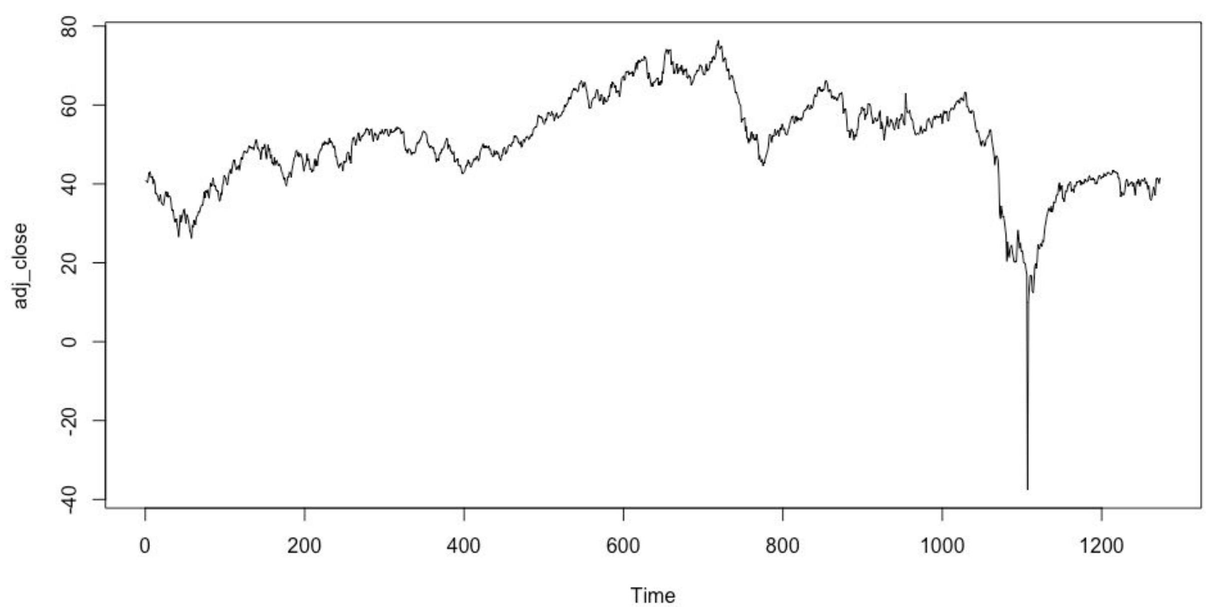
MATH 4070 Final Project

Group 9: Nicholas Arledge, Casey Hird, Carlos Kelaidis

# Time-Series Analysis of Crude Oil Prices

## Introduction

In order to analyze price data for crude oil, we evaluated a time series of adjusted close prices. These values account for corporate actions, and are thus considered the most accurate representation of historical returns. We collected a 5 year time series of adjusted close prices of crude oil (CL=F) from Yahoo Finance (https://finance.yahoo.com/quote/CL=F?p=CL=F&.tsrc=fin-srch). The time series plot of this data can be seen below:

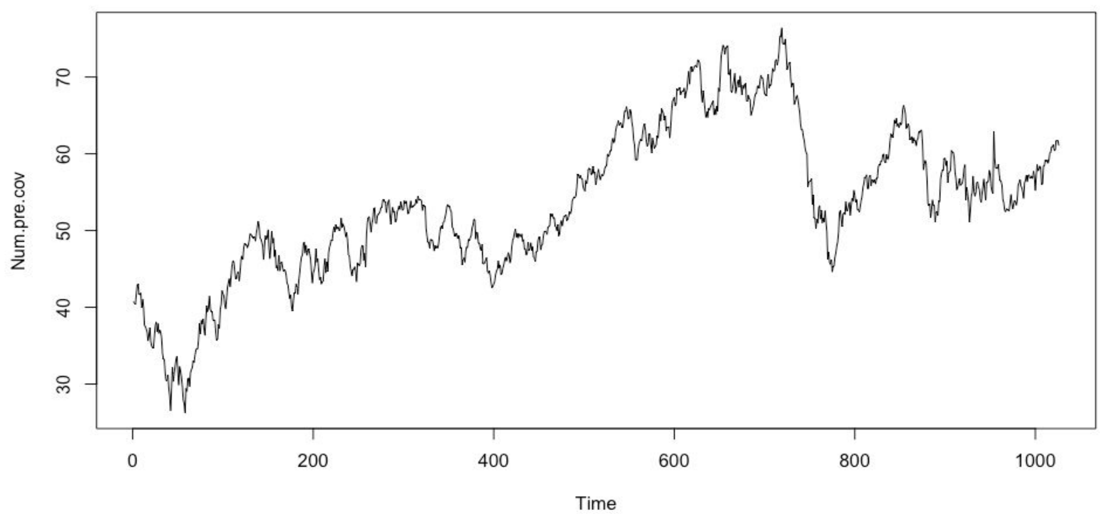


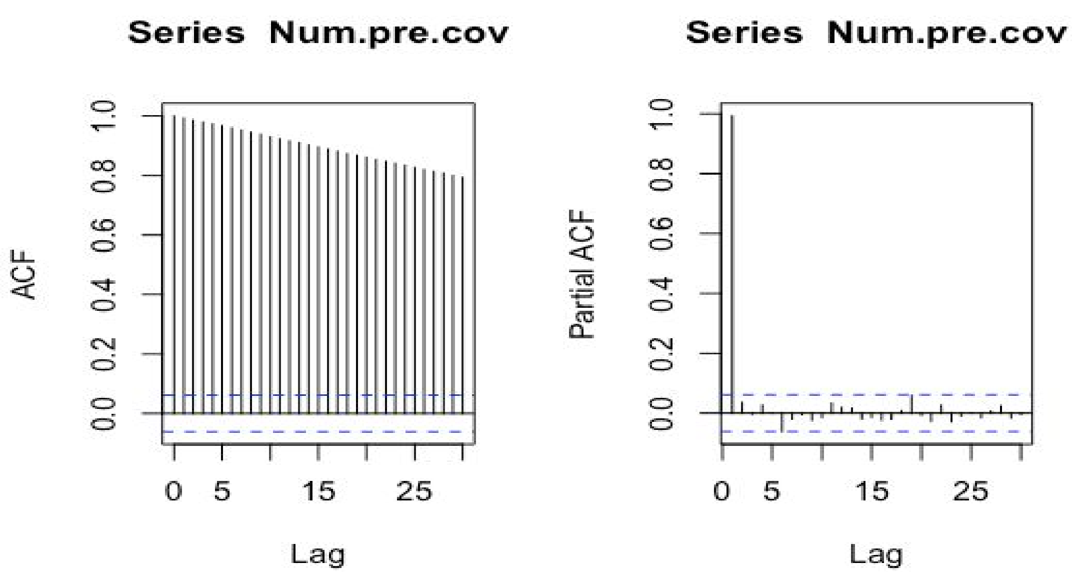

Note from the above time series that there is an abnormal drop and recovery coinciding with the economic effects of COVID-19. Since including this stretch of values would disrupt any model we might fit, we chose to split this data into two time periods: pre-COVID data from 11/18/2015 to 12/31/19 and post-COVID data from 6/1/2020 to 11/17/2020.

Note: In model consideration, if the coefficient for a parameter is less than two times the absolute value of the standard error for that parameter, the parameter is considered insignificant and dropped from the model. Outputs given in this report are after insignificant parameters have been dropped.

## Pre-Covid

The pre-covid data started on November 18 2015 and finished on December 31st of 2019. We started off by plotting a time series plot of the pre-covid data and its associated ACF and PACF to observe its behavior.

