Crude Oil Price Prediction Based on Multiple Linear Regression Model

Ziding Yuan*
[*] Corresponding author: zccazyu@ucl.ac.uk

University College London, Department of chemistry London, the UK

Abstract—The Russian-Ukraine conflict has had a significant impact on the world economy. This effect is mainly reflected in the energy market as the decisive role of Russia in fossil energy export, bringing a massive change to the price of the world energy market. This paper uses data from mainly EIA to build a multiple linear regression model and focuses on predicting the price of crude oil. The five independent variables chosen to create a multiple linear regression model are time supply, demand, CPI, and GDP. From time to time, the price of crude oil changes on account of many other reasons. Among various cause, supply and demand are two main reasons known to affect the price of crude oil. Thus, the paper studies the correlation of the price of selected target with CPI and GDP and also provides information on the possible future crude oil price. The price fluctuation of crude oil can significantly affect the market related to crude oil, for example, transporting sectors and some manufacturing sectors. The profit of these two sectors is closely associated with the price of crude oil. Therefore, the correct prediction of crude oil can help them suffer less loss during these times of the Russian-Ukraine conflict. These results shed light on guiding further exploration of oil price prediction.

Keywords-Multiple linear regression; price prediction; R-square test.

1 INTRODUCTION

Contemporaneity, the price of futures of energy resources proliferates. The price of energy resources is an essential feature of the world economy. It provides information for people in the manufacturing and transporting sectors and policymakers in governments and central banks. The oil price is a crucial factor in countries' budgets in their revenues for the OPEC. They mainly forecast oil prices by building a consensus forecast [1]. The United States ranks first in consuming crude oil and consumes 19,687,287 Barrels per day [2]. Therefore, it has a significant effect on all parts of the economy of the United States, including agricultural employment [3]. Furthermore, WTI is the primary source of oil in the U. S. Therefore, this paper will focus on the price of crude oil of WTI.

The linear regression model is an effective prediction model to predict the future price of futures. It illustrates the linear correlation between the price and the independent variable (those factors that affect the price, such as time and consumer price index). The extension considering multiple factors together is regarded as a multiple linear regression model (MLR), an addition of a linear regression model to model the linear relationship between several independent variables and one dependent variable [4]. Lu et al. summarizes some of relevant results recently [5]. Chen et al. compared three decision tree models (i.e., CART, M5P, and Random forest), with various

FFIT 2022, October 28-30, Shenzhen, People's Republic of China

DOI 10.4108/eai.28-10-2022.2328443

metrics [6]. Hao et al. accurately forecasted the price of natural oil, where avoiding of overfitting improves their forecasting performance. To be specific, the EWMC and MLR are applied, which is the same as this paper, to predict accurate oil prices [7]. Bwo-NungHuang et al. focused on the interaction between the spot and future price of the selected target and divided the data into three regimes based on the size of the databases. They explain investors' behavior using the VECM model in arbitrage and non-arbitrage commands [8]. Nalini Gupta uses an artificial neural network (ANN) to explain the correlation between factors affecting crude oil prices. The advantages and disadvantages of the ANN method were also discussed in that paper. Therefore, this idea was considered and affected the decision to choose factors in this paper's linear regression model [9].

In this paper, a multiple linear regression model is constructed to forecast the crude oil price. The rest part of the paper is organized as follows. The Sec. II will introduce the data origination, regression models as well as evaluation metrics. Subsequently, the results will be presented in Sec. III will analysis and discussion accordingly. Eventually, a brief summary will be given in Sec. IV.

2 METHODOLOGY

2.1 Data

The crude oil price index used in this paper is WTI, which is one of the three main benchmarks for the price of oil. WTI is a light sweet oil because it has low density (more miniature carbon chain) and contains 0.34% of sulfur [10]. Therefore, the relation between the price of crude oil with factors(x) was assumed to be the same for WTI and all other oil throughout the paper.

2.1.1 Historical Data Source

Both independent variable and dependent variable data were selected from 2011.1.1 to 2022.1.1 (11 years' data). For the dependent variable, data on WTI crude oil price came from investment.com. With regard to independent variable: CPI data for the U.S. comes from Fred's economic data (https://fred.stlouisfed.org/series/CPALTT01USM657N). Data for demand and supply of crude oil come from EIA (U.S. Energy information administration). The world bank provided U.S. GDP data.

