Stats 443, Group Project
Priyanka Sethi (53098133) Nicholas Ho (53693123) Jiaxuan Chen (85250158)

Introduction

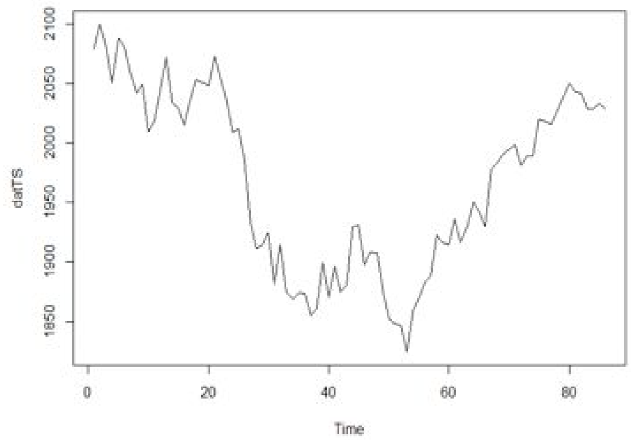
For our group project, we analyzed the S&P 500s futures from 30 November, 2015 till 28 March, 2016. The S&P 500 futures give a reasonable estimate of the changes in the market in the short term. With the large amount of data available, people can accurately predict changes in the S&P 500s, and hence, changes in the market to mitigate risk. Our motivations behind investigating this data were to see if we could, within a range of accuracy, predict future values of the S&P 500 futures using a smaller sample of data (around 100 data points), and whether our data could be modelled well in the frequency domain.

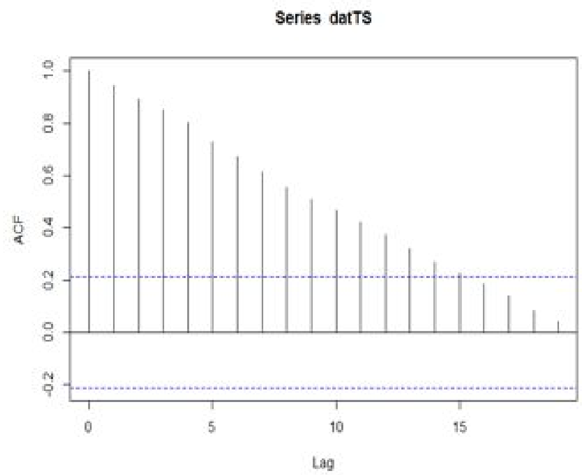
The data was collected from the website "http://ca.investing.com/", and then converted to a time-series indexed from 1 to 86.

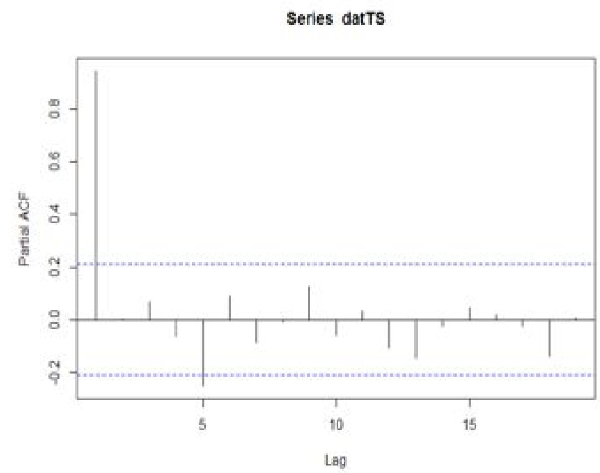
This study attempts to forecast the next day's closing price of a time series in S&P 500s futures using different models and attempts to explore the data within the frequency domain. The developed models are evaluated and the results are compared by comparing the values they predict for March 29, 2016 - April 2, 2016 with the actual observed values for that time period.

Analysis

The following plot, acf and pacf of the data are presented below:







At a first glance at the plot of the series, the series does not look like white noise and a trend is not clear. The ACF of the data decays slowly suggests high correlation between successive data points, and the PACF of the data seems to cut-off after lag 1. Therefore, looking at the ACF and PACF of the data, we can estimate an AR(1) model for the data.

