Regression Analysis of Influencing Factors on the Basis Spread Between Crude Oil Spot and Futures Price

YICHAO, IVAN, DAI
Wenzhou-Kean University, Junior, ID 1098325

West Texas intermediate Crushing, Oklahoma Crude Oil as one of the main benchmarks in oil pricing, becomes a popular higher commodity used by producers and refiners. Traders in the futures market might suffer some losses or gains caused by basis risks. This project aims to use the multi-linear regression analysis to predict the basis spread between the West Texas intermediate Crushing, Oklahoma Crude Oil spot price, and its 4-month futures contract price. Hedgers can partly eliminate the basis risks or even gain some arbitrage profits by knowing the basis in advance. In this project, researchers try to pick the most relative explanatory variables to build a multi-linear regression model to predict the basis of WTI OK crude oil. Oil production level, oil consumption level, speculative index, economic conditions, and several financial market factors had been taken into consideration. The research also excludes the impact of the Covid-19 outbreak on a WTI basis spread to reduce the impact on the accuracy of the prediction model. The predictive model built in this project is powerful to predict the movement direction or even the basis spread when there is a significant event in the world market such as a financial crisis, however, when there is no significant event over the world market, the predicted basis cannot be explained by the linear method. The final model has a relatively good adjusting R square equal to 25.6%, which can explain 25.6% of the WTI crude oil basis spread.

Keywords and Phrases: basis spread, crude oil, model selectin, OLS regression, LASSO regression

1 INTRODUCTION

West Taxes Intermediate crude oil has a relatively low density since it only has 0.24% sulfur content, therefore, WTI crude oil is also referred to as light crude oil. It's a specific grade of crude oil from Texas that spot and futures prices serve as one of the main benchmarks in oil pricing. The West Taxed Intermediate is the commodity of New York Mercantile Exchange's oil futures contract and is the main oil benchmark in North American. And the main delivery and price settlement point for West Taxed Intermediate crude oil in Cushing, Oklahoma. Also, WTI crude oil futures prices are also included in the Bloomberg Commodity Index and S&O GSCI commodity index. Therefore, WTI crude oil futures contracts gradually develop as hedging tools used by producers and refiners. Basis usually occurs when the hedgers are uncertain of the exact date when the asset will be bought or sold, and the hedgers may be required to close their current position before maturity days. And basis risk refers to financial risk that will happen while using hedging strategies which brings the potential for gains or losses. When there exist large amounts of shares or contracts in transactions, basis risk may bring significant losses or gains. In this project, the asset to be hedged and the asset under the futures contract is the same. Consequently, the basis should be zero at the maturity date. However, the basis may be positive or negative prior to maturity. Hence, explore the basis spread of WTI crude oil spot and futures price help investors who invest in WTI crude oil to eliminate some basis risks and minimize losses or even gain some arbitrage profit. However, previous researches have shown that basis prediction is a complex process, involving different factors in different markets (An et al., 2014). Some of the literature shows that the non-linear method is an optimal choice to predict the movement of basis. The multi-regression analysis helps determine correlations between one dependent variable and more independent variables and to have some predictions among these variables. (Unver & Gamgam, 1999, as cited in Uyanik & Güler, 2013).
The purpose of this research is to predict the basis spread between WTI Crushing, Oklahoma Crude spot price, and its 4-months futures contract price by using different multi-linear regression methods and chose the most fitted one.

2 LITERATURE REVIEW

Previous researches showed many great ideas on how to predict the futures price of crude oil and the factors influencing the spot market. However, only a few researches are exploring the basis spread between the crude oil spot price and the futures price. As a result, in this literature review, the researchers aim to explore the factor that can both affect the spot market and futures market of the WTI crude oil

2.1 Oil Production Capacity & Oil Consumption Level

Crude oil is one kind of exhaustible resource in the world (Miao et al., 2017). The Middle East and North Africa, as the main supplier, play an important role in the spot market. In 2009, Kaufmann and Ullman try to explore in their article "Interpreting causal relations among spot and futures prices" that how the innovations in oil prices first enter the market, after comparing the spot price and futures price in a different region, they find that the innovation first appears in the Middle East region and

then spread to other spot and futures price, which suggest the situation in Middle East region plays an important role in the fluctuation of the WTI spot and futures market.

2.2 Situation in Middle East and North Africa (MENA)

Miao and his team (2017) discussed the influential factors of the crude oil price forecasting and indicated the oil production capacity in the Middle East area had a strong relationship with the WTI crude oil price. Miao also found that the WTI spot price and futures price fluctuate because of some political factors, such as the total amount of terrorist attacks in the Middle East and North Africa. Considering that oil is actually an important energy resource, it would be easily targeted by terrorism, and it seems that the spot price and the futures price of crude oil would be vulnerable to be influenced by terrorist attacks. However, Holwerda and Scholtens (2016) found that the spot price and the futures price of crude oil did not have a significant reflection to the terrorist attacks, since the market had already included the terrorist attacks in the risk premium previously. It means that the market had already made the reaction to the terrorist attacks, and it makes the spot price and futures price of crude oil did not has too many changes. Of course, the spot price and the futures price of crude oil in the different markets would also have different reactions to the terrorist attacks. Kollias et al. (2013) tested the co-movement between oil returns and four market indexes under the influence of war and terrorism. It shows that the spot price and futures price of crude oil in S&P500, FTSE100 would have more capacity in absorbing the risk of the terrorist attacks, and the spot price and the futures price of crude oil CAC40, DAX did not.

2.3 Speculative Index: NTM1-CNCN Index

Kaufmann and Ullman (2009) also stated that the relationship between the spot market and the futures market was relatively weak, however, when the market initiated a long-term increase in oil price, this trend would be exacerbated heavily by the speculators. As a result, the number of speculators in the WTI oil futures market may have some contribution to the variance of the basis of the WTI crude oil. Other researchers (An et al., 2014) also argue that the relationship between co-movement of the spot & futures prices and the net non-commercial futures position, which could be measured as a speculative index, they found that the speculative index played an uncertain role in the spot and futures market. Although the previous research before 2008 has shown the price collapse would be accompanied by a significant drop in the speculative index, the oil price collapse in 2008 did not indicate a large drop in the speculative index (An et al., 2014). Moreover, Bu (2011) stated in his research that "reveal that the position changes held by speculative traders will cause crude oil price movement", especially when the financial crisis occurred, speculative traders would inject the futures price large vitality. Moreover, Stoll and Whaley (2015) found in their research that excessive speculation will cause the same direction movement in the spot and futures price of crude oil.

2.4 US Economics Condition: US GDP Growth Rate, US CPI, Dollar Index

US GDP Growth Rate & Dollar Index Economics condition plays a significant role to determine the spot price of crude oil. Miao (2017) stated in his research that the world economic growth is closely related to crude oil demand. Higher global economics would raise the spot price of the oil, however, a lower global economic growth rate would lead to the fall of the spot price. He also stated several factors such as the steel production and ISM manufacturing index which might have some effect on the spot price. Study (Wang & Wu, 2012) also show that a weaker dollar exchange rate may result in a higher spot price. The study also showed a similar relationship in the oil futures price, Algieri (2014) found in his research that a weaker US/Euro exchange rate could bring a decline in the futures commodity return. However, the relationship is not as strong as the price in the spot market. Furthermore, Amendola et al. (2017) explored that the expansionary monetary policy would have a negative impact on the crude oil future price and the fluctuation of the industrial production would make the crude oil future price have the opposite correlation. Then, Frondel et al. (2019) explained in their research that the production of crude oil in the US would also create an opposite influence on the spot price of crude oil. Furthermore, Basistha and Kurov (2015) also found that the changes in the crude oil price would create a significant influence on the federal funds target rate, but, with the analysis of the VAR model, it seems that the crude oil price might not have contemporaneous feedback with the federal funds rate shocks.

