# The University of Hong Kong

Academic Year 2020-2021 Semester Two

5 May 2021

STAT4601 Time Series Analysis

Project Report

**Title: The Statistical Analysis of the Brent Crude Oil Futures Prices daily data**

GUN Chun Ho
3035466320

Declaration for confirming the work has not been used for another project and relevant references are cited properly without committing plagiarism.

**Table of contents**


**Background** 3

**Test for stationarity** 4-7

- Time plot 4
- Identifying outliers 5
- Check for Stationarity 6
- Common differencing 7

**Specified models/model selections** 8-16

- Sample Auto-correction Function (ACF) and Sample Partial ACF (PACF) 8-10
- Akaike's information criterion (AIC) and Maximum Likelihood Estimation (MLE) 10-11
- Estimate of parameters of the fitted models 11-12
- Ljung Box Test 12-13
- Test for Normality 14
- Overparameterized method for the adequacy of the fitted models 15-16

**Forecasting** 16-17

**Forecasting Amendment** 17-18

**Reflection** 18

**Conclusion** 19

**Acknowledgement** 20

**References** 21

# 1. Background

The commodity price is extremely important to traders, producers or even a country as petroleum is very useful. It can serve as a fuel for aeroplanes, cars and ships. If we can successfully forecast the price of the oil futures. It is very useful for producers to earn money, traders to have the chance for arbitrage and risk management. The related oil company, oil production company and country's economy are seriously affected by the oil price. We have always heard of the 'Brent Crude oil price' from the media. 'Brent Crude oil price' is usually referring to Brent Crude Oil Futures price which this project will use. Since the general public cannot participate in the spot market, this project will mainly focus on the futures price. Brent Crude oil futures price is a benchmark for the oil trading market since ⅔ of the oil price in the world is referring to the Brent Crude oil price. There are two major oil markets, Brent Crude is one of them and it mainly targeted the european and African market. Compared to West Texas Intermediate (WTI) futures price, Brent Crude is usually a bit lower as the quality Brent Crude oil is a bit worse than the oil in WTI.

This data is downloaded from the Bloomberg terminal. It has the detailed daily price of the Brent Oil Futures data from 01 April 2011 to 31 March 2021 with the total sample size of 2580. Each futures contract refers to 1000 barrels. The following analysis is conducted by R. The detail of the code is archived in the attached RAR file.

# 2. Test for stationarity

## 2.1 Time Plot

The figure 1 below visualises the data in a time plot. The x-axis represents the time starting from each trading date in 01/04/2011 to that in 31/03/2021, and the y-axis represents the oil price in US dollars.

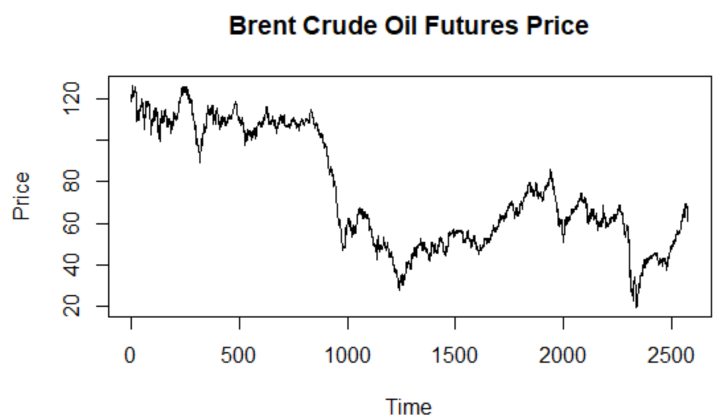



Figure 1

As we can see from Figure 1, there is no trend and cycle of the time series. However, we can see the left part of the time plot is higher than the right part. It may look right skewed. The variance may not be stationary. I have done the log transformation. Figure 2 below is the time plot after the log transform.

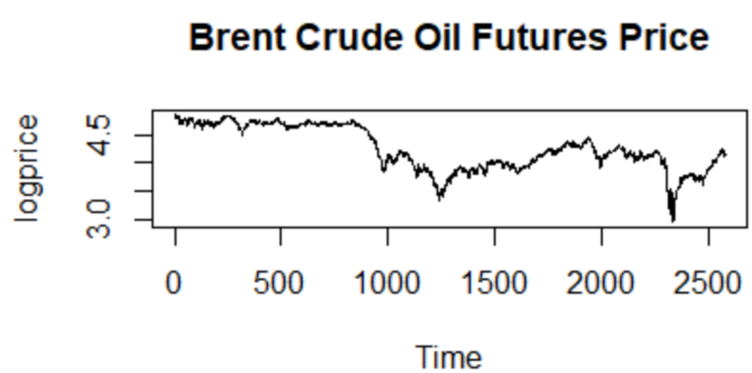



Figure 2

We do not see a huge difference between the logged time plot and the normal time plot in terms of trend or cycle. Therefore, we can reach the primary conclusion that the variance is non stationary at this point.

## 2.2 Identifying outliers

Although the price fluctuates a lot, there is no outlier for this time series data. The following Figure 2 recorded the details of the summary statistics of the data.

```
> summary(data)
   ï..Date              Price       
 Length:2580        Min.   : 19.33  
 Class :character   1st Qu.: 51.84  
 Mode  :character   Median : 65.50  
                    Mean   : 75.25  
                    3rd Qu.:107.42  
                    Max.   :126.65  
```

Figure 3

If we calculate the outliers by the IQR criterion $I = [Q1 - 1.5 * IQR, Q3 + 1.5 * IQR]$, we should get the range $[-31.53, 190.79]$. The following box-plot (figure 3) also shows that the data does not contain outliers.