MODEL 1 - AR(1):

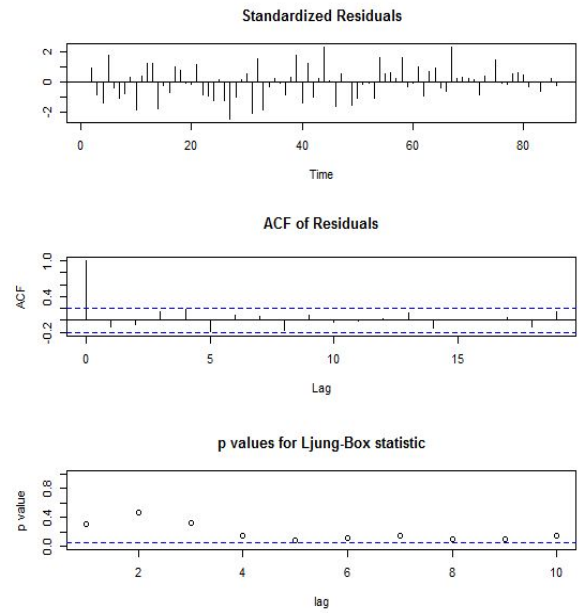
1.1

An ARIMA(1,0,0) or AR(1) model is fitted by minimizing the conditional sum-of-squares.The model diagnostic plots show that the p-values are not well above the significance bound. The parameter estimates are 0.9509 and 155.6248 (intercept) with standard errors 0.0300 and 46.5998 respectively. The variance is estimated to be 440 and log-likelihood to be -383.76.

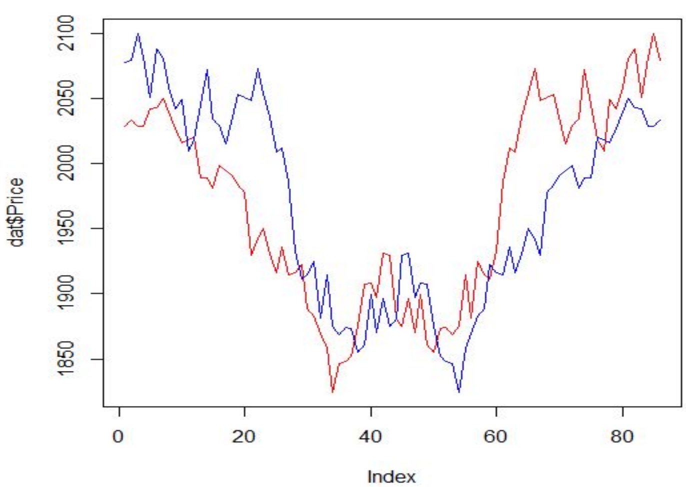
The predicted closing prices for the next 4 days would be 2024.919, 2021.514, 2018.276, 2015.197 with SEs 20.97584, 28.94464, 34.60438, 39.02120 while the actual closing prices are 2047.5, 2055.25, 2051.5, 2065.

1.2

This is an ARIMA (1,0,0) model fitted by the maximum-likelihood method. We again see that the p-values are not very high although exceeding the significance bound. The parameter estimates are 1 and 1987.889 (intercept) and the standard error for the intercept is 10366.583. The variance is estimated to be 448.8, log-likelihood of -385.11 and an aic of 776.22.



The predicted closing prices for the next 4 days would be 2028.499, 2028.497, 2028.496, 2028.494 with SEs 21.18401, 29.95819, 36.69050, 42.36580 while the actual closing prices are 2047.5, 2055.25, 2051.5. 2065.



The fitted series is in blue and the original series is in red. The two fitted AR models are almost identical with a slightly better predictive model being the AR(1) fitted by the maximum likelihood method.

MODEL 2 - ARIMA(0,1,0):

Upon differencing the data with lag 1, the plot of the differenced data seems like white noise, and the ACF and PACF of the data are consistent with those of white noise. Therefore, we fit an ARIMA(0,1,0) model to the data and obtained the following parameters: sigma^2 = 454.1, Log likelihood = -380.63 and aic = 763.27.

In order to assess how accurate this model is, model diagnostics were performed and the following results were obtained:

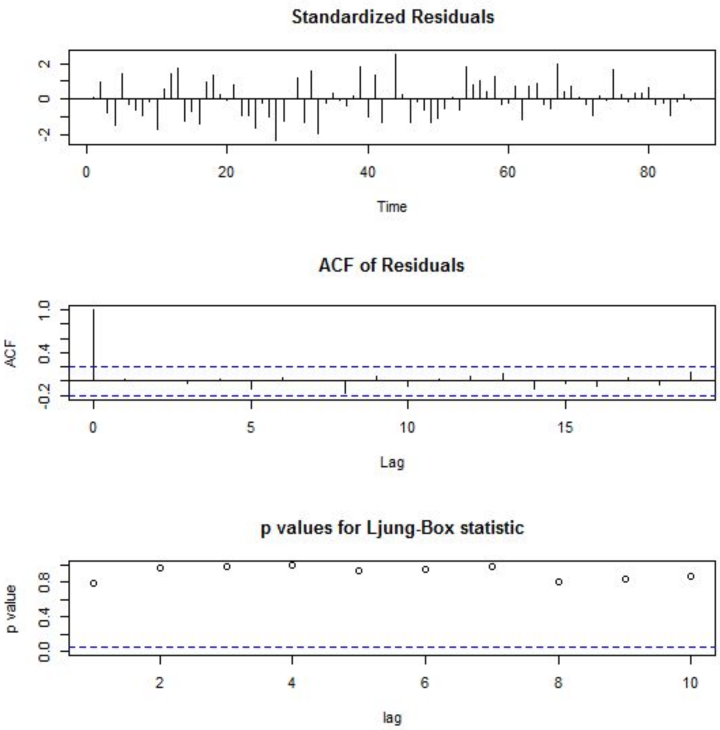
- The standardized residuals seem like white noise, and mostly lie between +-2
- The ACF of the residuals also follows the ACF of white noise

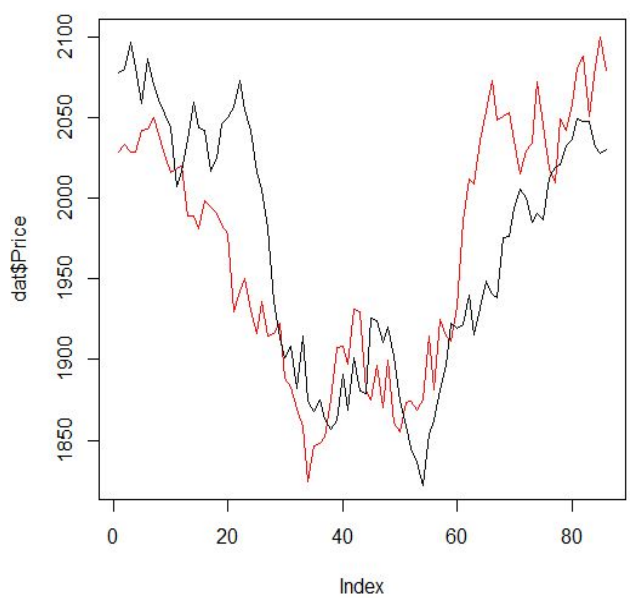
- The P-values, although significantly above 0, are not very high

The predicted closing prices for the next 4 days would be 2028.5, 2028.5, 2028.5, 2028.5 with SEs 21.30869, 30.13504, 36.90774, 42.61739 while the actual closing prices are 2047.5, 2055.25, 2051.5, 2065.

MODEL 3 – SARIMA(3,1,0) x (1,0,0)

The following SARIMA model is found by iteratively eliminating models with lower aic and/or insignificant coefficients. It appears that a SARIMA (3,1,0) x (1,0,0) model is appropriate for describing this time series since the ACF of Residuals shows that none of the sample autocorrelations from lag 1 to 20 exceed the significance bound and the p-values for the Ljung-Box statistic are also all well above the bound.