2.5 Financial Factors

Studies have shown that the financial factors could produce effects both on the spot price and futures price of the crude oil. Algieri (2014) stated in his research that Standard and Poor Index 500 positively affected the commodity futures market. Miao (2017) also found a significantly positive relationship between the oil stock price and the oil spot price. Moreover, according to the article by Jones and Kaul (1996) and Sadorsky (1999), they found that the stock market and oil prices tended to move in the same direction.

3 METHDOLOGY

3.1 Overall Approach

The project wound conduct time-series quantitative research, which aimed to explore the influencing factors on the basis risk of WTI Crushing OK Crude Oil. The overall approach was to use the unit root test and cointegration test to build a rational and valid ordinary least square (OLS) multi-regression model, and then the project will use the features-deduction method to avoid model overfitting. The project would also deal with the multicollinearity problem by checking the variance inflation factor (VIF). Moreover, The project would make some assumptions about the initial model, such as normality of the error, to fulfill the condition of use of the model and method. Last but not least, the project would use the adjusted R square and root mean square error (RMSE) to evaluate the overall model.

3.2 Method of Data Collection

All the data are quantitative data, which is collected through the existing data sources, Bloomberg and the U.S. Energy Information Administration (EIA), which is the official energy statistics from the U.S. Government (U.S. Energy Information Administration (EIA), 2020). These two data sources are reliable data sources for many researchers working on econometrics and finance. Because of Coronavirus-19, the crude oil spot market and the futures market are affected heavily in some time period (Ajifowoke, 2020). To eliminate the effect and corresponding effect of Coronavirus-19, the project will only select the data between January 2000 and December 2018, a 19-year range data set. All the explanatory variables would also be selected in this specific time period.

3.3 Data Description

The following table *(table 1)* shows a brief introduction to the response variables and explanatory variables:

Variable	Data Type	Time Series	Brief Description
Basis	Numeric	Monthly	WTI Oil Basis Spread
Dow Jones Oil Index	Numeric	Monthly	Stock Performance in the oil and gas sector
NTM1-CNCN Index	Numeric	Monthly	Oil Speculative Index
Open Interest-Crude Oil	Numeric	Monthly	Total Open interest for the oil futures
CPI US	Numeric	Monthly	A way to measure the US inflation
US Federal Rate	Numeric	Monthly	Factors Affect the consumers ST loans
OCED Oil Consumption	Numeric	Monthly	Demand factors, world oil consumption level
S&P 500 Index	Numeric	Monthly	US financial market performance
US Dollar Index	Numeric	Monthly	Measure Dollar Strength
US GDP Growth Rate	Numeric	Quarterly	US economic overall condition
OPEC Total Surplus	Numeric	Monthly	Supply factors, world oil production level
Terrorism Attack	Numeric	Monthly	Amount of attack in MENA

Table 1: Brief Data Descriptions

The response variable, basis, is calculated by using the WTI Crushing OK Crude Oil spot price and the 4-month WTI OK Crude Oil futures prices. Both the 4-month futures price and spot price were downloaded from the U.S. Energy Information Administration (EIA).

The explanatory variables are coming from a different aspect, however, the most influential factors should come from the demand and supply side. OPEC Total Surplus Index measures the production level over the world, it can reveal the supply ability of crude oil. Moreover, the OCED oil consumption index measures the consumption level over the world, it reveals the demand power from the consumers' side. Both of the data are collected from Bloomberg. NTM1-CNCN Index stands for the crude oil net non-commercial futures position. And Open Interest-Crude Oil reveals the total open WTI crude oil contracts in the futures market. A trader must report his or her position if, at the daily close of the market, their position reaches or exceeds the CFTC reporting level at the expiration of any futures month or option (Bloomberg). Under CFTC rules, if a trader uses a futures contract to hedge for a particular commodity, all futures positions reported by the trader are classified as commercial. Previous researches also use the net non-commercial position as a measure of the speculative level. The data is collected from Bloomberg. MENA Terrorism Attack measures the frequency of terrorist attacks in the Middle East and North Africa Region. The data comes from the Global Terrorism Database (GTD), which recorded the daily terrorism events in the world. U.S. GDP growth rate measures the US economic situation and US CPI measures the inflation level within US. Both of the data are collected from Bloomberg. However, GDP is the only quarter data in the dataset. We will use the linear interpolation method to fill the missing NA values of the GDP growth rate. The U.S. Dollar Index measures the value of the dollar relative to a basket of currencies in most of the country's most important trading partners (Investopedia). It can measure the weighted strength of the DOLLAR against other currencies. The data is collected from Bloomberg. The S&P 500 is widely seen as the best single measure of U.S. large-cap stocks and is the basis for a range of investment products (Bloomberg), which can be seen as a measure of the US equity market condition. Also, Dow Jones Oil Index is designed to measure the stock performance

of companies in the US oil and gas industry. Both of the data is collected from Bloomberg. A data sample table is showing as following.

Date	GDP Growth Rate	CPI US	DJ Oil Index	…	Basis
…	…	…	…	…	…
2000-03-01	4.2	3.8	306.98	…	3.03
2000-04-01	4.57	3.1	304.13	…	1.34
2000-05-01	4.93	3.2	334.44	…	1.45
2000-06-01	5.3	3.7	316.44	…	3.14
…	…	…	…	…	…
2018-09-01	3.1	2.3	1185.84	…	69.43
2018-10-01	2.9	2.5	1047.56	…	0.04
2018-11-01	2.7	2.2	1022.26	…	-0.15
2018-12-01	2.5	1.9	887.65	…	-0.05

Table 2: Sample Data Set

3.4 Method of Analysis

The project will typically use the STATA and R to do the data cleaning, data analysis, data visualization, features selection, and model building.

3.4.1 Multi-regression Linear Model Selection

In this project, the multi-regression model is built based on the ordinary least square method (OLS), which is a typical method to use several explanatory variables to predict a response variable. By using the ordinary least square method (OLS), we want to minimize the sum of square error (SSE) between the observed response variable and the value of the response variable predicted by the multi-regression model (Tufféry, 2011). The OLS linear regression model can be written as follow:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_n X_n + \varepsilon \sim N(0, \sigma^2)$$

Y is the response variable in the formula, which is also the basis risk of WTI, OK Crude Oil in this project. X is the explanatory variable in the formula, which should be tested and selected to predict the response variable. ε is the random error of the formula, which is assumed as a normal-distributed error with an expected value equal to 0 and variance equal to σ^2. β_0 is the intercept of the model, the other β_i are the corresponding coefficient of the explanatory variable. Every β_i havs a *t-value* and corresponding significant level, which decide whether the β_i is significant different from 0. If not, it means the variable do not have enough power to predict the outcome. The interpretation of the value of β_i is that for every unit increase in an explanatory variable X_i with other variables hold constant, the response variable will increase by the value of β_i.

In this project, we will also use features deduction method to update the OLS model. We would use the backward Akaike information criterion (AIC) method and least absolute shrinkage and selection operator (LASSO) method to do the dimension reduction to avoid the overfitting or multicollinearity problems. AIC method start with a full OLS model and remove one variable at a time to get a new AIC, which is calculated as follow:

$$AIC = 2k - \ln(\hat{L})$$

Where:

k: number of estimated parameters in the model.

$\hat{L}$: maximum value of the likelihood function for the model.