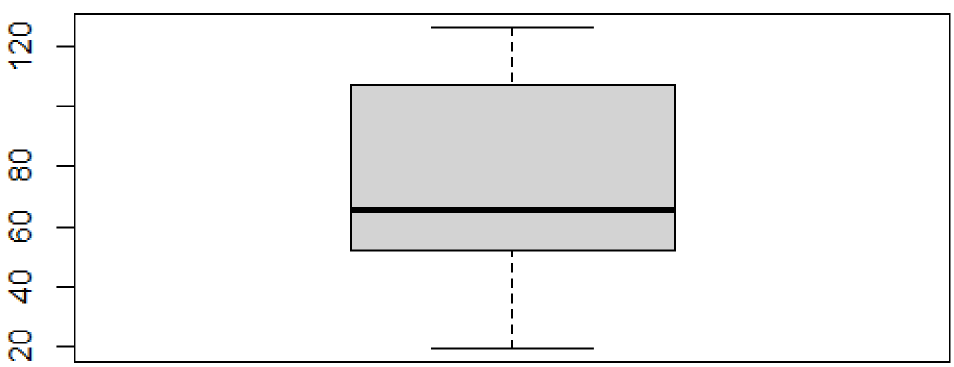



Figure 4

## 2.3 Check for Stationarity

The time series data split into two sets, the training dataset and the testing dataset. The testing dataset is the last 5 values of the data which is for testing the accuracy of the forecasting. The training dataset is to determine which model to use for this dataset.

```
train<-data[1:(nrow(data)-5),]
test<-data[(nrow(data)-4):nrow(data),]
```

Figure 5

For the stationarity test, there are a few ways to check the stationarity of the time series data. The sample Autocorrelation function (ACF) plot and the sample partial autocorrelation function (PACF) plot and the Argumented Dicky Fuller (ADF) test can help us to determine whether the time series data is stationary. The ADF test is listed below.

```
> adf.test(train)

        Augmented Dickey-Fuller Test

data:  train
Dickey-Fuller = -1.7871, Lag order = 13, p-value = 0.6685
alternative hypothesis: stationary
```

Figure 6

As we can see from the figure 5, the result of the Argumented Dicky-Fuller test shows us the p-value is 0.6685 which is larger than the 0.05 95% confidence. Therefore we cannot reject the null hypothesis.

Moreover, I have plotted the ACF. As we can see the ACF plot here. The ACF decreases slowly. It shows that the time series is non-stationary.

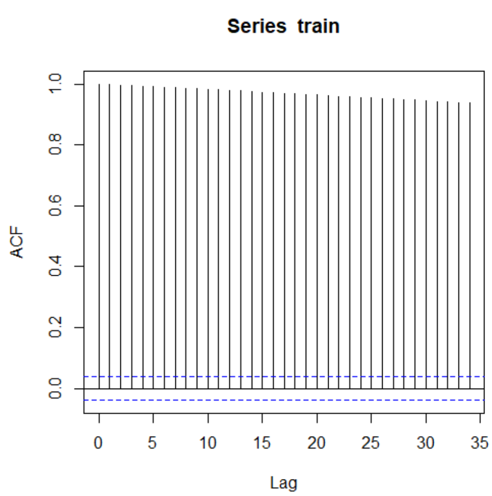
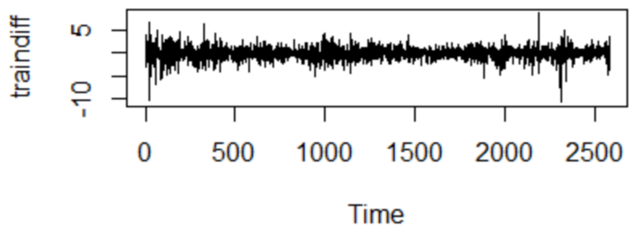



Figure 7

The time series data is non-stationary. Hence, we need to conduct the differencing transform the data into stationary and find out the number of d in the ARIMA (p,d,q) model.

## 2.4 Common differencing




```
> traindiff<-diff(train, differences=1)
> plot.ts(traindiff)
```

Figure 8

According to Figure 6, the time plot after the common differencing for 1 time. The data looks more stationary. We will conduct the ADF test again to check the stationarity of it.

```
> adf.test(traindiff)

        Augmented Dickey-Fuller Test

data:  traindiff
Dickey-Fuller = -13.264, Lag order = 13, p-value =
0.01
alternative hypothesis: stationary
```

Figure 9

The result of the Argumented Dicky-Fuller test after the differencing one time shows us that the p-value is smaller than 0.05. Therefore, we can reject the null hypothesis. The time series data after one time differencing is now stationary. Therefore, we can conclude the d is equal to 1 in the ARIMA (p,d,q) model. Since we need to have the stationary time series data to do forecasting correctly, we need to make the data to be stationary. Because stationary means the properties like mean, variance and autocovariance function do not change over time. Many useful analytical tools and statistical tests and models rely on the properties of stationarity. Since all the models we learnt (including ARIMA) are defined for stationary time series, we need to convert the time series into a stationary time series.