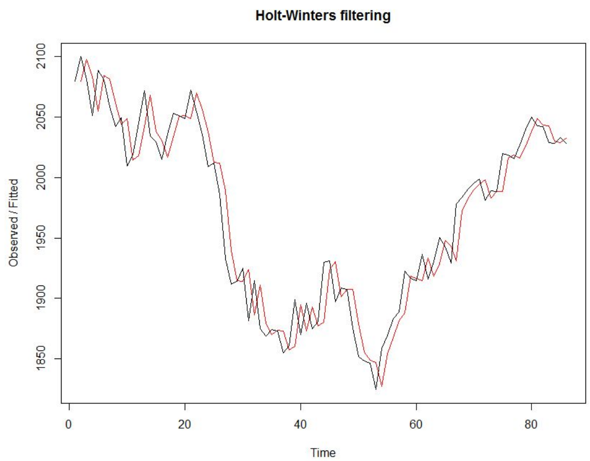


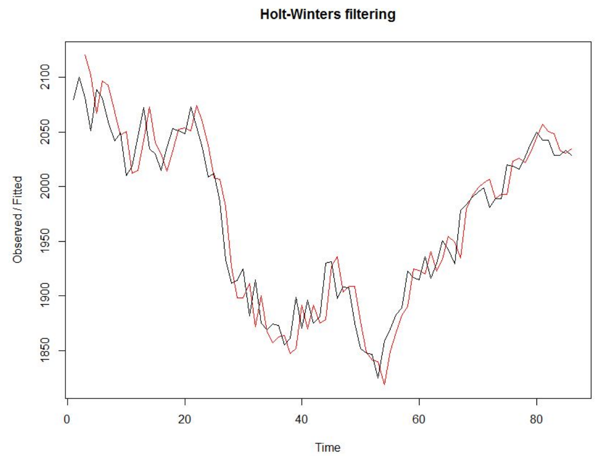

```
Coefficients:
         ar1      ar2     ar3     sar1
      0.4728  -0.3442  0.3653  -0.6077
s.e.  0.1441   0.1469  0.1196   0.1328

sigma^2 estimated as 415:  log likelihood = -376.96,  aic = 763.92
```

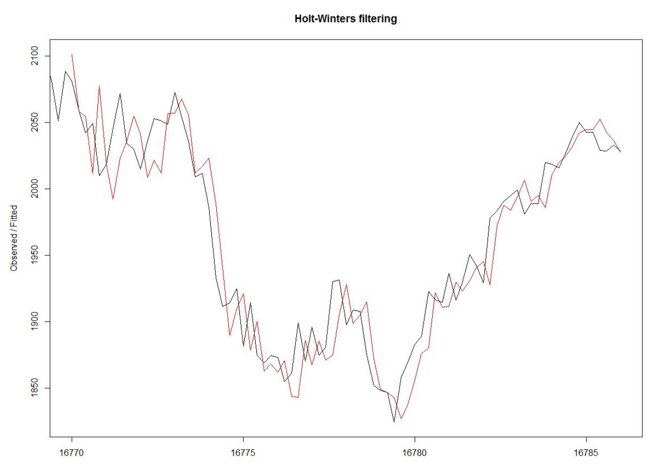
The predicted closing prices for the next 4 days would be 2025.779, 2027.103, 2027.392, 2025.798 with SEs 20.37271, 26.93776, 31.76584, 37.68692 while the actual closing prices are 2047.5, 2055.25, 2051.5, 2065.

MODEL 4 - Exponential, Holt and Holt-Winters Smoothing

To make predictions, we learnt exponential smoothing, double exponential smoothing (Holt's method) and Holt-Winters smoothing. Exponential smoothing is useful only for stationary series. An arima (0,1,0) model previously fitted for this time series suggests that it has some properties of a random walk process, which is non-stationary, so exponential smoothing might not be a good choice for forecasting. Instead, from the graph of original series and the fitted SARIMA model, we observe certain trend and seasonality pattern, which indicates that Holt's method and Holt-Winter smoothing are worth exploring.







The plots above are the smoothed series using exponential smoothing, Holt smoothing and Holt-Winters smoothing in clockwise order (the original series is in black, and the smoothed series is in red). The smoothed series all follow closely behind the original series.

The parameters for three smoothing methods are:

Exponential smoothing: alpha = 0.8807

Holt smoothing: alpha = 0.8557, beta = 0.1527

Holt-Winters smoothing: alpha = 0.701, beta = 0.0489, gamma = 0.65

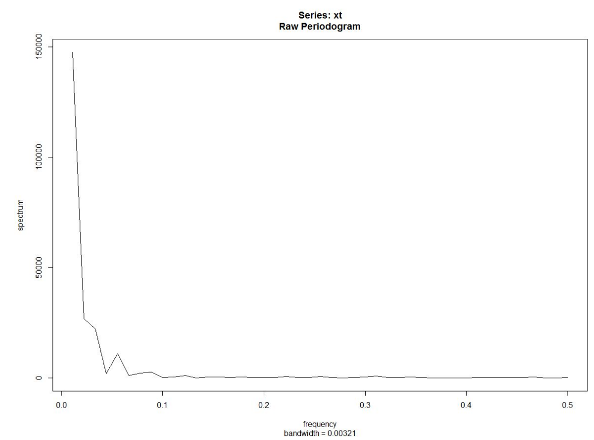
The predicted values for the four subsequent days using exponential smoothing are 2029.002 for all four days with SEs 42.56327, 56.71735, 67.98673, and 77.63673.