The method aims to explore whether the model can achieve a lower AIC, model with a lower AIC would be better. The final model will be selected if deleting any one of the remaining variables cannot achieve a lower AIC.

Another dimension reduction method is LASSO. LASSO method will do the adjustment to the corresponding coefficients of each variables by using following formula:

$$L_{lasso} = \min\left[\sum_{i=1}^{n}\left(y_i - \sum_{j} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right]$$

When lambda is equal to 0, no parameters are eliminated. The estimation of the coefficient is exactly equal to the previous one with OLS linear regression. However, As lambda increases, more and more coefficients are influenced and trend to be zero, when lambda is tend to infinite, all coefficients will be tend to 0 and be eliminated. LASSO method selection process will choose the best lambda with its corresponding model as the predictive model.

3.4.2 Unit Root Test

In a time-series project, it is important to avoid spurious regression, which is caused by a similar local trend. It is possible to build a significant model even if the relationship none exist. In time-series data, it is a common occurrence when the data is not stationary (Hill, Griffiths & Lim, 2017). As a result, the project should do the unit root test to make sure all the variable should be stationary, including the response variable. The common method is (Augmented) Dickey-Fuller tests (Hanck & Czudaj, 2015). The null hypothesis of the ADF test is the presence of the unit root (non-stationary), and the alternative hypothesis state that the variable is stationary. According to the empirical significant level in past researches, the significance level for this project was set as 5%, as a result, when the p-value is higher than the significant level, we cannot reject the null hypothesis and state the variable is non-stationary. If the variable is non-stationary, we may need to transform the variable to a stationary variable to adjust the model.

3.4.3 Cointegrations Test

In common case, the model should not contain the non-stationary variable, however, there is a special case, When Two nonstationary time series are cointegrated if they tend to move together through time, which can be shown as following formula, we can said they are cointegrated stationary (Hill, Griffiths & Lim, 2017):

$$x, y \sim I(1)$$
$$e = y - \beta_0 + \beta_1 x \sim I(0)$$

$I(1)$ means that x and y are non-stationary variable but $\Delta x, \Delta y$ are stationary variable. $I(0)$ means the variable is a stationary variable. In this case, it suggests that x and y are said to be cointegrated. Co-integration helps determine how sensitive two variables are to the same average price over a given period of time. Thus, co-integration does not reflect whether two pairs move in the same or opposite direction, but it can state whether the distance between them remains the same over time (Rahim, 2020). However, if one variable is I(0) variable, and anther variable is a I(1) variable, one should be transformed in the model. If both variable is I(0) variable, then the regression analysis should be valid.

3.4 Assumption of the Model

Although ordinary least squares (OLS) are used in a variety of economic and financial analysis, to be able to reasonably explain the parameters and outputs of the model, the least-squares method needs to satisfy three main assumptions: normal error, homoscedastic, and without multicollinearity. Assumptions are sometimes very difficult to satisfy, which needed certain tests that would allow the project to fulfill those assumptions.

Firstly, the observed data should have a normal distribution error. Non-normality residuals can be misunderstood by the confidence interval model for its predictive ability and create skewed distributions by increasing the appearance of outliers (Pham, 2018). In technical terms, "the Assumption of Normality claims that the sampling distribution of the mean is normal" (Mordkoff, n.d.). Nearly all the inferential statistics such as t-test and ANOVA reply on the assumption of normality

Secondly, the linear relationship should exist between the response variable and explanatory variables. With other explanatory variables holding constant, there should be a linear function to decide the relationship between the response variable and each explanatory variable. Moreover, The effects of each variable on the predictive value of the dependent variable should be additive (Pham, 2018), otherwise, the estimation will be misleading.

Thirdly, multicollinearity refers to the highly correlated relationship among the explanatory variables. Multicollinearity destroys the statistical significance of an independent variable (Allen, 1997), which means we may construct a variable as a linear function of other variables. The common evaluation of Multicollinearity is to use the variance inflation factor (VIF) (Pham, 2018), if VIF is equal to 1, we can state that no explanatory variable are correlated. However, when VIF is more than 10, It indicates a high correlation and is cause for concern. The ideal condition should be VIF less than 3, however, it does not cause concern when VIF is less than 10.

3.5 Evaluation of the Model

Based on the output of the model, we would like to evaluate the overall model performance by checking the Adjusted R square. The basic R square formula is shown as following formula:

$$R^2 = \frac{Expected\ Variance}{Total\ Variance}$$

The formula suggest that R^2 can indicate to what extent can the model explain the variance of the response variable. For example, a R-square value 0.8 indicates that 80% of the variance of dependent variables can be explained by the selected explanatory variables. However, the project would use the adjusting R square rather than simple R square in that the project

use multi-regression model with multi variables; consequently, adjusting R square is the modified simple R square. The adjusting R square formula is showing as:

$$R^2_{adj} = 1 - \left[\frac{(1-R^2)(n-1)}{n-k-1}\right]$$

In Finance field, the model with an adjusted R square above 0.7 would generally be seen as a high predictive model. However, for times series data, because all the variables should be converted to the stationary variables, as a result, an adjusting R square higher than 0.25 can be considered as a quite good model.

Another way to know the model performance is calculating the root mean square error, which is a general way in machine learning to evaluate the model accuracy. The following equation shows how to calculate the RMSE by using the predicted values and the observed values.

$$RMSE = \sqrt{\sum \frac{(\bar{y}_i - y_i)^2}{n}}$$

Where $\bar{y}_i$ is the predicted value by using the model, y_i is the observed value in the dataset. N the number of the observed value.

4 RESULTS

In this section, the research will show the result of the unit root test, model assumption checking, model selection processs, model performance and the comparison among different models.

4.1 Unit root test and cointegration test

Times-series data need the unit root test to avoid spurious regression. The following figure *(figure 1)* shows the times serious of the response variables and another 11 explanatory variables. These variables need to check whether there exists a unit root. The following table *(table 3),* shows the ADF test statistics and the significance level. The significant level is overall 5% in this project. The null hypothesis on the stationary test is that there exists a unit root, and the alternative hypothesis is the variable is a stationary variable. The null hypothesis on the drift test is that there exists a unit root with drift, the alternative hypothesis is that one of these two conditions in the null hypothesis does not meet. If the test statistics is larger than the 5% critical value (absolute value), then we can reject the null hypothesis and accept the alternative. As result, we have found that the response variables and four other explanatory variables including NTM1-CNCN Index, OCED Oil Consumption, US GD, and CPI US are stationary variables, which is also named as I(0) variable. Other explanatory variables are considered as the non-stationary variables. We would like to convert the non-stationary variables to their first difference and check whether the first differences are stationary.