# 3 Specified models/model selections

## 3.1 Sample Auto-correction Function (ACF) and Sample Partial ACF (PACF)

The ACF and PACF plots are listed below for figure 10 and figure 11.

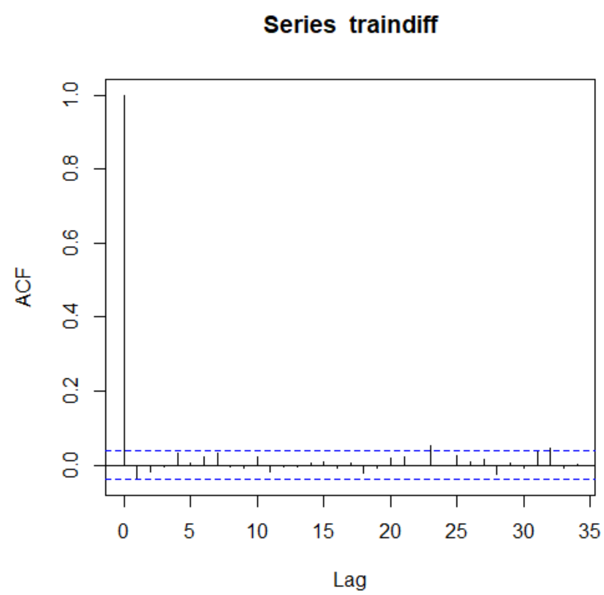



Figure 10

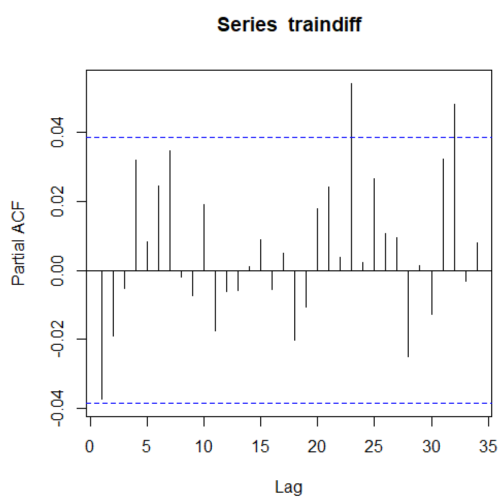



Figure 11

According to the result from the sample ACF and sample PACF plot, lag 23 in sample ACF plot and lag 32 in sample PACF plot exceeds the significance bounds. However, the model is too large, if not useless. Therefore I decided to do one more differencing. Although this might lead to the over differecing problem as this time series is already stationary after 1 differencing. After the consultation from the tutor, the lag 1 is not over negative 0.5 for the sample ACF plot. The Time series can conduct one more time of differencing. The time plot, sample ACF and sample PACF plot after the second time differencing are listed below.

