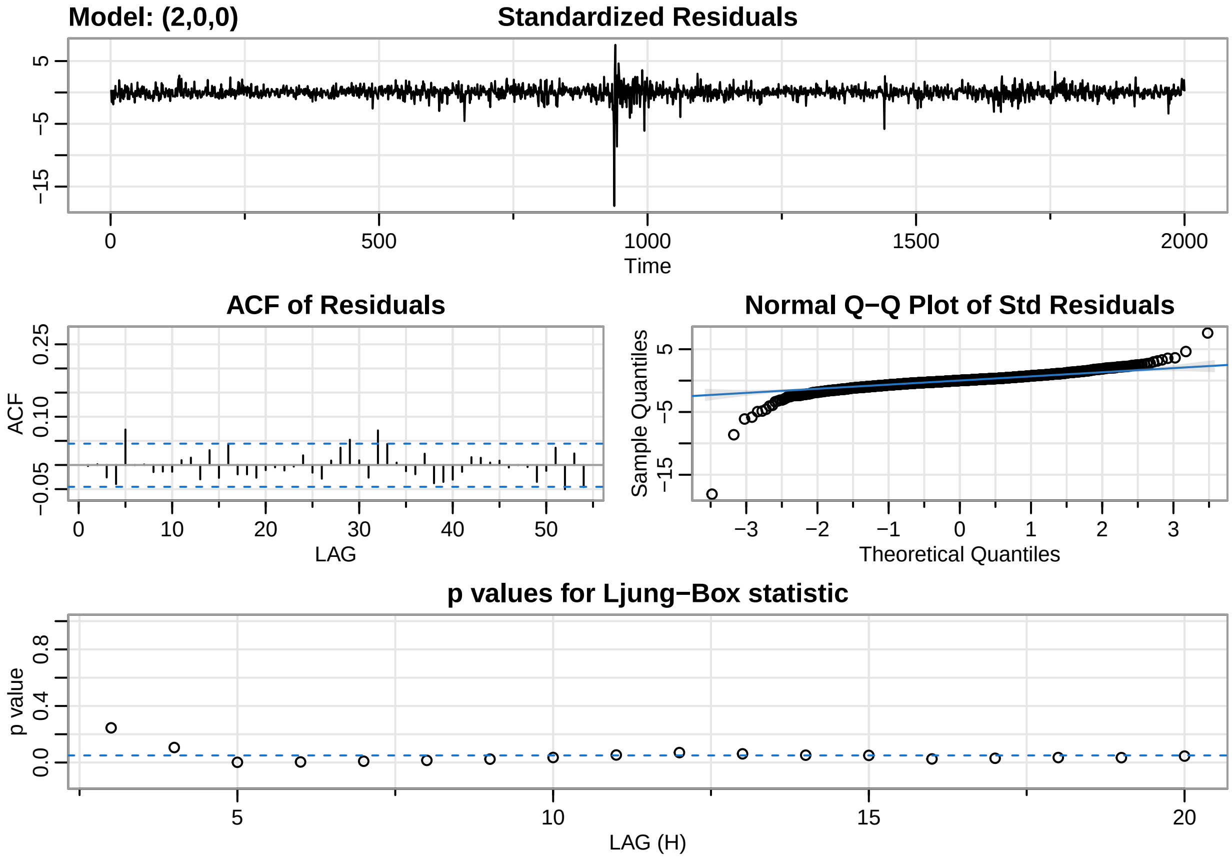
```r
sarima(nyse,2,0,0, details=FALSE)
```

```
## $fit
##
## Call:
## arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D, Q), period = S),
##     xreg = xmean, include.mean = FALSE, transform.pars = trans, fixed = fixed,
##     optim.control = list(trace = trc, REPORT = 1, reltol = tol))
##
## Coefficients:
##          ar1      ar2  xmean
##       0.1083  -0.0652  5e-04
## s.e.  0.0223   0.0223  2e-04
##
## sigma^2 estimated as 9.58e-05:  log likelihood = 6415.39,  aic = -12822.77
##
## $degrees_of_freedom
## [1] 1997
##
```

```
## $ttable
##       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
##
## $AIC
## [1] -6.411386
##
## $AICc
## [1] -6.41138
##
## $BIC
## [1] -6.400184
```

```
#fit the ARIMA(2,0,0) model
ARIMA2 <- sarima(nyse, 2, 0, 0)
```

```
## initial  value -4.619376
## iter   2 value -4.626662
## iter   3 value -4.626737
## iter   4 value -4.626738
## iter   4 value -4.626738
## iter   4 value -4.626738
## final  value -4.626738
## converged
## initial  value -4.626632
## iter   1 value -4.626632
## final  value -4.626632
## converged
```



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. Therefore, we select the ARIMA(2,0,2) model for prediction.