Variable	ADF Test on Stationary	ADF Test On Drift	Constant 5% Critical Value	Constant and Trend 5% Critical Value
Basis	-4.5437***	6.9471***	-3.43	4.75
Dow Jones Oil Index	-1.8973	1.6993	-3.43	4.75
NTM1-CNCN Index	-3.4675***	4.0867	-3.43	4.75
Open Interest-Crude Oil	-3.0976	4.0676	-3.43	4.75
CPI US	-4.4786***	6.7045***	-3.43	4.75
US Federal Rate	-1.263	1.8419	-3.43	4.75
OCED Oil Consumption	-4.7136***	7.4353***	-3.43	4.75
S&P 500 Index	-1.8281	2.5567	-3.43	4.75
US Dollar Index	-1.3244	1.0478	-3.43	4.75
US GDP Growth Rate	-3.8136***	4.8894***	-3.43	4.75
OPEC Total Surplus	-2.5598	2.2842	-3.43	4.75
Terrorism Attack	-1.7006	1.1564	-3.43	4.75

Table 3: ADF Test For Original Data Set

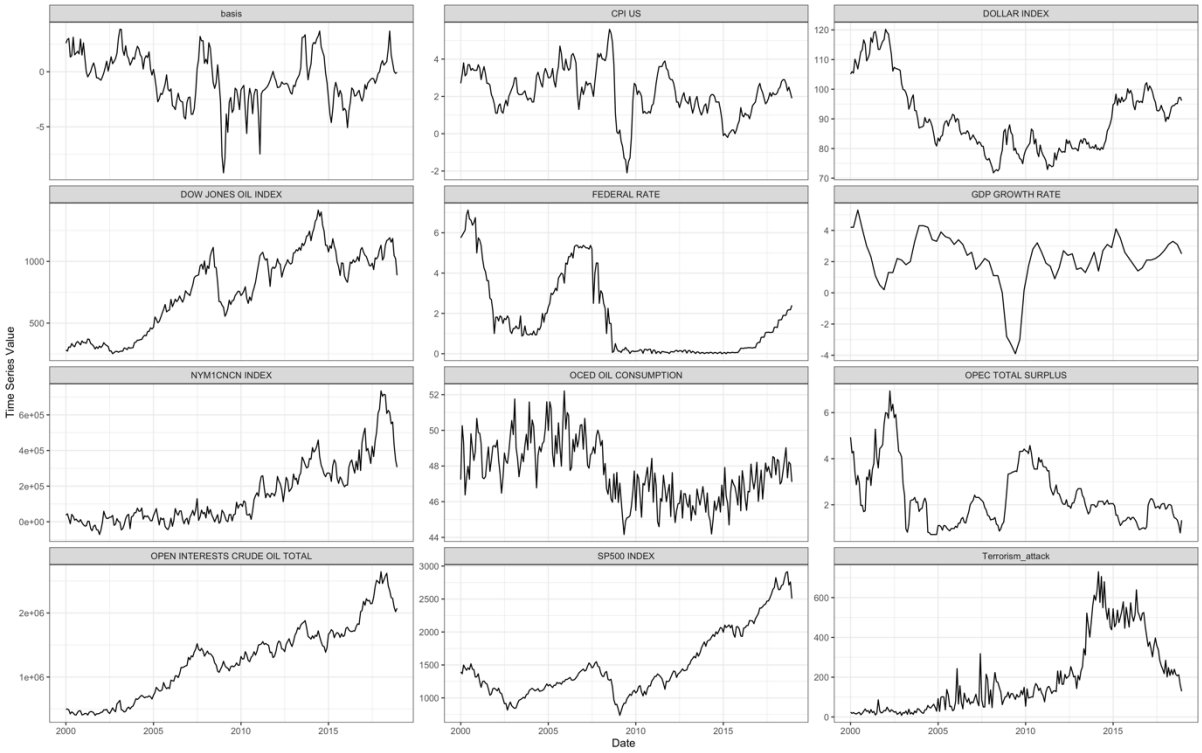

Figure 1: Original Time Series

The following table shows the test statistics for their first difference. We found all the first different of the stationary variables are stationary. As a result, we have converted the whole model dataset to a stationary dataset. The following figure *(figure 2)* shows the times series after converting non-stationary variables into their first difference. Because the response variable is an I(0) variables, as a result, there is no need to do the cointegration test.

Variable	ADF Test on Stationary	ADF Test On Drift	Constant 5% Critical Value	Constant and Trend 5% Critical Value
Dow Jones Oil Index	-8.8393***	26.0969***	-3.43	4.75
Open Interest-Crude Oil	-10.0222***	33.4821***	-3.43	4.75
US Federal Rate	-13.1882***	57.9829***	-3.43	4.75
S&P 500 Index	-10.1739***	34.5773***	-3.43	4.75
US Dollar Index	-9.9518***	33.0143***	-3.43	4.75
OPEC Total Surplus	-8.9708***	26.8305***	-3.43	4.75
Terrorism Attack	-15.5329***	80.4342***	-3.43	4.75

Table 4: ADF Test For the first difference of non-stationary variables

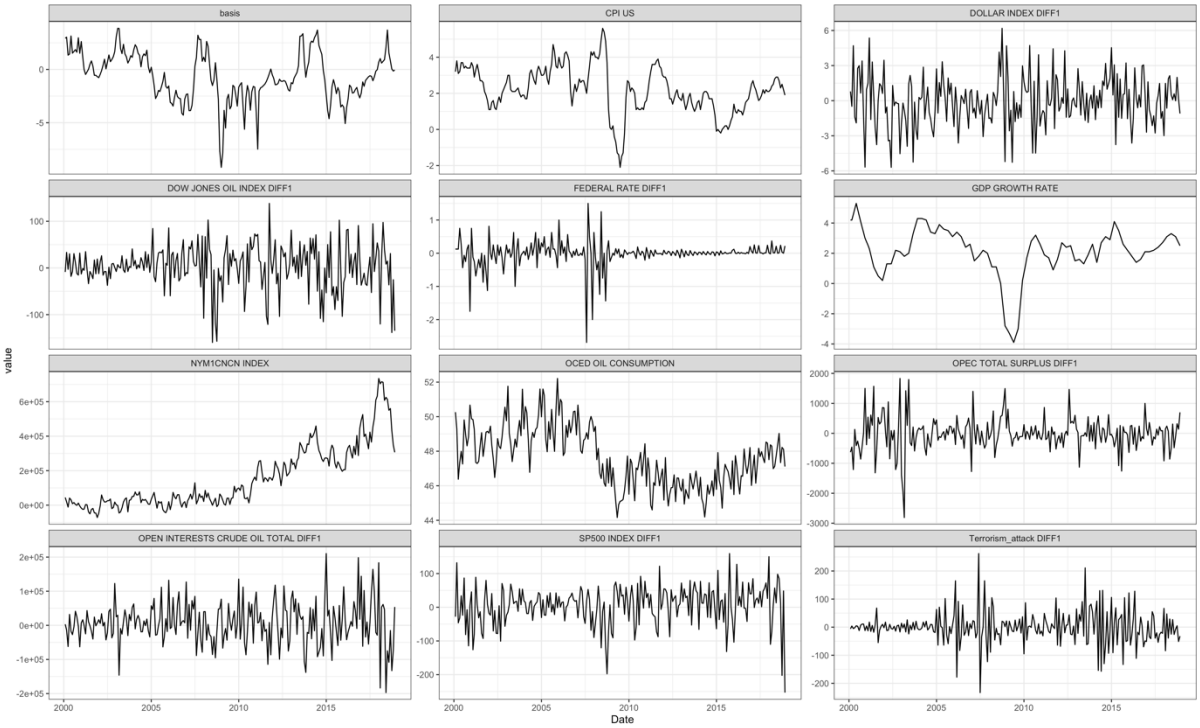

Figure 2: Times Series after converting to stationary

4.2 Checking Model Assumption

In this section, we want to check the assumption of the multi-linear regression model. It will mainly focus on whether there exist multicollinearity problems, whether there is a normal residual, and whether there is homoscedasticity. The following matrix *(figure 3)* is the correlation matrix of the explanatory variables. We can see that most of the correlation between explanatory variables, however, Dow Jones Oil Index and S&P 500 have a relatively high correlation. The variance inflation factor (VIF) of each variable is showing in the following table *(Table 5).* All the VIF of explanatory variables is between 1 to 10, which indicates there is not multicollinearity problem in the OLS model.