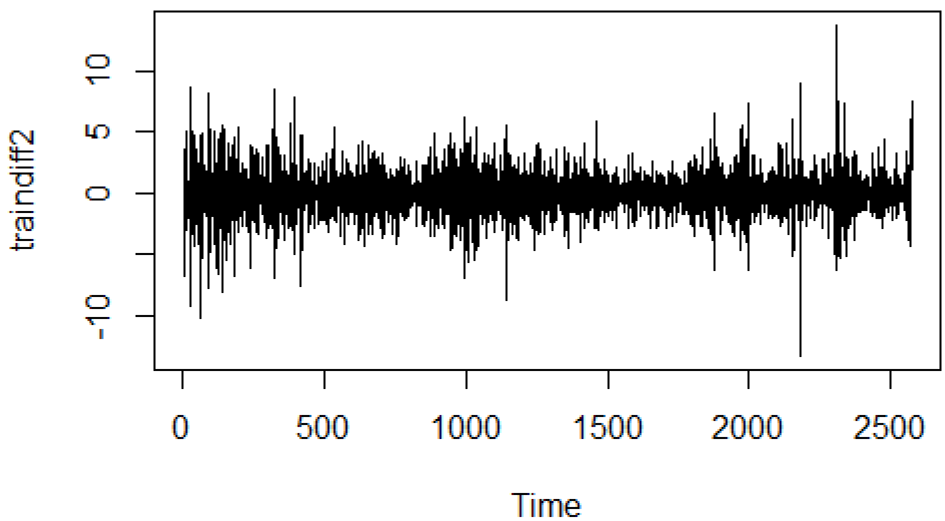



Figure 12

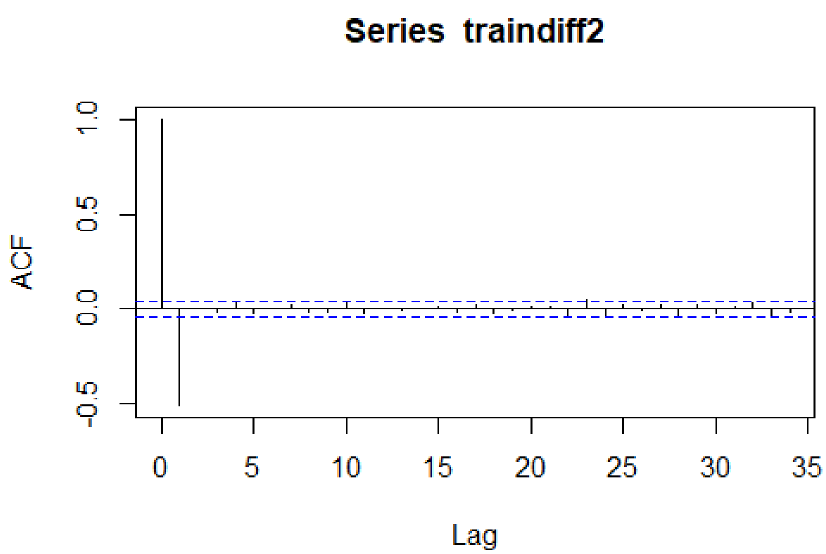


Figure 13

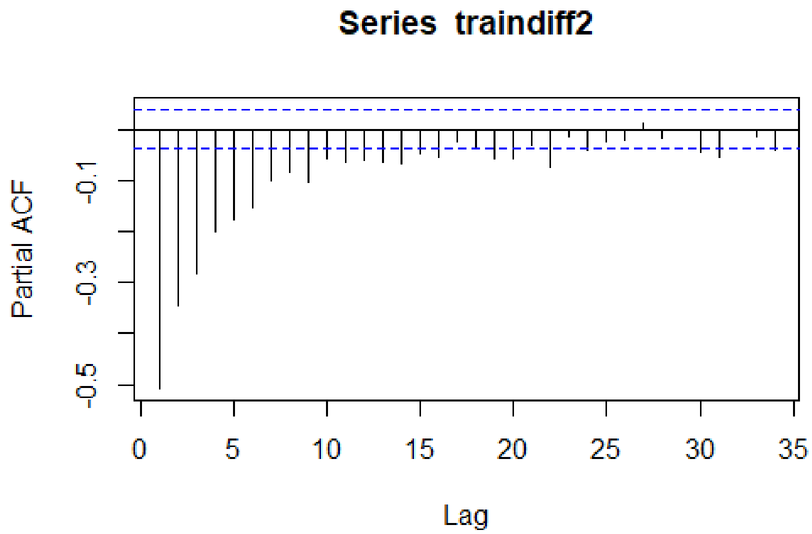



Figure 14

According to the second differencing time plot, it still looks more stationary in mean and variance. For the sample ACF plot, lag 1 exceeds the significance bounds. For the sample PACF plot, lag 1 exceeds the significance bounds the most. Some people may argue that the lags like 2,3,4,5., etc also exceed the 95 % significance bounds for a bit. This might be because of the error and lag 1 is the most significant. Therefore, by the result of the sample ACF plot and sample PACF plot. I should use the ARIMA (1,2,1) mode.

## 3.2 Akaike's information criterion (AIC) and Maximum Likelihood Estimation (MLE)

This method is for comparing the smallest AIC values between the different ARIMA models automatically and selecting the model with the smallest AIC value. After That, using the MLE to find out all the parameters of the selected model. The details of the method are listed in the following figure.

Data without differencing

```
> AICMLE<-auto.arima(train,ic=c("aic"),trace=T,method="ML")

 Fitting models using approximations to speed things up...

 ARIMA(2,1,2) with drift         : Inf
 ARIMA(0,1,0) with drift         : 8913.728
 ARIMA(1,1,0) with drift         : 8912.11
 ARIMA(0,1,1) with drift         : 8911.973
 ARIMA(0,1,0)                    : 8912.341
 ARIMA(1,1,1) with drift         : 8913.468
 ARIMA(0,1,2) with drift         : 8913.223
 ARIMA(1,1,2) with drift         : 8915.226
 ARIMA(0,1,1)                    : 8910.644
 ARIMA(1,1,1)                    : 8912.159
 ARIMA(0,1,2)                    : 8911.917
 ARIMA(1,1,0)                    : 8910.777
 ARIMA(1,1,2)                    : 8913.92

 Now re-fitting the best model(s) without approximations...

 ARIMA(0,1,1)                    : 8910.644

 Best model: ARIMA(0,1,1)

> AICMLE
Series: train
ARIMA(0,1,1)

Coefficients:
          ma1
      -0.0387
s.e.   0.0201

sigma^2 estimated as 1.864:  log likelihood=-4453.32
AIC=8910.64   AICc=8910.65   BIC=8922.35
```

Data with differencing

```
> AICMLE<-auto.arima(traindiff,ic=c("aic"),trace=T,method="ML")

 Fitting models using approximations to speed things up...

 ARIMA(2,0,2) with non-zero mean : Inf
 ARIMA(0,0,0) with non-zero mean : 8913.728
 ARIMA(1,0,0) with non-zero mean : 8912.11
 ARIMA(0,0,1) with non-zero mean : 8911.973
 ARIMA(0,0,0) with zero mean     : 8912.341
 ARIMA(1,0,1) with non-zero mean : 8913.468
 ARIMA(0,0,2) with non-zero mean : 8913.223
 ARIMA(1,0,2) with non-zero mean : 8915.226
 ARIMA(0,0,1) with zero mean     : 8910.644
 ARIMA(1,0,1) with zero mean     : 8912.159
 ARIMA(0,0,2) with zero mean     : 8911.917
 ARIMA(1,0,0) with zero mean     : 8910.777
 ARIMA(1,0,2) with zero mean     : 8913.92

 Now re-fitting the best model(s) without approximations...

 ARIMA(0,0,1) with zero mean     : 8910.644

 Best model: ARIMA(0,0,1) with zero mean

> AICMLE
Series: traindiff
ARIMA(0,0,1) with zero mean

Coefficients:
          ma1
      -0.0387
s.e.   0.0201

sigma^2 estimated as 1.864:  log likelihood=-4453.32
AIC=8910.64   AICc=8910.65   BIC=8922.35
```