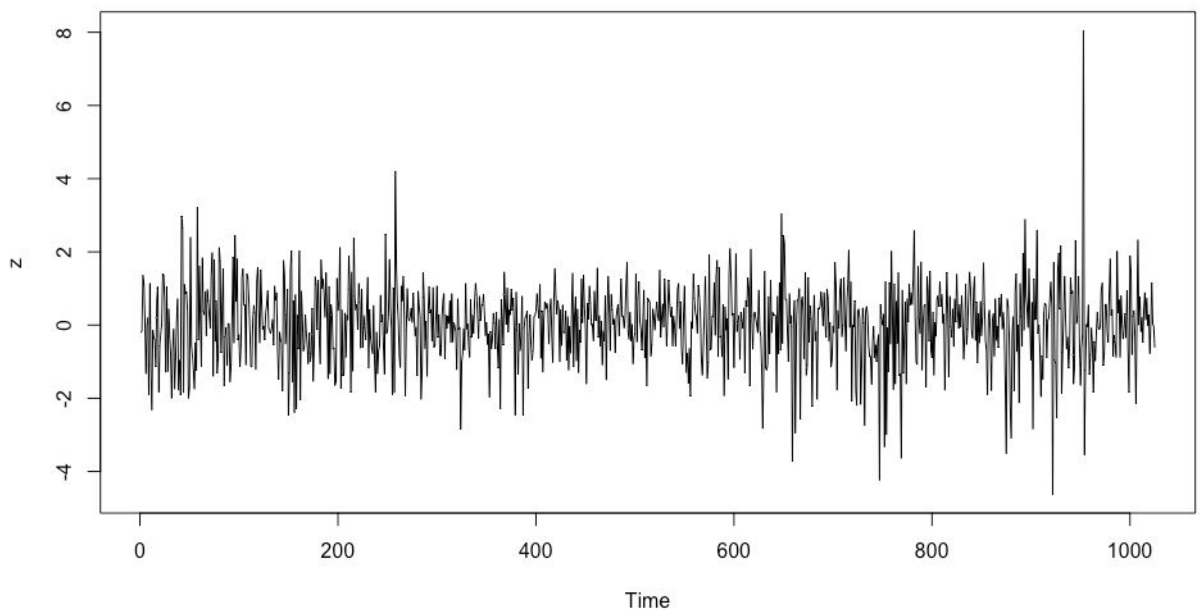


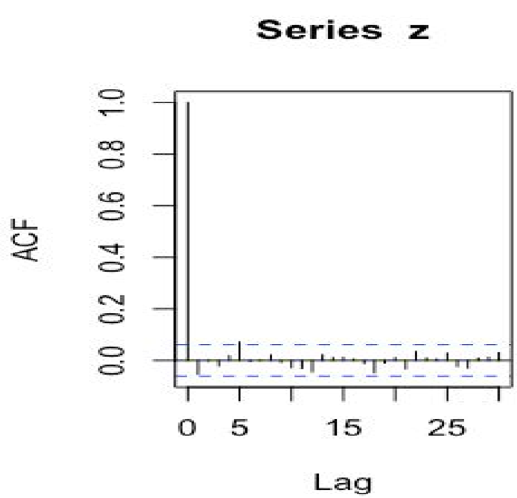



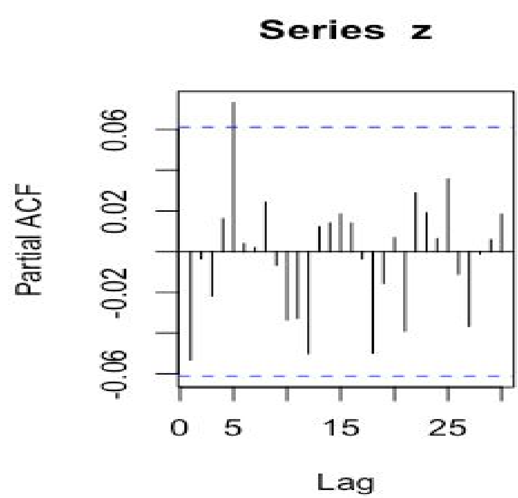

The time series plot of the data has an increasing trend indicating non-stationarity. Further, we can observe a gradual constant decrease over time in the ACF indicating that our data has non-constant variance and non-stationarity.

In order to stabilize the variance and make our data stationary we decided to take the first differences of the data. We did not observe any seasonal spikes in the PACF that could indicate a difference at a particular lag.

Below is the time series plot, ACF and PACF of our first differences:




The time series plot exhibits no trend anymore, further the ACF and PACF die down quickly indicating that we have made our data stationary and stabilized the variance.

Now looking at our PACF, we see a slight spike at lag 5 with the remaining lags inside the bounds. From this we tried a pure AR(5).
Noticing that the ACF also had a slight spike at lag 5, we tried a pure MA(5). Further we considered a possible seasonal trend at lag 5, since both the ACF and PACF are outside the bounds there we tried a model with just a seasonal MA at lag 5, a model with just a seasonal AR at lag 5 and a model with a seasonal MA and a seasonal AR at lag 5.
Finally we tried an AR(1) coupled with a seasonal AR at lag 5.

Hence our competing models were the following:

- Pure AR(5)
- Pure MA(5)
- Seasonal MA at lag 5
- Seasonal AR at lag 5
- Seasonal AR & MA at lag 5
- AR(1) with seasonal AR at lag 5

Below are the summaries of each of our prospective models along with their respective Box Test:

AR(5): Note that some parameters have been set to zero due to non-significance

```
Call:
arima(x = Num.pre.cov, order = c(5, 1, 0), fixed = c(0, 0, 0, 0, NA))
Coefficients:
          ar1  ar2  ar3  ar4       ar5
          0    0    0    0    0.0715
s.e.      0    0    0    0    0.0311
sigma^2 estimated as 1.174:  log likelihood = -1536.8,  aic = 3077.59


Box-Ljung test
data:  pre_ar5$residuals
X-squared = 16.309, df = 20, p-value = 0.6973

BIC = 3087.456
```

MA(5): Note that some parameters have been set to zero due to non-significance

```
Call:
arima(x = Num.pre.cov, order = c(0, 1, 5), fixed = c(0, 0, 0, 0, NA))

Coefficients:
      ma1  ma2  ma3  ma4     ma5
        0    0    0    0  0.0763
s.e.    0    0    0    0  0.0321
sigma^2 estimated as 1.174:  log likelihood = -1536.62,  aic =
3077.24


Box-Ljung test
data:  pre_ma5$residuals
X-squared = 16.088, df = 20, p-value = 0.7112

BIC = 3087.11
```