Not all model parameters are statistically significant for the ARIMA(2,0,0) model; the ma1 parameter estimate has a p-value greater than $\alpha = 0.05$, which shows that adding an extra ma parameter does not significantly change the result. The AIC, AICc and BIC are nearly identical for both models. Based on the above, we are choosing the ARIMA(2,0,2) model for forecast as well.

```r
sarima(nyse,2,0,2, details=FALSE)
```

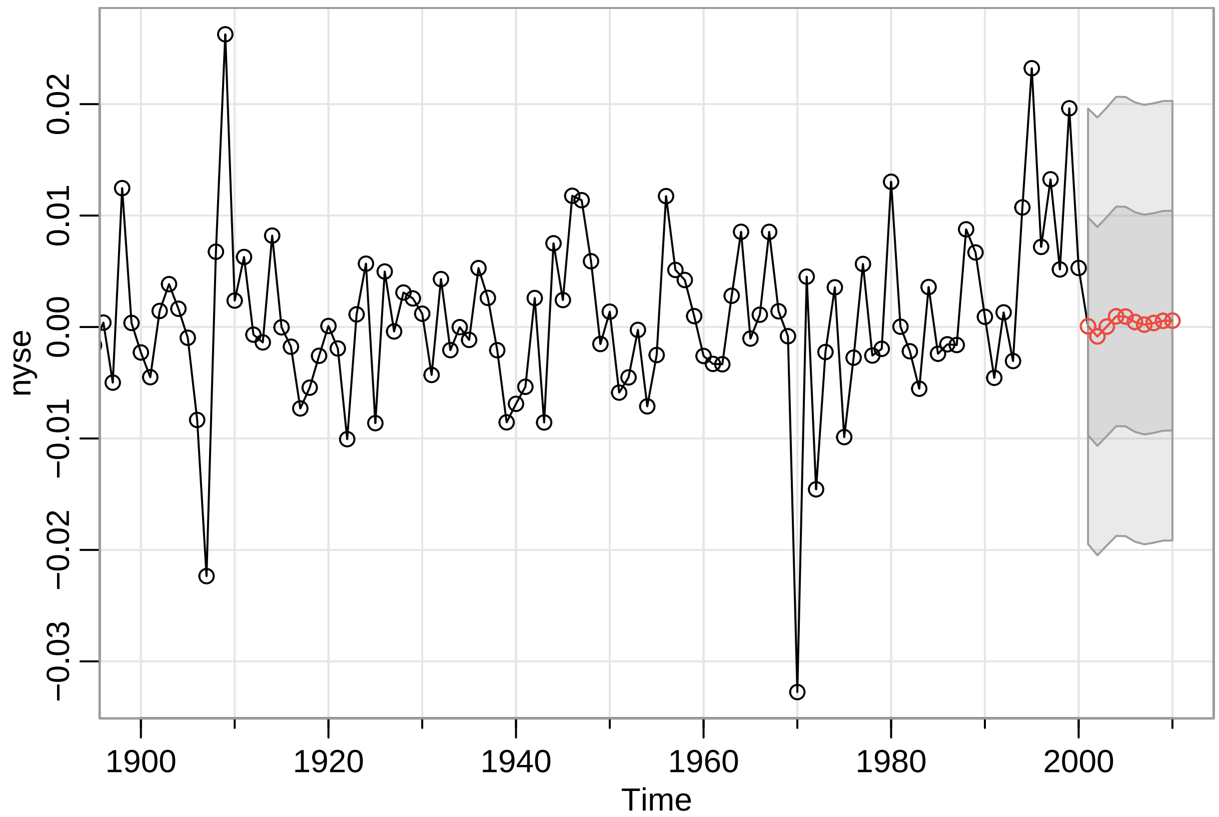
```
## $fit
##
## Call:
## arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D, Q), period = S),
##     xreg = xmean, include.mean = FALSE, transform.pars = trans, fixed = fixed,
##     optim.control = list(trace = trc, REPORT = 1, reltol = tol))
##
## Coefficients:
```

```
##          ar1      ar2      ma1     ma2  xmean
##       0.4927  -0.5279  -0.3944  0.4413  5e-04
## s.e.  0.1592   0.1394   0.1678  0.1473  2e-04
##
## sigma^2 estimated as 9.549e-05:  log likelihood = 6418.58,  aic = -12825.15
##
## $degrees_of_freedom
## [1] 1995
##
## $ttable
##       Estimate     SE t.value p.value
## ar1     0.4927 0.1592  3.0941  0.0020
## ar2    -0.5279 0.1394 -3.7856  0.0002
## ma1    -0.3944 0.1678 -2.3512  0.0188
## ma2     0.4413 0.1473  2.9952  0.0028
## xmean   0.0005 0.0002  2.1308  0.0332
##
## $AIC
## [1] -6.412576
##
## $AICc
## [1] -6.412561
##
## $BIC
## [1] -6.395773
```