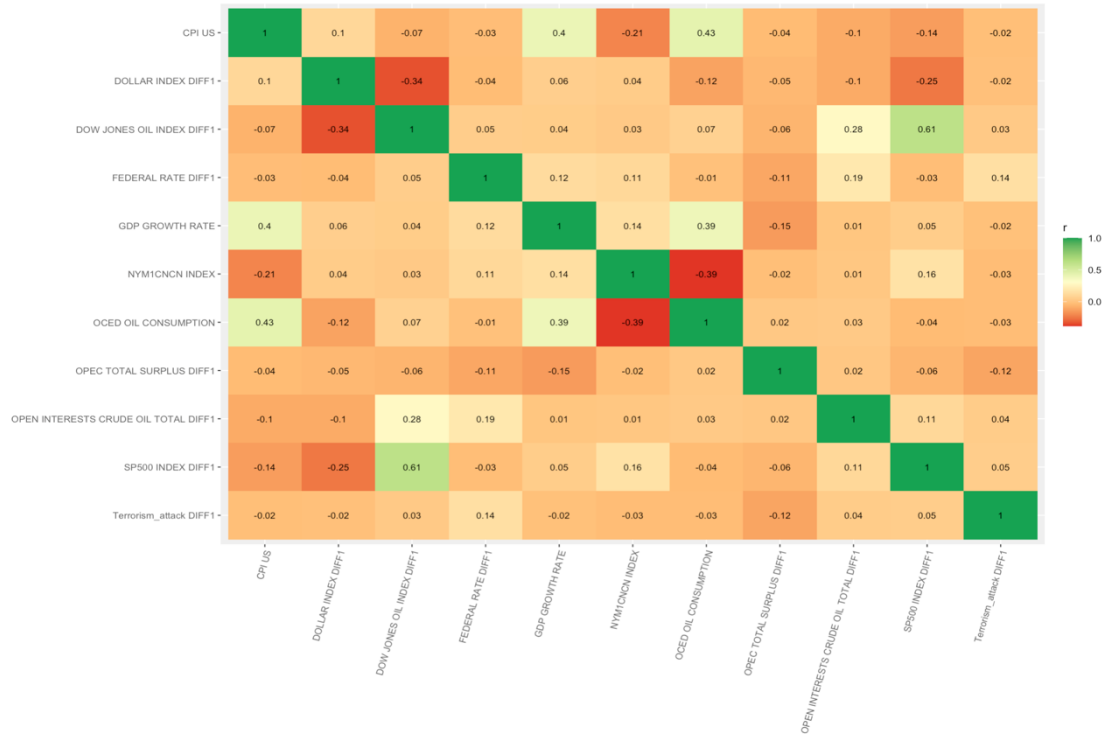

Figure 3: Correlation Matrix

Variables	VIF	Variables	VIF
OCED OIL CONSUMPTION	1.697254	NYM1CNCN INDEX	1.405055
OPEC TOTAL SURPLUS DIFF1	1.063096	DOLLAR INDEX DIFF1	1.188221
SP500 INDEX DIFF1	1.707549	GDP GROWTH RATE	1.566492
CPI US	1.440911	FEDERAL RATE DIFF1	1.106847
DOW JONES OIL INDEX DIFF1	1.850783	OPEN INTERESTS CRUDE OIL DIFF1	1.147305
Terrorism_attack DIFF1	1.04191		

Table 5: VIF of each variables

We can see the following pictures *(figure 4)* to see the performance and diagnostics of the model. The residual and fitted plot shows that the residual is around 0, also, the Quantile-quantile plot shows the residuals almost follow a normal distribution, which meets the assumption of the model. Also, by looking at the Scale-location plot, we found a nearly horizontal line in the plot, which suggests the model does meet the assumption of equal variance (homoscedasticity). The Residual vs Leverage plot does not show any extreme value at the upper right corner, which suggests that there is not any spot that can be influential to the regression line.

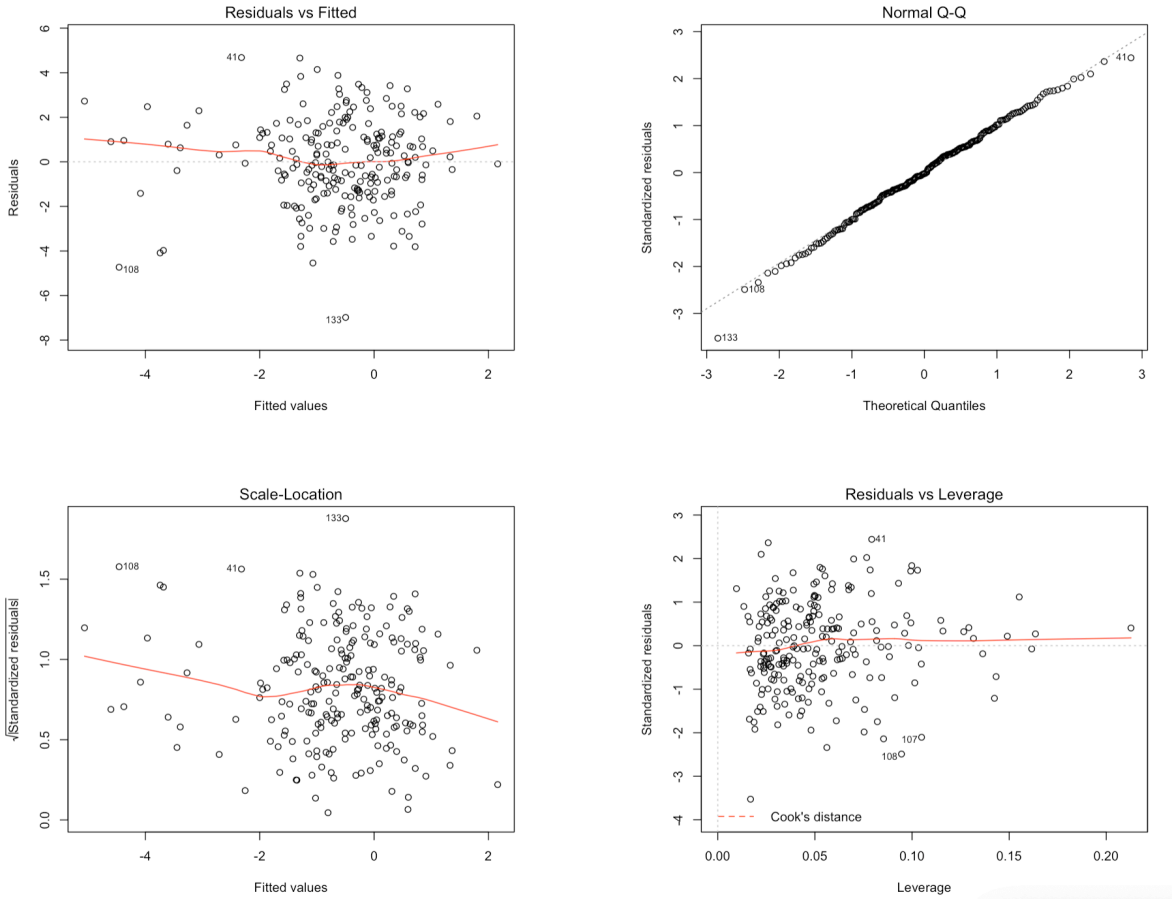

Figure 4: Model Diagnostics

4.3 Model Selection

In this section, we will first build up the full model by using all the explanatory variables by using the ordinary least square (OLS) method. Then the research will do the dimension reduction by using the backward Akaike information criterion (AIC) method and least absolute shrinkage and selection operator (LASSO) method. The section will also show the performance of each model by using the adjusted R square and the root means square error (RMSE).