Seasonal MA at lag 5:

```
Call:
arima(x = Num.pre.cov, order = c(0, 1, 0), seasonal = list(order =
c(0, 0, 1), period = 5))
Coefficients:
        sma1
      0.0763s.e.  0.0321

sigma^2 estimated as 1.174:  log likelihood = -1536.62,  aic =
3077.24


Box-Ljung test
data:  pre_sma5$residuals
X-squared = 16.088, df = 24, p-value = 0.8849

BIC = 3087.11
```

Seasonal AR at lag 5:

```
Call:
arima(x = Num.pre.cov, order = c(0, 1, 0), seasonal = list(order =
c(1, 0, 0), period = 5))
Coefficients:
        sar1
      0.0715
s.e.  0.0311
sigma^2 estimated as 1.174:  log likelihood = -1536.8,  aic = 3077.59


Box-Ljung test
data:  pre_sar5$residuals
X-squared = 16.309, df = 24, p-value = 0.8766

BIC = 3087.456
```

Seasonal AR & MA at lag 5:

```
Call:
arima(x = Num.pre.cov, order = c(0, 1, 0), seasonal = list(order =
c(1, 0, 1), period = 5))
Coefficients:
         sar1    sma1
      -0.9455  0.9754
s.e.   0.0732  0.0547
sigma^2 estimated as 1.171:  log likelihood = -1535.51,  aic =
3077.02


Box-Ljung test
data:  pre_sarma5$residuals
X-squared = 17.289, df = 23, p-value = 0.7948

BIC = 3091.814
```

AR(1) with seasonal AR at lag 5:

```
Call:
arima(x = Num.pre.cov, order = c(1, 1, 0), seasonal = list(order =
c(1, 0, 0), period = 5))
Coefficients:
          ar1    sar1
      -0.0545  0.0728
s.e.   0.0312  0.0311
sigma^2 estimated as 1.171:  log likelihood = -1535.27,  aic =
3076.54


Box-Ljung test
data:  pre_ar1sar5$residuals
X-squared = 13.806, df = 23, p-value = 0.9323

BIC = 3091.334
```

All of our models passed the Box Test and hence removed correlation adequately up to lag 25. We now turn to the AIC, BIC, sigma^2, and parsimony in order to pick the best model.

Being that AIC's and BIC's differed by just a small amount, we decided to start by sticking to the simpler models. That is we excluded models with more than one parameter namely the AR(1) with a seasonal AR at lag 5 and the model with a seasonal AR and seasonal MA at lag 5. These two models had a lower AIC than the other models but a higher BIC, so it came down to excluding them based on the simplicity of the model.
Within those that were left, the seasonal AR at lag 5 and the AR(5) had a slightly higher AIC and BIC than the seasonal MA at lag 5 and the MA(5) so we excluded these two as well.
The seasonal MA at lag 5 and the MA(5) had the same AIC, BIC and sigma^2 so we were left to pick between these two and decided to pick the seasonal MA at lag 5.

The final model used for predictions for the pre-covid data is given on the following page.

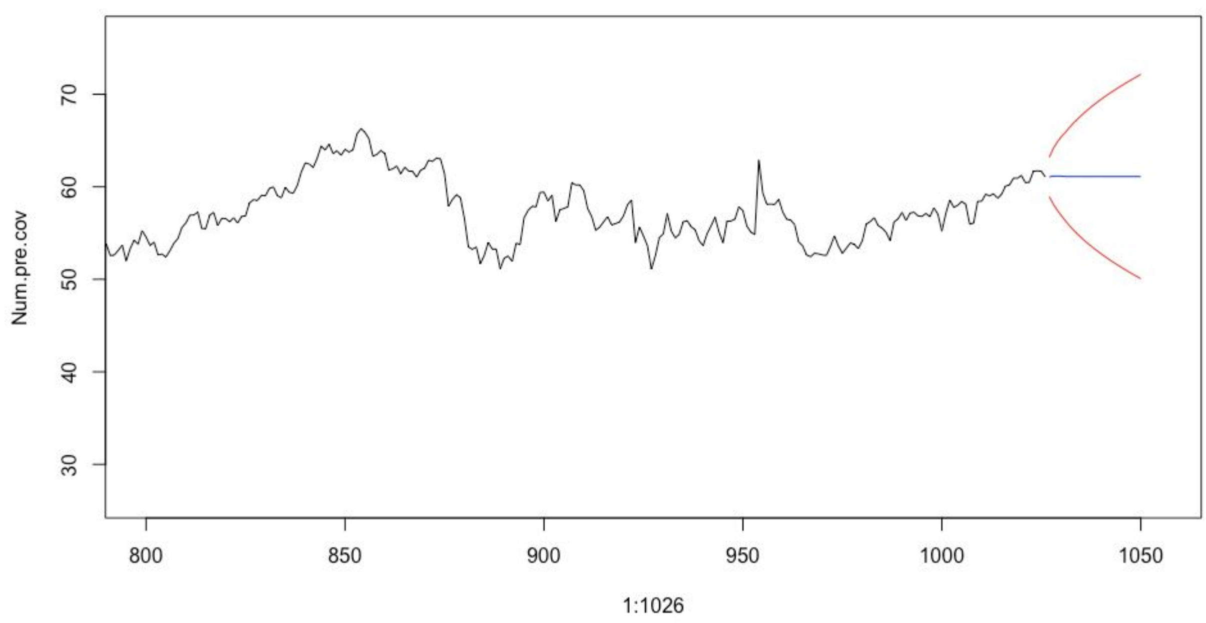
Seasonal MA at lag 5:

```
Call:
arima(x = Num.pre.cov, order = c(0, 1, 0), seasonal = list(order =
c(0, 0, 1), period = 5))
Coefficients:
        sma1
      0.0763
s.e.  0.0321
sigma^2 estimated as 1.174:  log likelihood = -1536.62,  aic =
3077.24
```

```
Box-Ljung test
data:  pre_sma5$residuals
X-squared = 16.088, df = 24, p-value = 0.8849

BIC = 3087.11
```