The parameter estimates are $\hat{\phi}_1 = 0.4927$, $\hat{\phi}_2 = -0.5279$, $\hat{\theta}_1 = -0.3944$, $\hat{\theta}_2 = 0.4413$. The ϕ estimates show that stock returns with time lag $h = 1$ apart are positively correlated by factor of 0.4927 and stock returns with time lag $h = 2$ apart are negatively correlated by a factor of 0.5279. The θ estimates show that a random shock on a stock return affects the $h = 1$ future stock return shocks by a factor of -0.3944 and affects the $h = 2$ future stock return shock by a factor of 0.4413.

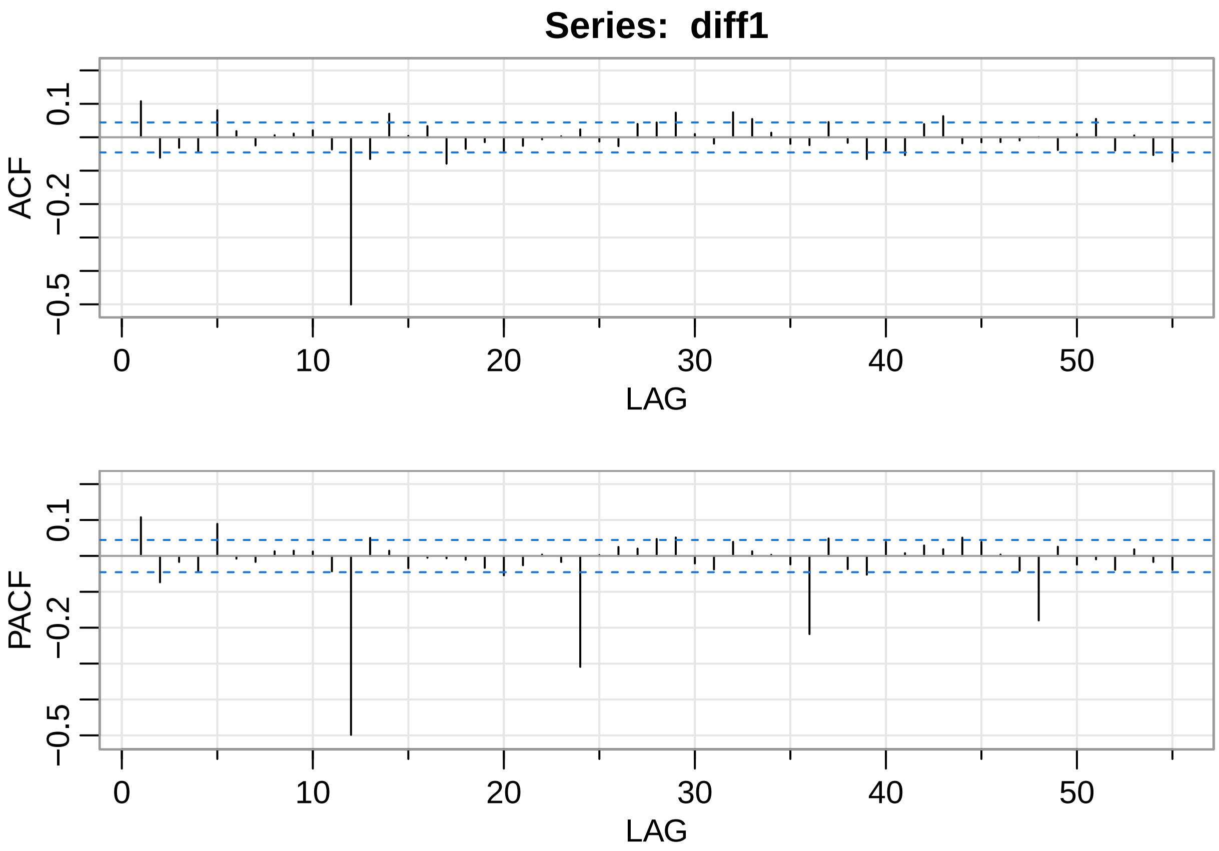
All parameter estimates have p-values below $\alpha = 0.05$ threshold, meaning that all model parameters are statistically significant.

```
#Forecast next ten-time periods
pred1 <- sarima.for(nyse, 10, 2, 0, 2)
```