4.3.1 OLS Full Model

The initial model is showing below:

$$quality = \beta_0 * x_1 + \beta_1 * x_2 + \cdots + \beta_{11} * x_{11} + \varepsilon \sim N(0, \sigma)$$

Where x_i is the stationary variables after the unit root test.
The following table show the summary of the OLS full model *(Table 6),* where the estimated is the coefficient for each explanatory variables. And another table *(Table 7)* shows the performance of the OLS full model.

	Estimated	Std. Error	Pr(>|t|)	Signif.
OCED OIL CONSUMPTION	6.535e-02	1.023e-01	0.639	
NYM1CNCN INDEX	1.733e-06	8.753e-07	1.980	*
OPEC TOTAL SURPLUS DIFF1	-5.160e-04	2.424e-04	-2.129	*
DOLLAR INDEX DIFF1	-1.223e-01	6.868e-02	-1.780	.
SP500 INDEX DIFF1	-1.595e-03	2.954e-03	-0.540	
GDP GROWTH RATE	3.250e-01	1.031e-01	3.151	**
CPI US	4.554e-01	1.247e-01	3.652	***
FEDERAL RATE DIFF1	-3.222e-01	3.490e-01	-0.923	
DOW JONES OIL INDEX DIFF1	1.275e-03	3.729e-03	0.342	
OPEN INTERESTS OIL DIFF1	-5.882e-06	2.277e-06	-2.584	*
Terrorism_attack DIFF1	1.999e-03	2.304e-03	0.867	

*Table 6: Regression Summary, Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1, Multiple R-squared: 0.2598, Adjusted R-squared: 0.222, F-statistic: 6.861, p-value: 6.978e-10*

	Adjusting R Square	MSE	RMSE
OLS FULL Model	0.222	3.989727	1.99743

Table 7: OLS Full model Performance

4.3.2 AIC Backward Selection Model

The model will delete one variable at one time, to achieve a better model with a lower AIC. Backward AIC final model through the stepwise-backward method eliminates multicollinearity problems. The selection process is showing as the following table *(Table 8):*

	Adj R-Square	AIC	RMSE
DOW JONES OIL INDEX DIFF1	0.2251	970.0052	1.9929
SP500 INDEX DIFF1	0.2281	968.1909	1.9892
OCED OIL CONSUMPTION	0.2301	966.6417	1.9866
Terrorism_attack DIFF1	0.2313	965.3361	1.9850
FEDERAL RATE DIFF1	0.2326	963.9847	1.9834

Table 8: AIC Backward Selection Process

The final model achieved by the stepwise backward method is as following table *(table 9)* with a lowest AIC = 963.9847. The estimated column show the estimated coefficients of the model. All the explanatory variables are linear significant. Moreover, the whole multi-regression model is in a significant level (F-statistics), and the adjusted R square is about 0.23, which means the model can explained 23% of variance of the WTI basis spread.

	Estimated	Std. Error	Pr(>|t|)	Signif.
NYM1CNCN INDEX	1.358e-06	7.771e-07	0.081997	.
OPEC TOTAL SURPLUS DIFF1	-5.044e-04	2.365e-04	0.034016	*
DOLLAR INDEX DIFF1	-1.273e-01	6.335e-02	0.045692	*
GDP GROWTH RATE	3.374e-01	9.352e-02	0.000382	***
CPI US	4.838e-01	1.194e-01	7.02e-05	***
OPEN INTERESTS OIL DIFF1	-5.976e-06	2.134e-06	0.005550	**

*Table 9: Regression Summary, Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1, Multiple R-squared: 0.2529, Adjusted R-squared: 0.2326, F-statistic: 12.41, p-value: 4.915e-12*

	Adjusting R Square	MSE	RMSE
AIC BACKWARD Model	0.2326	3.933876	1.9834

Table 10: AIC Model Performance

4.3.3 LASSO Model

The plot *(figure 5)* on the right shows that most of R square were explained for quite heavily shrunk coefficients. But at the end, a little bit increase in R square will cause huge growth of some variables. It can be seen as a signal of model overfitting. The figure *(figure 6)* also shows the model selection process of the LASSO regression, each lambda value will have the corresponding whole path of coefficients. The cross validation process will pick up the best model with lowest MSE. The model shown the relationship between the value of logarithm of lambda and the mean squared error (MSE) of the model. As a result, we can directly observe the best model is when the value of logarithm of lambda is near -3. The table *(table 11)* shows the coefficient of variables in LASSO model. We can see some variable have been eliminated.