Here are the predictions for the next 24 observations along with a plot showing the 95% confidence interval of our predictions:

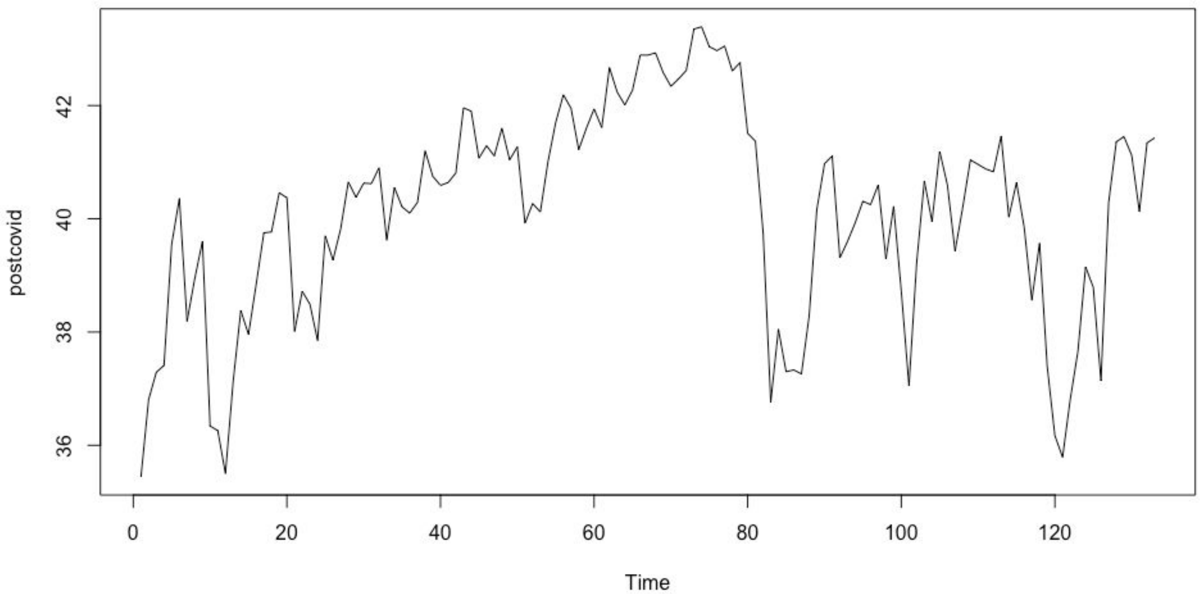



The next five predictions are:

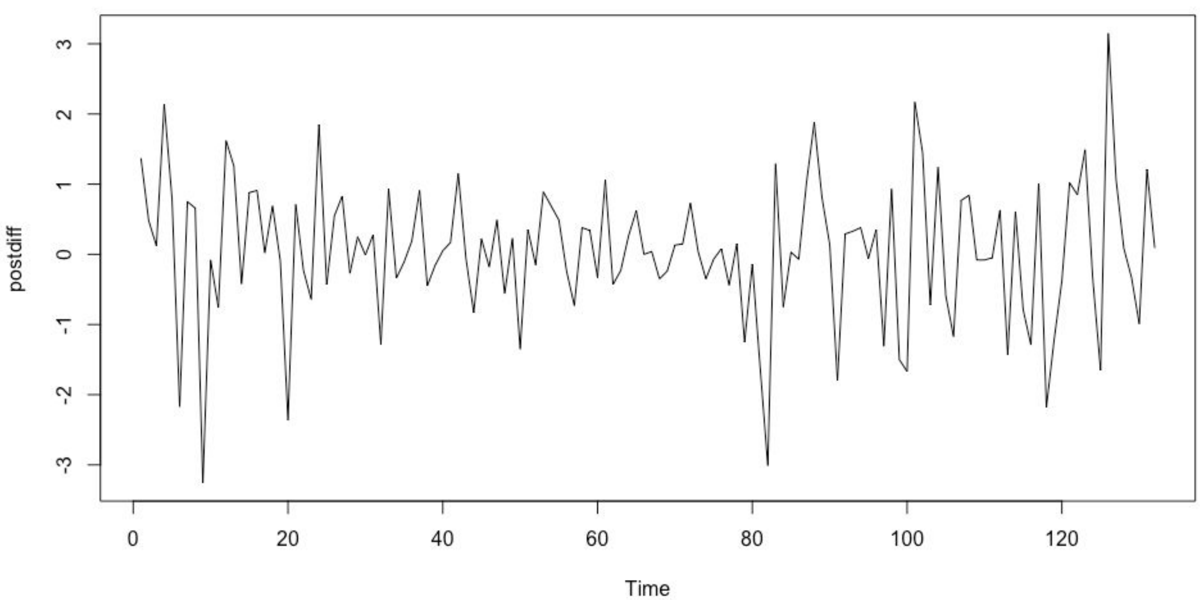
| | |
|---|---|
| 61.06528 | t = 1027 |
| 61.14958 | t = 2028 |
| 61.15240 | t = 2029 |
| 61.14785 | t = 2030 |
| 61.10548 | t = 2031 |

## Post-Covid

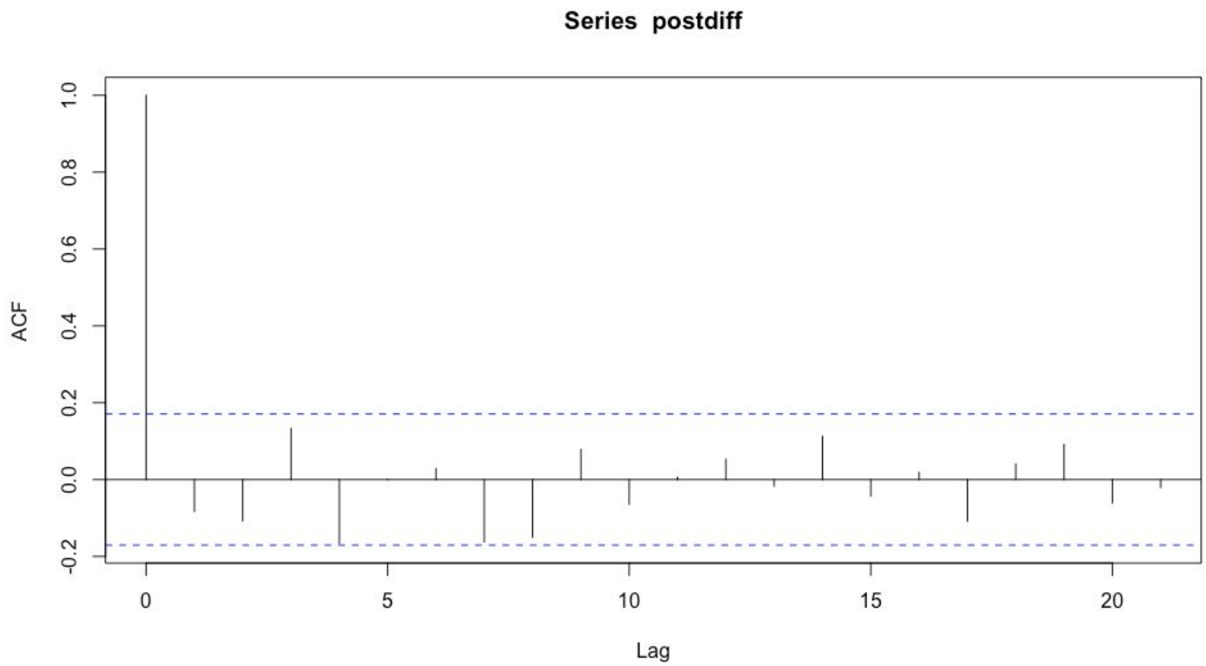
For analyzing the post-covid data, we continued to use the first differences. The time-series for the original post-covid data is shown below.