There seems to be a seasonal trend, which is expected for stock exchange returns. Further revision to model:

```
diff1 = diff(nyse,12)
acf2(diff1)
```



```
##      [,1]  [,2]  [,3]  [,4] [,5]  [,6]  [,7] [,8] [,9] [,10] [,11] [,12] [,13]
## ACF  0.11 -0.06 -0.03 -0.04 0.08  0.02 -0.02 0.01 0.01  0.02 -0.04  -0.5 -0.07
## PACF 0.11 -0.07 -0.02 -0.04 0.09 -0.01 -0.02 0.01 0.01  0.01 -0.04  -0.5  0.05
##      [,14] [,15] [,16] [,17] [,18] [,19] [,20] [,21] [,22] [,23] [,24] [,25]
## ACF   0.07  0.00  0.03 -0.08 -0.03 -0.01 -0.04 -0.03 -0.01  0.00  0.02 -0.01
## PACF  0.01 -0.03 -0.01 -0.01 -0.01 -0.03 -0.05 -0.03  0.00 -0.02 -0.31  0.00
##      [,26] [,27] [,28] [,29] [,30] [,31] [,32] [,33] [,34] [,35] [,36] [,37]
## ACF  -0.03  0.04  0.04  0.07  0.01 -0.02  0.07  0.05  0.01 -0.02 -0.02  0.05
## PACF  0.03  0.02  0.05  0.05 -0.02 -0.04  0.04  0.01  0.00 -0.02 -0.22  0.05
##      [,38] [,39] [,40] [,41] [,42] [,43] [,44] [,45] [,46] [,47] [,48] [,49]
## ACF  -0.02 -0.07 -0.04 -0.05  0.04  0.06 -0.02 -0.02 -0.01 -0.01  0.00 -0.04
## PACF -0.04 -0.05  0.04  0.01  0.03  0.02  0.05  0.04  0.00 -0.04 -0.18  0.03
##      [,50] [,51] [,52] [,53] [,54] [,55]
## ACF   0.01  0.05 -0.04  0.01 -0.05 -0.07
## PACF -0.02 -0.01 -0.04  0.02 -0.02 -0.04
```

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.