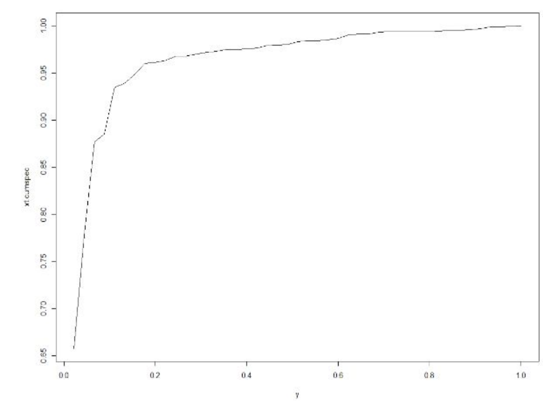
The predicted values for the next four days using Holt smoothing are 2030.247, 2031.125, 2032.003, and 2032.881, with SEs 45.16531, 63.43878, 81.05357, and 98.71837.

The predicted values for the next four days using Holt-Winters smoothing are 2028.127, 2031.852, 2039.724, and 2045.399 with SEs 50.15459, 62.25459, 73.25204, and 83.61786.

Comparing the predicted values with the actual prices 2047.5, 2055.25, 2051.5, 2065, we could see that in this case the Holt-Winters smoothing makes a better predictions in values. This Holt-Winters model was fit by using frequency=5 for the data, since the data was measured daily, and 5 observations were made per week.

FREQUENCY DOMAIN





The raw periodogram exhibits the dominance of frequency below 0.6 of this series. In particular,there are four frequencies contributing to the spectrum, among which the first one has the largest contribution. Reading the output when the variable xt.spec$spec is requested in R, we extracted the the four frequencies as 0.0116 (largest), 0.0232, 0.0349, and 0.0581 for spectrum density 1.475012e+05, 2.682687e+04, 2.224041e+04, and 1.106186e+04 in the same order. The corresponding wavelengths are 86, 43, 28.7, and 17.2.

Fitting the simple model $X(t) = a_0 + a_p\cos(w_p t) + b_p\sin(w_p t) + \varepsilon$ to our data, the same as what we dealt with in the second assignment, we found that this model does not work for all these significant frequencies, with coefficients almost equalling zero. So only one frequency at a time to construct the model is not sufficient to explain our series.

Now we perform the white noise test using the cumulative periodogram and Dc critical value at the 5% significance level which equals 0.21 in this case.

The cumulative spectrum clearly does not resemble the line y=x, which suggests that the series does not follow a white noise process. The test statistic verifies the conjecture, with a significant value 0.82 greater than the critical value 0.21.

Conclusion

Broadly looking at all the models discussed above, it was found that in Model 1, the MSE (mean-squared error) of the AR(1) model by “CCS” between the observed future values and the predicted values for the data points between March 29, 2016 - April 2, 2016 was 1308.048. Comparatively, the MSE between the 4 observed and predicted values for the AR(1) model by “ML” was 734.6583. The actual values for that time period were within the 95% confidence interval (CI) of the predicted values, but the p-values for both these fitted models were not very high.

In comparison, in Model 2, the ARIMA(0,1,0) model, the four observed data points also fell within the 95% CI of the predicted values, although the p-values for the model weren’t very high. The MSE between the observed future values and the predicted values for those 4 data points was found to be 734.4531.

In Model 3, the SARIMA(3,1,0)x(1,0,0) model, the MSE was found to be 845.512. The observed values for the 4 data points fell within the 95% CI of the predicted values and the p-values for the model were very high.

In Model 4, the MSE for the 4 data points with exponential smoothing was found to be 708.2874, the MSE for the 4 data points with Holt’s smoothing was found to be 572.8612 and the MSE using Holt-Winters was 361.4132. The actual values for the 4 data points fell inside the 95% CI of the predicted values.

Overall, it seems that Model 4, the one found using Holt-Winters smoothing, as the best fit for the data since the predictions made from it had the smallest MSE (361.4132). Moreover, all of the predicted values found with the models above were similar to the actual observed values. We can then conclude that we can predict future values of the S&P 500 futures using a relatively small sample of data (around 100 data points) within a range of accuracy, with a Holt-Winters smoothed model being the most accurate predictor with our data.

Analysis in the frequency domain shows that the prime wavelength is exactly the same length of our observations and the series has more than one featuring frequency. From the cumulative periodogram, the series is suggested not to be a realization of white noise process.