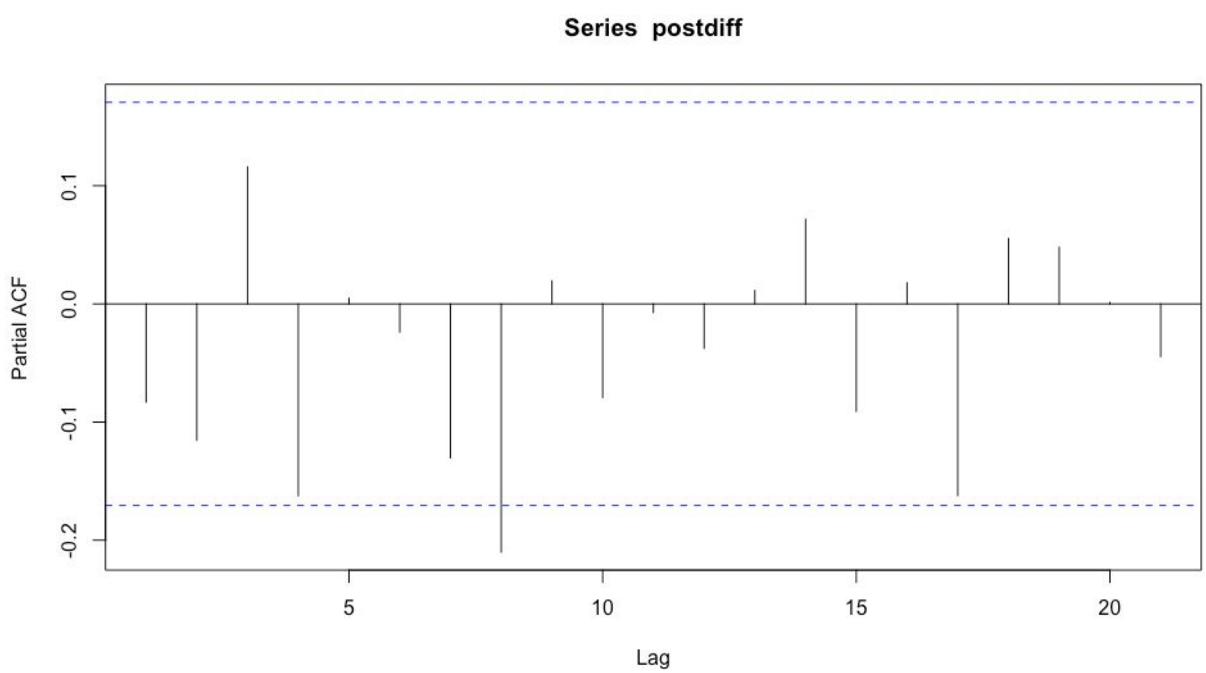



Here we see non-constant variance and non-constant mean. Taking first differences yields the following time series plot:



We see this time-series has a constant mean and constant variance, so it is safe to assume our data is now stationary.
Plotting the ACF and PACF of the first differenced post-covid data yields the following:




In the ACF plot, we see the values are inside the bounds for all lags. This suggests a white noise model for our first differences. We also see values growing larger around lags at multiples of 4 in the ACF. This suggests a seasonal MA model at lag 4. In the PACF plot, we see the values are inside the bounds for all lags except at lag 8. This suggests two models: an AR(8) or a seasonal AR at lag 8.

Thus, the four competing models are as follows:

1. White Noise
2. Seasonal MA at lag 4
3. AR(8)
4. Seasonal AR at lag 8

The fits and a discussion for each competing model is given below.

White Noise:

```
Call:
arima(x = postcovid, order = c(0, 1, 0))
sigma^2 estimated as 1.006:  log likelihood = -187.7,  aic = 377.39

Box-Ljung test
data:  z$residuals
X-squared = 17.433, df = 10, p-value = 0.06532

BIC: 380.2777
```

Seasonal MA at lag 4

```
Call:
arima(x = postcovid, order = c(0, 1, 0), seasonal = list(period = 4,
order = c(0, 0, 1)))
Coefficients:
      Sma1
       -0.2606
s.e.   0.1042
sigma^2 estimated as 0.9607:  log likelihood = -184.79,  aic = 373.59

Box-Ljung test
data:  postma1s4$residuals
X-squared = 10.674, df = 8, p-value = 0.2209

BIC: 379.3532
```

AR(8) note: not all parameters were statistically significant; such have been dropped

```
Call:
arima(x = postcovid, order = c(8, 1, 0), fixed = c(0, 0, 0, NA, 0, 0,
0, NA))
Coefficients:
     ar1  ar2  ar3   ar4      ar5  ar6  ar7    ar8
     0    0    0    -0.2128   0    0    0     -0.2222
s.e. 0    0    0     0.0877   0    0    0      0.0930
sigma^2 estimated as 0.9336:  log likelihood = -183.03,  aic = 372.06

Box-Ljung test
data:  postar8$residuals
X-squared = 9.5947, df = 2, p-value = 0.008252
```