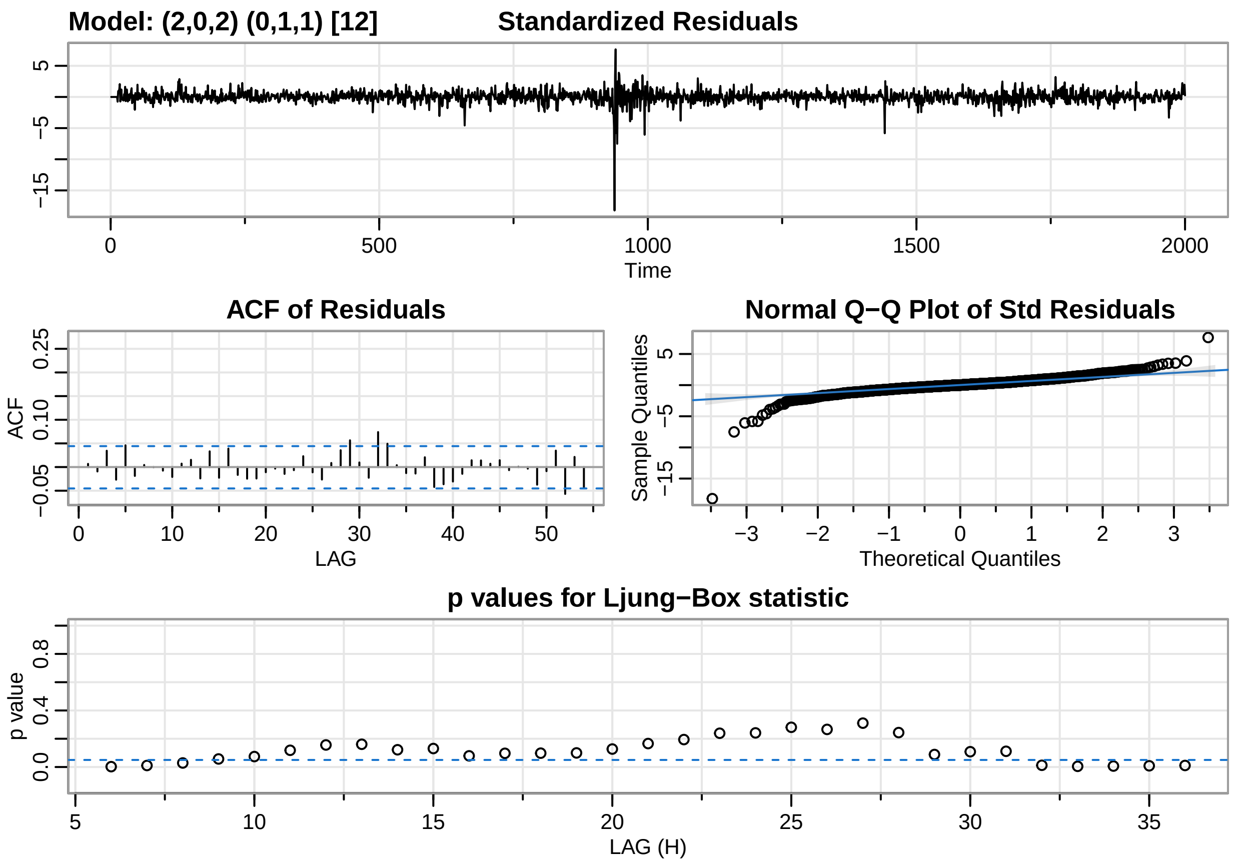
```
sarima(nyse, p=2,d=0,q=2, P=0,D=1,Q=1, S=12, details=FALSE)
```

```
## $fit
##
```