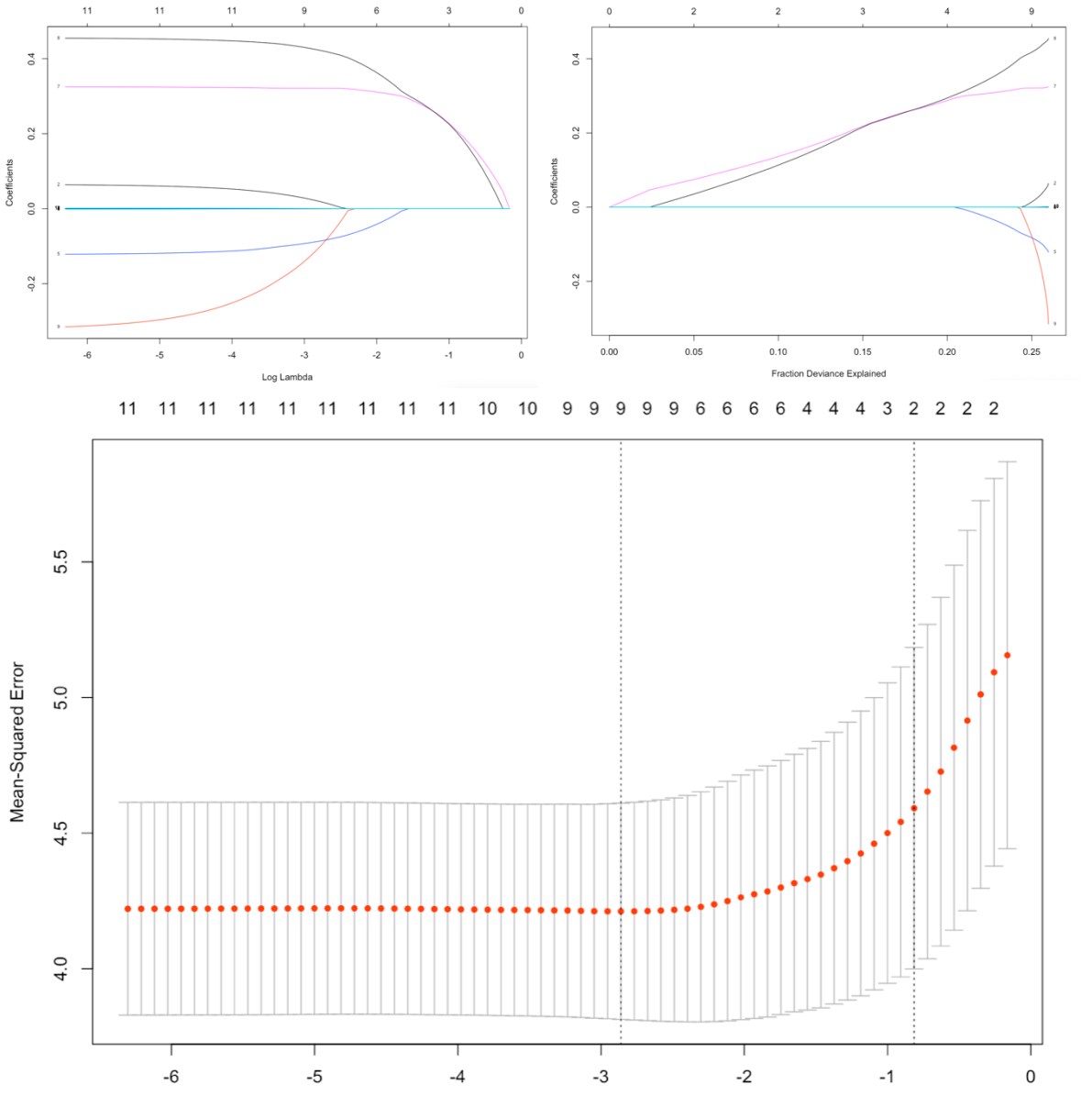

Figure 5: LASSO Model Selection

Variables	Estimated
OCED OIL CONSUMPTION	2.274374e-02
NYM1CNCN INDEX	1.071267e-06
OPEC TOTAL SURPLUS DIFF1	-4.130028e-04
DOLLAR INDEX DIFF1	-8.878033e-02
SP500 INDEX DIFF1	.
GDP GROWTH RATE	3.212513e-01
CPI US	4.252777e-01
FEDERAL RATE DIFF1	-1.162570e-01
DOW JONES OIL INDEX DIFF1	.
OPEN INTERESTS OIL DIFF1	-4.952636e-06
Terrorism_attack DIFF1	7.457815e-04

Table 11: LASSO Model Coefficients

The following table *(table 12)* shows the model performance of the LASSO regression.

	Adjusting R Square	MSE	RMSE
LASSO Model	0.2560962	3.813779	1.95289

Table 12: LASSO Model Performance

5 DISSUSION

By comparing the adjusted R square and the root mean square error (RMSE), we can see the performance of each model. The model with a higher adjusting R square can explain more variability of the response variable. By comparing three different models, the LASSO model has the highest adjusting R square equal to 0.256, which indicates it can explain almost 26% of the variability of the basis spread between the WTI crude oil spot price and WTI 4-month futures price. Moreover, the model with a lower root mean square error means the model is accurate. By comparing the RMSE, the LASSO model also has the lowest RMSE equal to 1.9529. As a result, the LASSO model has the lowest adjusting R square and lowest RMSE, we should choose the LASSO model among these three to predict the basis spread in the future.

	Adjusting R Square	MSE	RMSE
OLS FULL Model	0.222	3.989727	1.9974
AIC BACKWARD Model	0.233	3.933876	1.9834
LASSO Model	0.256	3.813779	1.9529

Table 12: Model Comparison

Following figure *(figure 6)* show the time series of predicted basis spread by using LASSO model and the real basis spread. We can see the predicted basis is less fluctuated than the real basis spread. This is fair enough because the LASSO model only explains about 26% of the variability of real basis. Because the original data set are time-series data, by removing the random effect, all the variables used in the predicted models are stationary. As a result, it is hard to have an adjusted R square which is higher than 0.7, actually, according the empirical rule of other previous time-series research, a predicted model with an adjusted R square higher than 25% can be considered as a quite good model.

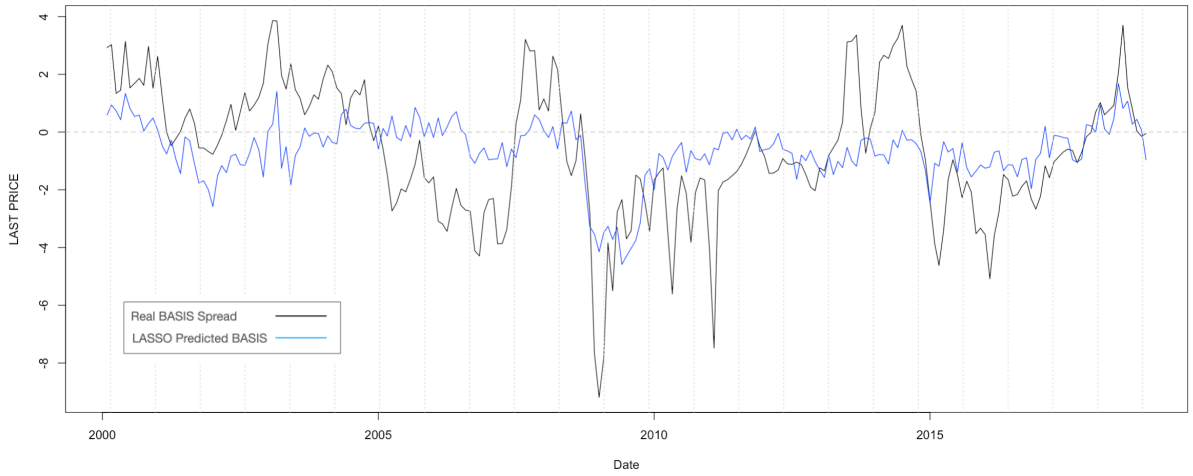

Figure 6: Observed vs Predicted

Moreover, by looking close to the figure, we can find the LASSO regression model does not do well in predicting the relatively large spread. Starting at the beginning, the real basis spread and predicted basis is moving in the roughly same direction at the same time. However, at around 2003 - 2004, the predicted basis start to move behind the real basis spread and fluctuate around 0 after 2005. At this period, the predicted model becomes inaccurate, even sometimes the predicted basis spread move in the opposite direction of the real basis. Around the 2008 financial crisis, the movement of both is highly correlated. Later it becomes inaccurate and unregulated again. Then around the 2014 financial crisis, the movement of both basis is similar again. There is three main financial crisis in the 21 century. The first time is the 2001–2002 Argentine Economic Crisis， the second time is the 2007–2009 Global Financial Crisis, and the last time is the 2014 Russian Financial Crisis. The figure shows the highly similar co-movement of the real basis and predicted basis.

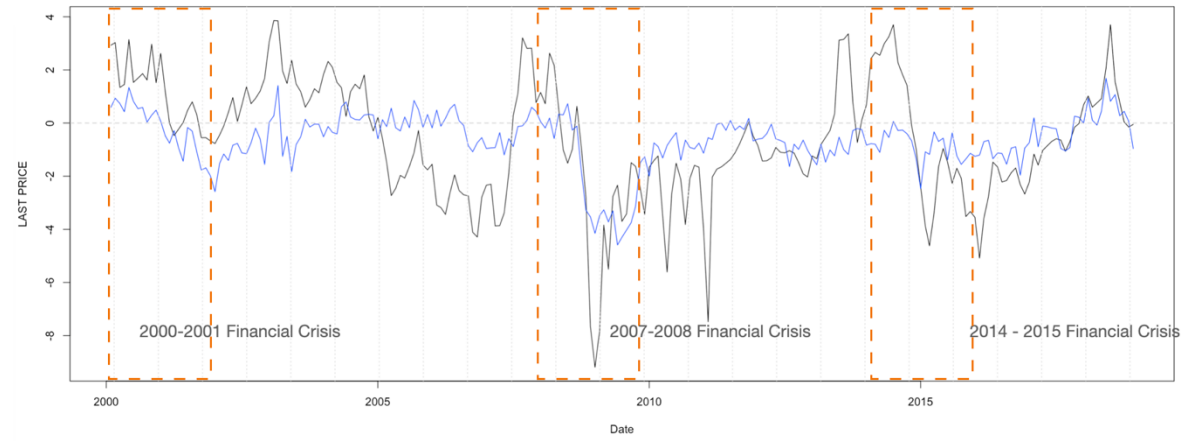

Figure 7: Observed vs Predicted

One possible reason is that most of the significant explanatory variables can be easily affected by the financial crisis and put these influences in the same direction of the basis. Moreover, basis spread has a large uncertainty, it depends on both the WTI crude oil spot market and the WTI crude oil futures market, some of the variability cannot be explained by the linear model. As a result, the predicted basis becomes more accurate and reasonable if there is a large event that happened in the market, such as the financial crisis. Whereas, when there are not influential events happening in the market, the predicted basis become more like random work.