Figure 15

```
> summary (AICMLE_withoutdiff)
Series: train
ARIMA(0,1,1)

Coefficients:
          ma1
      -0.0387
s.e.   0.0201

sigma^2 estimated as 1.864:  log likelihood=-4453.32
AIC=8910.64   AICc=8910.65   BIC=8922.35

Training set error measures:
                      ME    RMSE       MAE         MPE     MAPE      MASE         ACF1
Training set -0.02194014 1.36477 0.9765898 -0.05293452 1.497826 0.9971566 0.0004836412
```

Figure 16

According to Figure 12, the result shows us we should use the ARIMA (0,1,1) model. Therefore we will conclude that ARIMA (0,1,1) and ARIMA (1,2,1) are the two models we will evaluate for this time series.

## 3.3 Estimate of parameters of the fitted models

I have used the Maximum likelihood to estimate the parameters of the ARIMA (1,2,1) and ARIMA (0,1,1).

```
> fit<-arima(train, order=c(1,2,1), method="ML")
> fit

Call:
arima(x = train, order = c(1, 2, 1), method = "ML")

Coefficients:
          ar1      ma1
      -0.0372  -1.0000
s.e.   0.0197   0.0021

sigma^2 estimated as 1.864:  log likelihood = -4455.79,  aic = 8917.58
```

```
> summary (AICMLE_withoutdiff)
Series: train
ARIMA(0,1,1)

Coefficients:
          ma1
      -0.0387
s.e.   0.0201

sigma^2 estimated as 1.864:  log likelihood=-4453.32
AIC=8910.64   AICc=8910.65   BIC=8922.35

Training set error measures:
                      ME    RMSE       MAE         MPE     MAPE      MASE         ACF1
Training set -0.02194014 1.36477 0.9765898 -0.05293452 1.497826 0.9971566 0.0004836412
```

Figure 17

## 3.4 Ljung-Box test

We then conduct the Ljung-Box test for both ARIMA (1,2,1) and ARIMA (0,1,1). The result is shown in the figure below.

**ARIMA (1,2,1)**

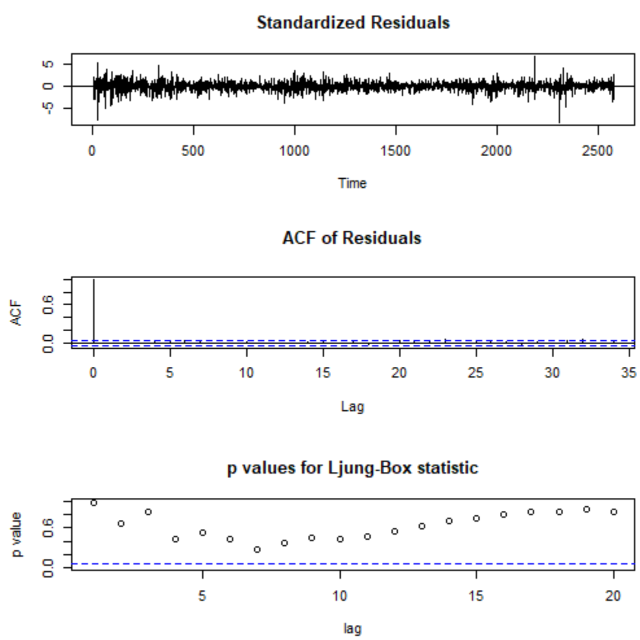


Figure 18

**ARIMA (0,1,1)**

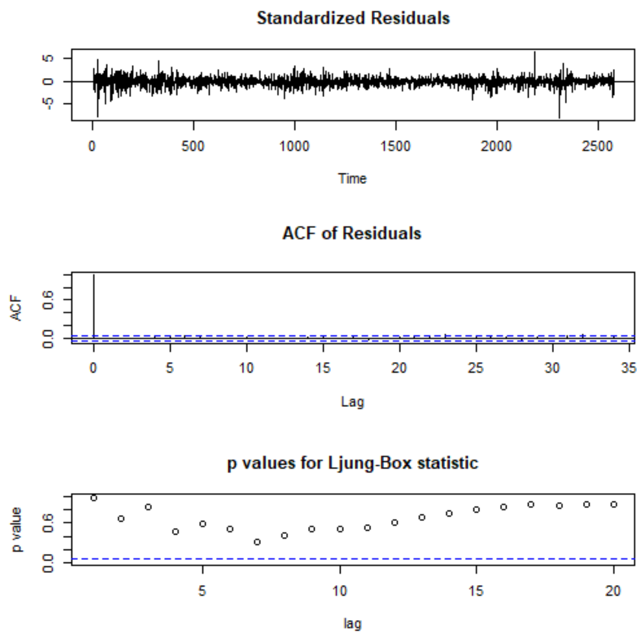
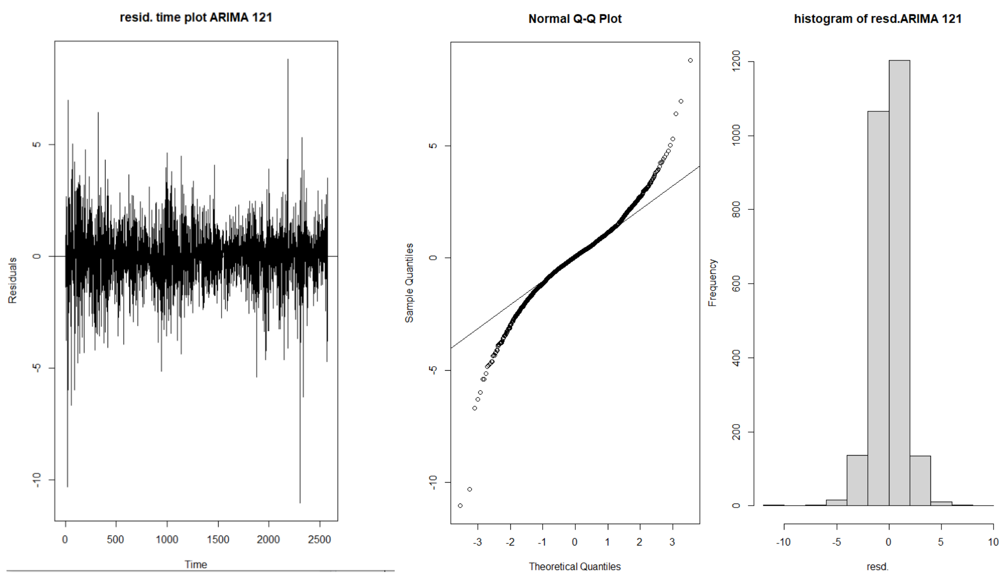