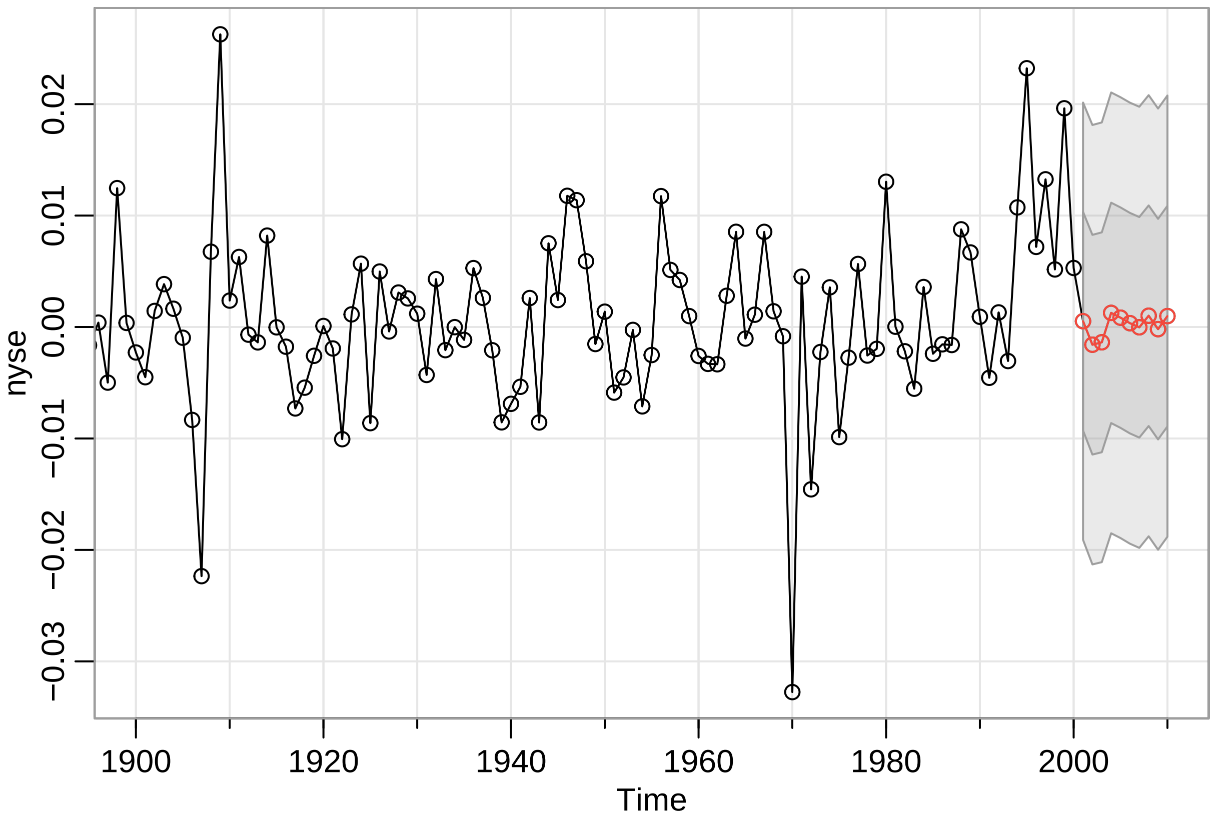
```
## Call:
## arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D, Q), period = S),
##     xreg = constant, transform.pars = trans, fixed = fixed, optim.control = list(trace = trc,
##         REPORT = 1, reltol = tol))
##
## Coefficients:
##          ar1      ar2      ma1     ma2    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
##
## $AIC
## [1] -6.378934
##
## $AICc
## [1] -6.378913
##
## $BIC
## [1] -6.359234
```

```
#fit the SARIMA(2,0,2)x(0,1,1)_{12} model
ARIMA3 <- sarima(nyse, p=2,d=0,q=2, P=0,D=1,Q=1, S=12, no.constant = TRUE)
```

```
## initial  value -4.279148
## iter   2 value -4.477695
## iter   3 value -4.540959
## iter   4 value -4.588366
## iter   5 value -4.589356
## iter   6 value -4.590390
## iter   7 value -4.590992
## iter   8 value -4.591078
## iter   9 value -4.592036
## iter  10 value -4.593136
## iter  11 value -4.593770
## iter  12 value -4.593816
## iter  13 value -4.593861
## iter  14 value -4.593870
## iter  15 value -4.593886
## iter  16 value -4.593935
## iter  17 value -4.594069
## iter  18 value -4.594462
## iter  19 value -4.595265
## iter  20 value -4.595533
## iter  21 value -4.595765
## iter  22 value -4.596033
## iter  23 value -4.596153
## iter  24 value -4.596209
## iter  25 value -4.596251
## iter  26 value -4.596272
## iter  27 value -4.596311
## iter  28 value -4.596347
## iter  29 value -4.596381
## iter  30 value -4.596423
## iter  31 value -4.596448
## iter  32 value -4.596494
## iter  33 value -4.596536
## iter  34 value -4.596624
## iter  35 value -4.596635
## iter  36 value -4.596657
## iter  37 value -4.596660
## iter  38 value -4.596663
## iter  39 value -4.596671
## iter  40 value -4.596674
## iter  41 value -4.596678
## iter  42 value -4.596679
## iter  43 value -4.596680
## iter  44 value -4.596682
## iter  45 value -4.596686
## iter  46 value -4.596687
## iter  47 value -4.596688
## iter  48 value -4.596688
## iter  49 value -4.596691
## iter  50 value -4.596691
## iter  51 value -4.596691
## iter  52 value -4.596693
## iter  53 value -4.596694
## iter  