Appendix

```
data <- read.table("sp500futures.csv", header=T, sep=",")
x <- ts(rev(data$Price), frequency = 5, start = as.Date("2015-11-30"))
library(hydroGOF)
# Plot the original time series, acf, pacf
plot(x)
acf(x)
pacf(x)
```

Stats 443, Group Project
Priyanka Sethi (53098133) Nicholas Ho (53693123) Jiaxuan Chen (85250158)

Model 1:

```
# Fitting an ARIMA model with method "CSS"
ArimaModel2 <- arima(x, order = c(1,0,0), method = "CSS")
predict(ArimaModel2 , n.ahead = 4)
mse(c(2024.919, 2021.514, 2018.276, 2015.197), c(2047.5, 2055.25, 2051.5, 2065))
# Fitting an ARIMA model with method "ML"
ArimaModel3 <- arima(x, order = c(1,0,0), method = "ML")
predict(ArimaModel3 , n.ahead = 4)
library(forecast)
# Plot fitted model against the original series
plot(data$Price,col="red", type="l")
lines(fitted(ArimaModel1),col="blue")
# MSE for the model
mse(c(2028.499, 2028.497, 2028.496, 2028.494), c(2047.5, 2055.25, 2051.5, 2065))
```

Model 2:

```
#find and analyze the differenced time series
y <- diff(x)
acf(y); pacf(y)
#fit, observe and analyze the ARIMA model
m2 <- arima(x, order=c(0,1,0))
m2
tsdiag(m2)
#predict the next 4 values
predict(m2,4)
```

Model 3:

```
# SARIMA model fitting
sarima1 <- arima(x, order = c(3, 1, 0), seasonal = list(order = c(1, 0, 0)))
tsdiag(sarima1)
# Plot fitted model against the original series
plot(data$Price,col="red", type="l")
lines(fitted(sarima1 ),col="blue")
predict(sarima1 , n.ahead = 4)
# MSE for the model
mse(c(2025.779, 2027.103, 2027.392, 2025.798), c(2047.5, 2055.25, 2051.5, 2065))
```

Model 4:

```
# exponential smoothing
x.exp <- HoltWinters(x, beta=FALSE, gamma=FALSE)
plot(x.exp)
forecast(x.exp, n.ahead=4)
```

Stats 443, Group Project
Priyanka Sethi (53098133) Nicholas Ho (53693123) Jiaxuan Chen (85250158)

```
# MSE for the model
mse(c(2029.002, 2029.002, 2029.002, 2029.002), c(2047.5, 2055.25, 2051.5, 2065))

# Holt smoothing
x.holt <- HoltWinters(xt,gamma=FALSE)
plot(x.holt)
forecast(x.holt, n.ahead=4)
# MSE for the model
mse(c(2030.247, 2031.125, 2032.003, 2032.881), c(2047.5, 2055.25, 2051.5, 2065))

# Holt-Winters smoothing
x.hw <- HoltWinters(x, seasonal="additive")
plot(x.hw)
forecast(x.hw, n.ahead=4)
# MSE for the model
mse(c(2028.127, 2031.852, 2039.724, 2045.399), c(2047.5, 2055.25, 2051.5, 2065))
```

<u>Frequency domain:</u>

```
# looking for the most important frequencies
x.spec <- spec.pgram(xt, log="no")
x.spec$spec # reading the output for the biggest spectrum densities
dom_f1 <- 1/86; dom_f2 <- 2/86; dom_f3 <- 3/86; dom_f4 <- 5/86
wavelength1 <- 1/dom_f1; wavelength2 <- 1/dom_f2; wavelength3 <- 1/dom_f3; wavelength4 <- 1/dom_f4

# fitting a simple model for the most contributing frequency 1/86; other frequencies use similar commands
p1 <- 86 / wavelength1 # p is the number of cycles completed over the duration of the data
omega_p1 = (2 * pi * p1) # given by the formula (2*pi)/omega_p = N/p
t <- 1:86
cs1 <- cos(omega_p1 * t)
sn1 <- sin(omega_p1 * t)
lma1 <- lm(xt ~ cs1 + sn1)
summary(lma1)
plot(xt)
lines(lma1$coefficients[1] + lma1$coefficients[2]*cs1 + lma1$coefficients[3]*sn1, col='red', lty = 2)

# white noise test
x.cumspec <- cumsum(x.spec$spec)/sum(x.spec$spec)
y <- x.spec$freq
plot(y, x.cumspec, type="l")
D <- max(abs(x.cumspec - 1:45/45)) # 0.8233388
D_c <- 1.358 / sqrt(43 - 1) # 0.2095439
```

Stats 443, Group Project
Priyanka Sethi (53098133) Nicholas Ho (53693123) Jiaxuan Chen (85250158)