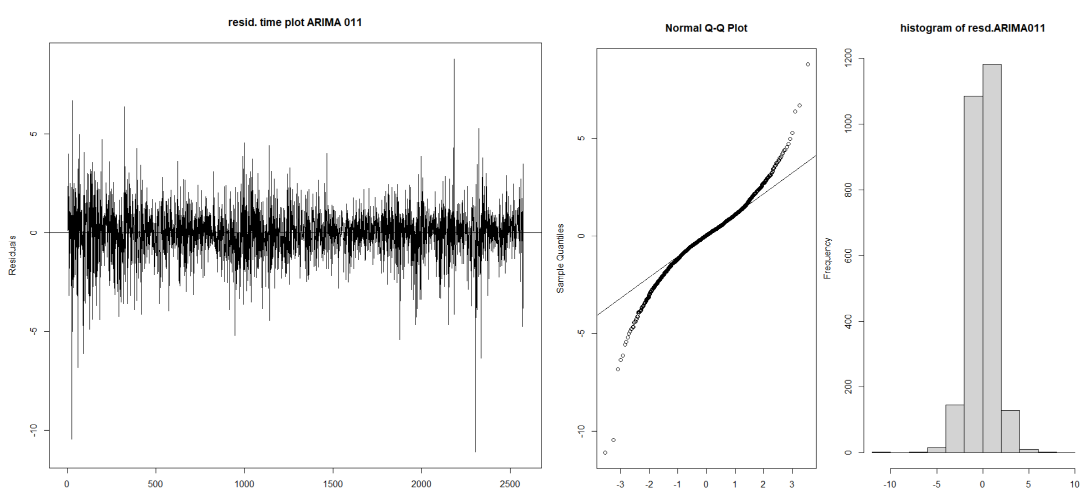

Figure 19

We can conclude that there is no correlation for the residuals from K = 0 to K = 20. The residuals are white noise. For the ACF of residuals, some people may argue that the lag 32 also exceeds the significance bound by a little bit. It might be the 5% error from the 95% confidence interval. Overall, we can conclude that both models are adequate and the parameters and model information are extracted sufficiently

## 3.5 Test for normality




The above figures are the residual timeplot and the QQ-plot and histogram for the residual after fitting ARIMA (1,2,1) with the time series data. We can see that the residual has no trend of normal distribution. The p-value of the Shapiro test is less than 2.2* e^-16. It means that we can reject the null hypothesis and the residuals are not normally distributed.




The above figures are the residual timeplot and the QQ-plot and histogram for the residual after fitting ARIMA (0,1,1) with the time series data. We can see that the residual has no trend of normal distribution. The p-value of the Shapiro test is less than 2.2* e^-16. It means that we can reject the null hypothesis and the residuals are not normally distributed.

## 3.6 overparameterized method for the adequacy of the fitted models

By the overparameterized method, I chose ARIMA (2,2,1) and ARIMA (1,2,2) for the ARIMA (1,2,1) model and ARIMA (1,1,1) and ARIMA (0,1,2) for ARIMA (0,1,1) models. The results are shown in the following figure.

```
> arima(train, order=c(1,1,1), method="ML")

Call:
arima(x = train, order = c(1, 1, 1), method = "ML")

Coefficients:
         ar1      ma1
      0.2258  -0.2651
s.e.  0.2929   0.2899

sigma^2 estimated as 1.863:  log likelihood = -4453.08,  aic = 8912.16
> arima(train, order=c(0,1,2), method="ML")

Call:
arima(x = train, order = c(0, 1, 2), method = "ML")

Coefficients:
          ma1      ma2
      -0.0380  -0.0163
s.e.   0.0197   0.0191

sigma^2 estimated as 1.863:  log likelihood = -4452.96,  aic = 8911.92

> arima(train, order=c(2,2,1), method="ML")

Call:
arima(x = train, order = c(2, 2, 1), method = "ML")

Coefficients:
          ar1      ar2      ma1
      -0.0378  -0.0186  -1.0000
s.e.   0.0197   0.0198   0.0022

sigma^2 estimated as 1.863:  log likelihood = -4455.35,  aic = 8918.7
> arima(train, order=c(1,2,2), method="ML")

Call:
arima(x = train, order = c(1, 2, 2), method = "ML")

Coefficients:
         ar1      ma1     ma2
      0.2281  -1.2673  0.2673
s.e.  0.3000   0.2974  0.2974

sigma^2 estimated as 1.863:  log likelihood = -4455.48,  aic = 8918.97
> |
```