54 value -4.596694
## iter  54 value -4.596694
## iter  54 value -4.596694
## final  value -4.596694
## converged
## initial  value -4.601762
## iter   2 value -4.608201
## iter   3 value -4.609490
## iter   4 value -4.609606
## iter   5 value -4.609686
## iter   6 value -4.609737
## iter   7 value -4.609743
## iter   8 value -4.609800
## iter   9 value -4.609908
## iter  10 value -4.610045
## iter  11 value -4.610175
## iter  12 value -4.610239
## iter  13 value -4.610241
## iter  14 value -4.610242
## iter  15 value -4.610243
## iter  16 value -4.610244
## iter  17 value -4.610308
## iter  18 value -4.610328
## iter  19 value -4.610381
## iter  20 value -4.610519
## iter  21 value -4.610728
## iter  22 value -4.610810
## iter  23 value -4.610861
## iter  24 value -4.610974
## iter  25 value -4.611030
## iter  26 value -4.611036
## iter  27 value -4.611043
## iter  28 value -4.611074
## iter  29 value -4.611148
## iter  30 value -4.611282
## iter  31 value -4.611426
## iter  32 value -4.611800
## iter  33 value -4.611889
## iter  34 value -4.611923
## iter  35 value -4.611924
## iter  36 value -4.611925
## iter  37 value -4.611925
## iter  38 value -4.611925
## iter  38 value -4.611925
## iter  38 value -4.611925
## final  value -4.611925
## converged
```



```
#Forecast next ten-time periods
pred2 <- sarima.for(nyse, 10, 2,0,2, 0,1,1, 12)
```



```
#z value = 1.96 for 95% CI (1.96 std.dev.)
#upper bound
U = pred2$pred+1.96*pred2$se
#lower bound
L = pred2$pred-1.96*pred2$se

tabl <- data.frame('Prediction'=c(pred2$pred),'Upper Bound'=c(U), 'Lower Bound'=c(L))
tabl
```

```
##       Prediction Upper.Bound Lower.Bound
## 1   5.226411e-04  0.01974979 -0.01870451
## 2  -1.587699e-03  0.01773157 -0.02090697
## 3  -1.371143e-03  0.01796210 -0.02070438
## 4   1.269128e-03  0.02064992 -0.01811166
## 5   8.389911e-04  0.02022172 -0.01854374
## 6   3.459720e-04  0.01973683 -0.01904489
## 7  -2.626973e-05  0.01936953 -0.01942207
## 8   1.013031e-03  0.02040898 -0.01838291
## 9  -1.910512e-04  0.01920673 -0.01958883
## 10  9.786252e-04  0.02037664 -0.01841939
```