6 CONCLUSION

In this research, we considered the 11 different variables from a different aspect, including demand factors, supply factors, economic condition, financial market performance, and the dollar strength, to predict the basis spread between WTI OK Crushing crude oil sport price and the 4-month WTI OK Crushing crude oil futures price. We start with the full multi-linear regression model to predict the basis spread and then do the feature reduction to reduce the effect of overfitting. The final result shows that the LASSO model has the best model performance, which can explain 26% of the basis variability with the lowest root mean square error. The adjusting R square is a little bit higher than the empirical value of 25% for times-series and

stationary explanatory variables. As a result, the LASSO regression model can be considered as a quite good model. By comparing the predicted basis and real basis, we have seen that the predictive model is powerful to predict the movement direction or even the basis spread when there is a significant event in the world market. The time period during the three main financial crises in 21 century shows strong evidence that the predicted model is more accurate when there are influential events. For the other side, it more likely to indicate that the WTI crude oil basis spread is predictable when there is an influential event. However, if there are not influential factors in the market, the predicted basis spread becomes inaccurate and more like random work. Due to COVID-19, the financial market has been suffered for some time. As a result, further research can conduct more research about how basis of crude oil is driven by the COVID-19. In this research, we used the data ranged from January 2000 to December 2018 excluding the driven power by COVID-19. However, by applying the LASSO predictive model to the data during COVID-19, we have found the huge error and smallest adjusting R square. It may indicate the further research should separate their data and observe the COVID-19' effect respectively. Moreover, further research also can be conducted to use the non-linear way to predict the basis of WTI crude oil.

7 REFERENCE:

Ajifowoke, M. G. (2020). Weekly economic index: MPC raises cash reserve requirement, coronavirus affects oil prices.

Algieri, B. (2014). The influence of biofuels, economic and financial factors on daily returns of commodity futures prices. Energy Policy, 69, 227–247. https://doi.org/10.1016/j.enpol.2014.02.020

Allen, M. P. (Ed.). (1997). The problem of multicollinearity. In Understanding Regression Analysis (pp. 176–180). Springer US. https://doi.org/10.1007/978-0-585-25657-3_37

Amendola, A., Amendola, A., Candila, V., Candila, V., Scognamillo, A., & Scognamillo, A. (2017). On the influence of US monetary policy on crude oil price volatility. Empirical Economics, 52(1), 155-178. doi:10.1007/s00181-016-1069-5

An, H., Gao, X., Fang, W., Ding, Y., & Zhong, W. (2014). Research on patterns in the fluctuation of the co-movement between crude oil futures and spot prices: A complex network approach. Applied Energy, 136, 1067–1075. https://doi.org/10.1016/j.apenergy.2014.07.081

Basistha, A., & Kurov, A. (2015). The impact of monetary policy surprises on energy prices. The Journal of Futures Markets, 35(1), 87-103. doi:10.1002/fut.21639

Bildirici, M., Guler Bayazit, N., & Ucan, Y. (2020). Analyzing Crude Oil Prices under the Impact of COVID-19 by Using LSTARGARCHLSTM. Energies, 13(11), 2980.

Bu, H. (2011). Price Dynamics and Speculators in Crude Oil Futures Market. Systems Engineering Procedia, 2, 114–121. https://doi.org/10.1016/j.sepro.2011.10.014

Frondel, M., & Horvath, M. (2019). The U.S. fracking boom: Impact on oil prices. The Energy Journal (Cambridge, Mass.), 40(4), 191. doi:10.5547/01956574.40.4.mfro

Hanck, C., & Czudaj, R. (2015). Nonstationary-volatility robust panel unit root tests and the great moderation. AStA Advances in Statistical Analysis, 99(2), 161–187. Retrieved from: https://doi.org/10.1007/s10182-014-0235-3

Haushalter, G. D. (2000). Financing Policy, Basis Risk, and Corporate Hedging: Evidence from Oil and Gas Producers. The Journal of Finance, 55(1), 107–152. https://doi.org/10.1111/0022-1082.00202

Hill, R., Griffiths, W., Lim, G. (2017). Principles of Econometrics, 5Th Edition. Chapter 12: Regression With Time-Series Data: Non-stationary Data. LCCN 2017056927 (eBook).

Holwerda, D., & Scholtens, B. (2016). The financial impact of terrorist attacks on the value of the oil and gas industry: An international review. (pp. 69-80). Cham: Springer International Publishing. doi:10.1007/978-3-319-32268-1_5

Homepage—U.S. Energy Information Administration (EIA). (2020). Retrieved November 22, 2020, from https://www.eia.gov/index.php

Jones, C. M., & Kaul, G. (1996). Oil and the Stock Markets. The Journal of Finance, 51(2), 463–491. https://doi.org/10.1111/j.1540-6261.1996.tb02691.x

Kaufmann, R. K., & Ullman, B. (2009). Oil prices, speculation, and fundamentals: Interpreting causal relations among spot and futures prices. Energy Economics, 31(4), 550–558. https://doi.org/10.1016/j.eneco.2009.01.013

Kollias, C., Kyrtsou, C., & Papadamou, S. (2013). The effects of terrorism and war on the oil price-stock index relationship. Energy Economics, 40, 743–743. https://doi.org/10.1016/j.eneco.2013.09.006

Miao, H., Ramchander, S., Wang, T., & Yang, D. (2017). Influential factors in crude oil price forecasting. Energy Economics, 68, 77–88. https://doi.org/10.1016/j.eneco.2017.09.010

Mordkoff, J. T. (n.d.). The Assumption(s) of Normality. 6.

Pham, C. S. (2018). Multiple regression model for cotton price returns: Analysis of the impact of weather, oil price return, and China's economy. https://aaltodoc.aalto.fi:443/handle/123456789/33972

Rahim, Abd. (2020). Re: What is Cointegration?. Retrieved from: https://www.researchgate.net/post/what_is_Cointegration/5f3509c84275c25ce56ab139/citation/download.

Sadorsky, P. (1999). Oil price shocks and stock market activity. Energy Economics, 21(5), 449–469. https://doi.org/10.1016/S0140-9883(99)00020-1

Sarwat, S., Kashif, M., Aqil, M., & Ahmed, F. (2019). determination of causality in prices of crude oil. International Journal of Energy Economics and Policy, 9(4), 298-304. doi:10.32479/ijeep.7724

Stoll, H. R., & Whaley, R. E. (2015). Commodity Index Investing and Commodity Futures Prices (SSRN Scholarly Paper ID 2693084). Social Science Research Network. https://papers.ssrn.com/abstract=2693084

Tufféry, S. (2011). Data Mining and Statistics for Decision Making. John Wiley & Sons, Incorporated. Retrieved from: http://ebookcentral.proquest.com/lib/kean/detail.action?docID=792450

Wang, Y., & Wu, C. (2012). Energy prices and exchange rates of the U.S. dollar: Further evidence from linear and nonlinear causality analysis. Economic Modelling, 29(6), 2289–2297. https://doi.org/10.1016/j.econmod.2012.07.005