Seasonal AR at lag 8

```
Call:
arima(x = postcovid, order = c(0, 1, 0), seasonal = list(period = 8,
order = c(1, 0, 0)))
Coefficients:
     sar1
      -0.1786
s.e.   0.0934
sigma^2 estimated as 0.9772:  log likelihood = -185.91,  aic = 375.81

Box-Ljung test
data:  postma1s8$residuals
X-squared = 17.584, df = 8, p-value = 0.02457
```

From the Box test on each model, we see the two AR models have a p-value of less than .05, so they are not adequately removing the correlation, and will not be considered. The remaining two models -- white noise and seasonal MA at lag 4 -- are both adequate at removing the correlation. Thus, we turn to AIC and BIC for model selection between these two. The AIC and BIC are both lower for the seasonal MA at lag 4 model, so the seasonal parameter is justified to add, and this model is preferred over the white noise model.
The final model for the post-COVID data is given on the following page and used for predictions.

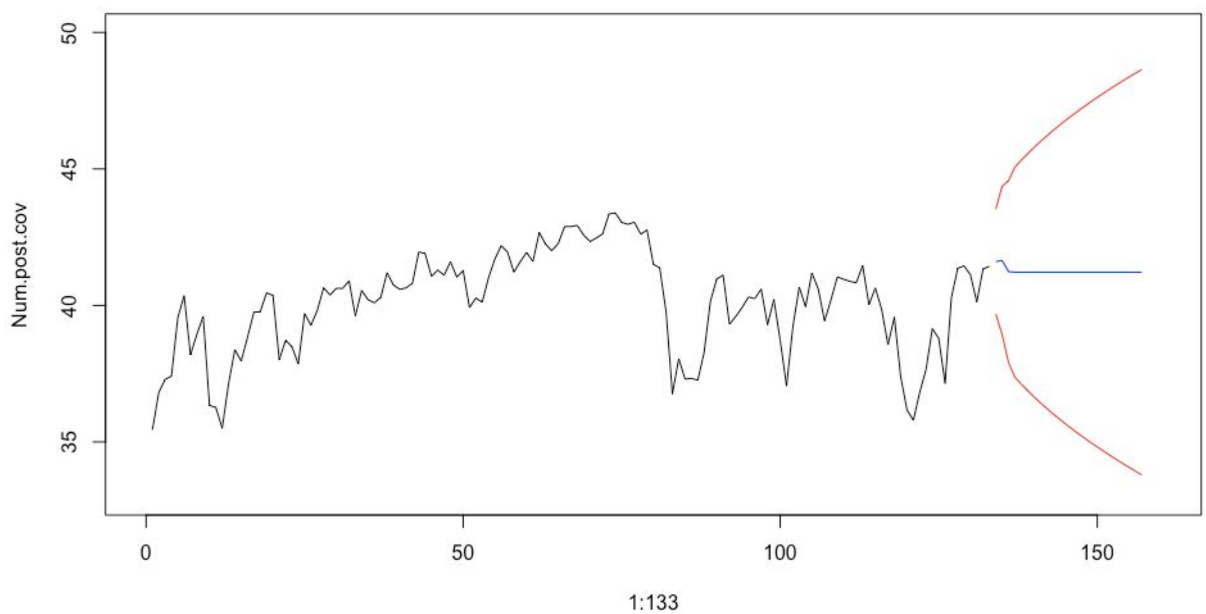
Seasonal MA at lag 4:

```
Call:
arima(x = postcovid, order = c(0, 1, 0), seasonal = list(period = 4,
order = c(0, 0, 1)))
Coefficients:
       Sma1
      -0.2606
s.e.   0.1042
sigma^2 estimated as 0.9607:  log likelihood = -184.79,  aic = 373.59

Box-Ljung test
data:  postma1s4$residuals
X-squared = 10.674, df = 8, p-value = 0.2209

BIC: 379.3532
```

Below are the predictions for the next 24 observations along with a plot showing the 95% confidence interval of our predictions:




The next five predictions are:

| | |
|---|---|
| 41.60709 | t = 134 |
| 41.64552 | t = 135 |
| 41.23774 | t = 136 |
| 41.21756 | t = 137 |
| 41.21756 | t = 138 |

## Conclusion

Note that in our analysis of the pre-Covid and post-Covid data we selected two different models. From this we can conclude that either the movement of oil prices somehow changed after Covid, or more likely the prices are fairly random since both data sets seem to be near a white noise model. Also, neither of these models can adequately capture the economic impact of the Covid-19 pandemic, so there are clearly some movements in the prices of crude oil that we cannot fully explain with the current models.