Figure 20

According to the result, coefficients of the AR(1) in ARIMA (1,1,1) model is not significantly different from zero and it is closer to 0 compared to the MA(1). The coefficients of the MA(2) model in ARIMA (0,1,2) model is very close to 0. The AR (2) model in the ARIMA (2,2,1) model does not change the parameters in the original estimate of the ARIMA (1,2,1) model. The MA (2) model in the ARIMA (1,2,2) model also does not change the original estimate of the ARIMA (1,2,1) model. The coefficient of AR (2) and MA(2) are not significantly different from

0. The AIC values of the overparameterized models are also larger than that in the ARIMA (0,1,1) model.

By comparing the AIC value of the ARIMA (0,1,1) model and the ARIMA (1,2,1) model, the AIC value of the ARIMA (0,1,1) model is smaller.  Therefore, we can conclude and choose that the ARIMA (0,1,1) model is more streamlined and adequate, indicating that the problem of overfitting does not exist.

# 4 Forecasting

Foresting is the most important part in the whole time series process as it can help us to make the decision in the future. The future 5 value forecast is listed in the following figure.

```
> forecast(fitARIMA011,h=5,level=c(99,5))
     Point Forecast     Lo 5     Hi 5    Lo 99    Hi 99
2576       64.27578 64.19018 64.36138 60.75969 67.79187
2577       64.27578 64.15704 64.39451 59.39844 69.15312
2578       64.27578 64.13132 64.42024 58.34164 70.20992
2579       64.27578 64.10952 64.44203 57.44646 71.10510
2580       64.27578 64.09027 64.46128 56.65573 71.89583
```

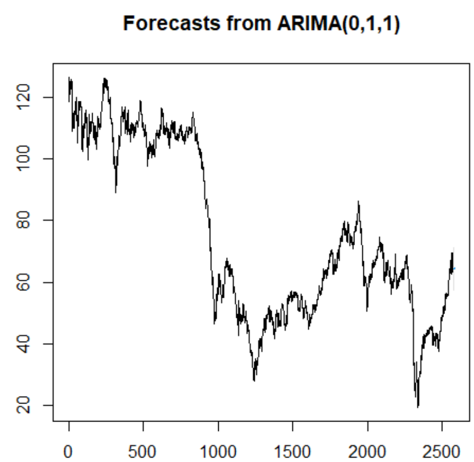



Figure 21

It shows that the forecasted 5 values are 64.27578. It is the same value because the model does not have any trend for the ARIMA model and it is MA(1) model. The following figure shows the prediction accuracy of the forecast. The result shows that the relative error or the prediction error for the first one is relatively large and the next 4 one are relatively smaller. But overall, the prediction result is good. The first data might be affected by other factors like the news of the increasing extraction from the OECD earlier that week. But generally, the time series can accurately predict the near future.

It is normal that for a lot of financial data we will eventually come up with the MA(1) model as it is how a martingale under an efficient market looks like. Therefore, the best predictable price (T+1) is today's price (T). This is a good model for predicting T+1 but not for the further future as there are no better predictions that $Y_{t+1}$. Indeed

$$E_t(Y_{t+2}) = Et(Y_{t+1} + \beta\epsilon_{t+1} + \epsilon_{t+2}) = E_t(Y_{t+1}) = Y_t + \beta\epsilon_t$$ . This

will hold for any $K \geq 2$.

```
> cbind(test[1:5],pre,error,percentageerror)
Time Series:
Start = 2576
End = 2580
Frequency = 1
     test[1:5]      pre      error percentageerror
2576     61.95 64.27578  2.3257792       3.7542844
2577     64.57 64.27578 -0.2942208      -0.4556618
2578     64.98 64.27578 -0.7042208      -1.0837501
2579     64.14 64.27578  0.1357792       0.2116919
2580     63.54 64.27578  0.7357792       1.1579780
```

Figure 22

# 5 Forecasting amendment

Since we need to forecast 5 future values, we need to use another model rather than ARIMA (0,1,1). As aforementioned, we have come up with the ARIMA (1,2,1) model. Here we will try to forecast with the ARIMA (1,2,1) model.