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.

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.

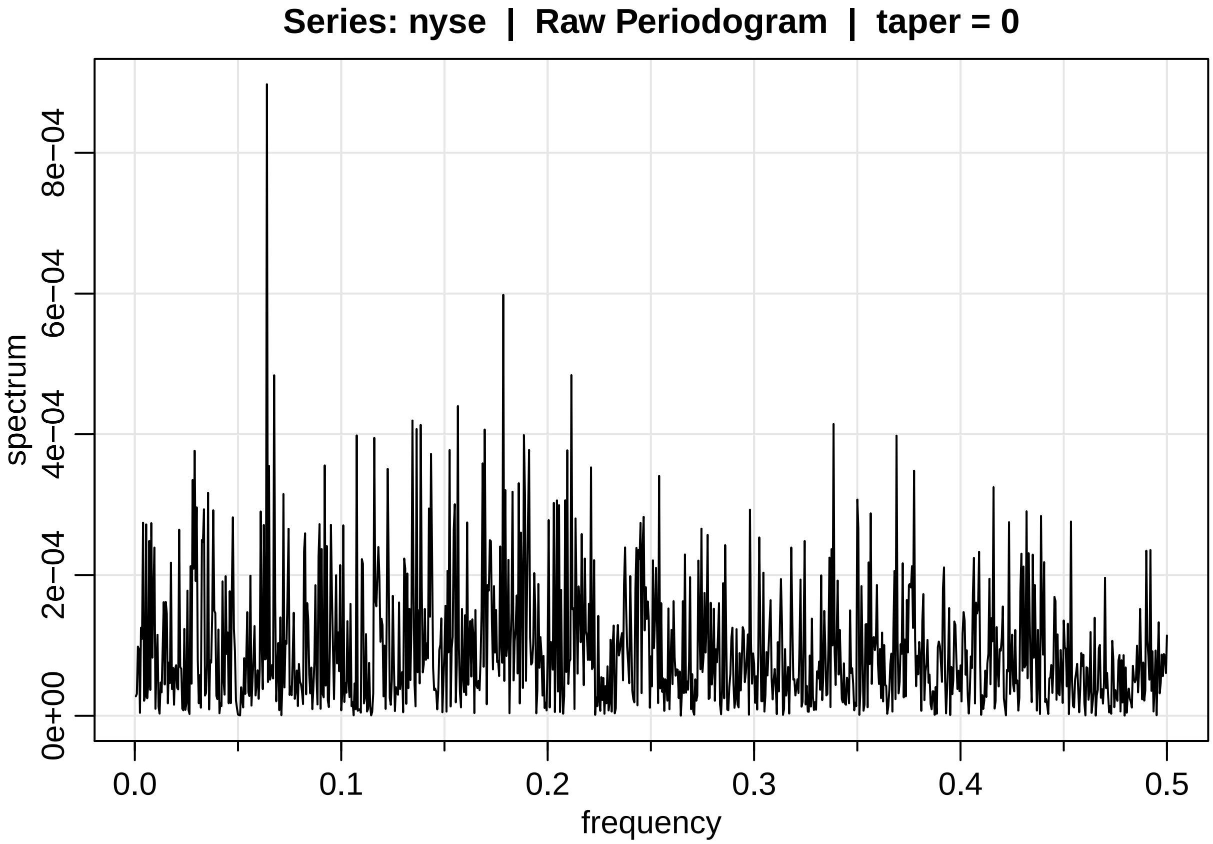
All parameter estimates have p-values below $\alpha = 0.05$ threshold, meaning that all model parameters are statistically significant.

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.

```
#Spectral Analysis
nyse.per = mvspec(nyse, log="no")
```



```
#The index for the predominant periods are 128, 357, 423
order(nyse.per$spec, decreasing = TRUE)[1:3]
```

```
## [1] 128 357 423
```

```
P1 <- nyse.per$details[order(nyse.per$details[,3],decreasing=TRUE),]
#Top three dominant frequencies
P1[1,1];P1[2,1];P1[3,1]
```

```
## frequency
##     0.064
```

```
## frequency
##    0.1785
```

```
## frequency
##    0.0675
```

```
# When do cycles occur
cat("Cycles occur at", 1/P1[1,1], 1/P1[2,1], 1/P1[3,1])
```

```
## Cycles occur at 15.625 5.602241 14.81481
```

```
#95% CIs
U = qchisq(0.025,2)
L = qchisq(0.975,2)
Freqs <- c(nyse.per$freq[128], nyse.per$freq[357], nyse.per$freq[423])
Specs <- c(nyse.per$spec[128], nyse.per$spec[357], nyse.per$spec[423])
LB <- c(2*nyse.per$spec[128]/L, 2*nyse.per$spec[357]/L, 2*nyse.per$spec[423]/L)
UB <- c(2*nyse.per$spec[128]/U, 2*nyse.per$spec[357]/U, 2*nyse.per$spec[423]/U)
tabl2 <- data.frame('Freq'=Freqs, 'Spec'=Specs, 'Lower Bound'=LB, 'Upper Bound'=UB)
tabl2
```

```
##     Freq         Spec  Lower.Bound Upper.Bound
## 1 0.0640 0.0008974890 0.0002432958  0.03544892
## 2 0.1785 0.0005983926 0.0001622153  0.02363525
## 3 0.2115 0.0004840169 0.0001312097  0.01911765
```

The periodogram supports our findings for the top three predominant periods.

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

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.

Discussion

The spectral analysis section tells us that the top three predominant 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 and the fluctuation from mean in the second previous trading day by a factor of 0.4364.

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.