The future 5 forecasted values are listed in the following figure.

```
> cbind(test[1:5],preamendment,erroramendment,percentageerroramendment)
Time Series:
Start = 2576
End = 2580
Frequency = 1
     test[1:5] preamendment erroramendment percentageerroramendment
2576     61.95     64.25350     2.30349704                3.7183164
2577     64.57     64.23735    -0.33264878               -0.5151754
2578     64.98     64.21599    -0.76401110               -1.1757635
2579     64.14     64.19482     0.05482046                0.0854700
2580     63.54     64.17364     0.63364481                0.9972377
```

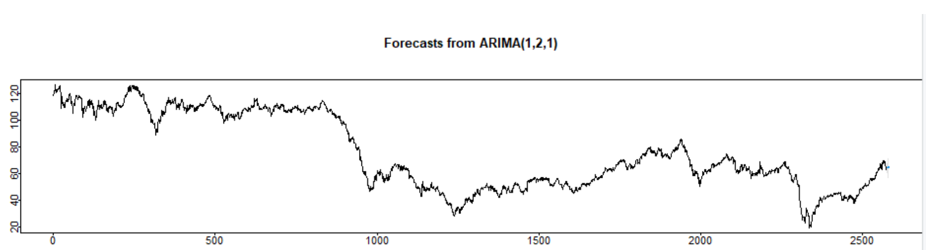


Figure 23

The figure above shows the prediction power for the ARIMA (1,2,1) model. The prediction accuracy is similar to the ARIMA (0,1,1) model for the first 3 values but the ARIMA (1,2,1) predicts better than ARIMA (0,1,1) for the 4th and 5th values. We can conclude that the ARIMA (1,2,1) model predicts better than ARIMA (0,1,1) when it comes to a longer future. It matched our assumption above.

# 6 Reflection

Before I selected the topic of Brent Crude oil futures price, I wanted to choose the Western Texas Intermediate (WTI) oil futures price. However, around the april 2000. The WTI futures has reached the historical lowest negative price because of the oil trade war between Saudi Arabia and Russia and the development of coronavirus. It has a lot of outliers. Therefore, I chose Brent Crude oil futures price as it serves as a benchmark for ⅔ of the oil prices in the world. Since it does not have any trend and like many other financial indices, it is non-stationary and It is very hard to fit any models. It turns out that the result is quite matched with what I expected. The best predictable price (T+1) is today's price (T) as it is how a martingale under an efficient market looks like. Therefore, the MA (1) model is a good model. However, this project requires us to forecast 5 future values. ARIMA (0,1,1) model will give the same result for all 5 values. It is only good for the prediction of the near future like T+1, T+2. But when it comes to a longer future like T+4, T+5. It needs to have another model. In this case we use the ARIMA (1,2,1) model. Although the AIC value of ARIMA (1,2,1) is bigger than that of ARIMA (0,1,1), we still choose to use this model for a longer future prediction and the result shows that ARIMA (1,2,1) predict better in a longer future like in T+4 and T+5.

To conclude, the residuals of both models are not normally distributed. Therefore, both of them might not be good models for our data. More complicated models like non-linear models are needed for our modelling.

By doing this project, I realized that we need to change our model based on the real life situations and different circumstances.

# 7 Conclusion

The oil futures price is heavily affected by the decision of The Organization of the Petroleum Exporting Countries (OPEC). When OPEC decided to increase or decrease the production of oil, the oil futures price usually fluctuate a lot. For example, OPEC decided to decrease the oil production on 7 Jan 2021. The futures price then increased in the following few trading days. The decisions of the OPEC are hard to predict, if not impossible. This will affect the performance of our models. These are the errors that we may encounter during the forecasting. By the analysis above, I can conclude that the ARIMA (0,1,1) model is better for the short term future or forecasting for T+1 but it is better to use the ARIMA (1,2,1) model for the longer futures like T+4 and T+5. Given the effect of the COVID-19 pandemic and the unstable political environment in Russia and Middle Eastern countries, the Brent Crude oil futures price is expecting to have a down trend for the long run. To conclude, the residuals of both models are not normally distributed. Therefore, both of them might not be good models for our data. More complicated models like non-linear models are needed for our modelling.

# 8. Acknowledgement

I would like to thank you Dr. Guodong Li and two of the tutors Miss Xiaoyu Zhang and Miss Feiqing Huang for your effort for teaching this course and the advice for the overdifferencing problem.

# 9 References


A-little-book-of-r-for-time-series.readthedocs.io. 2021. *Using R for Time Series Analysis — Time Series 0.2 documentation*. [online] Available at: <https://a-little-book-of-r-for-time-series.readthedocs.io/en/latest/src/timeseries.html#differencing-a-time-series> [Accessed 27 April 2021].

ARIMA(0, 1., Cavalcante, P. and Cavalcante, P., 2021. *ARIMA(0,1,1) repeating the same point estimate*. [online] Cross Validated. Available at: <https://stats.stackexchange.com/questions/424756/arima0-1-1-repeating-the-same-point-estimate> [Accessed 27 April 2021].

Bloomberg.com. 2021. *Bloomberg - CO1:COM, Brent Crude Oil Futures*. [online] Available at: <https://www.bloomberg.com/quote/CO1:COM> [Accessed 27 April 2021].

CNBC. 2021. *ENERGY Oil rises as supply constraints retain focus amid U.S. Capitol drama*. [online] Available at: <https://www.cnbc.com/2021/01/07/oil-markets-us-politics.html> [Accessed 27 April 2021].

E. E. Holmes, E., 2021. *5.3 Dickey-Fuller and Augmented Dickey-Fuller tests | Applied Time Series Analysis for Fisheries and Environmental Sciences*. [online] Nwfsc-timeseries.github.io. Available at: <https://nwfsc-timeseries.github.io/atsa-labs/sec-boxjenkins-aug-dickey-fuller.html> [Accessed 27 April 2021].

Moshiri, S. and Foroutan, F., 2006. Forecasting Nonlinear Crude Oil Futures Prices. *The Energy Journal*, 27(4).