2.1.2 Future data prediction

US GDP prediction gives 2.5% in 2022, 2.3% in 2023, 2.0% in 2024 from official website (https://www.philadelphiafed.org/surveys-and-data/real-time-data-research/spf-q2-2022) [11]. Demand for crude oil increases 3.01% in 2022, 1.81 in 2023, 1.09% in 2024 according to statistics (https://www.statista.com/statistics/271823/daily-global-crude-oil-demand-since-2006/). Data on crude oil Production came from a paper written by Admin: 0.78% increase in 2022, 7.22% increase in 2023[12]. US CPI was collected on (https://tradingeconomics.com) for 7.00% 2022, 1.90% 2023

2.2 Models

In general, the equation of a single linear regression model can be simply given as:

$$y = \beta 0 \ + \ \beta 1X1 \ + \ u \tag{1}$$

The value of B1 equals changes in y divided by change in x. However, if u and x are correlated, the conclusion is useless. As a matter of fact, it is necessary to make another assumption “mean-independence” assumption for all values of x. This means the mean of u for” each slice of the population” is constant. However, u and x are still allowed to be mutually dependent, and in this case, price won’t simply be related to a non-related independent variable. Therefore:

$$E\left[\frac{y}{x}\right] = E[\beta 0 + \beta 1x \ + \ x] = \beta 0 \ + \beta 1x \ + \ E[u|x] \tag{2}$$

Substituting E[u|x] = 0

$$E\ [y|x] = \ \beta 0 \ + \ \beta 1x \tag{3}$$

A diagram is given in Fig. 1, Where the small circle on the line is the mean value of x and they are linearly related to y.

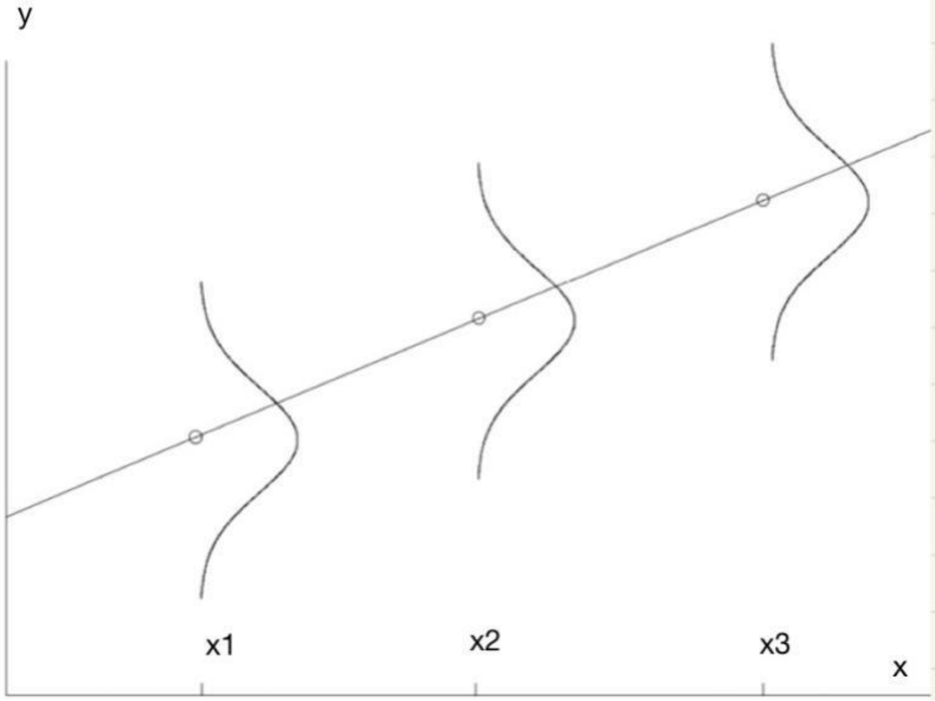

Figure 1. Linear relation of y to mean value of x.

2.3 Multiple regression

It is obvious that in this case price is not correlated by one factor only, i.e., E[u |Time] is not constant. Therefore, one needs to consider other factor for example exchange rate of Ruble to US dollar:

$$Log(price) = \ \beta 0 \ + \ \beta 1Time \ + \ \beta 2Exchange\ rate \ + \ u \tag{4}$$

where B2 is the change of price to change of exchange rate. Same for other factors, we can derive the multiple regression model as:

$$Log(price) = B0 + B1x1 + B2x2 + B3x3 + \cdots + u \tag{5}$$

Here, x1, x2, x3… are all factors linearly related to price of futures.

2.4 Calculate B0 and B1 using Ordinary Least Square (OLS) Estimates.

Assuming that E[u|x1, x2,…,xk] = 0 and using OLS as minimized of sum of squared residuals(SSR):

$$\mathrm{SSR} = \sum_{i=1}^{n} u^2 = \sum_{i=1}^{n} (y - \beta 0 - \beta 1 * x1 - \beta 2x2)^2 \tag{6}$$

For first-order conditions:

$$\frac{\delta SSR}{\delta \beta 0} = 2\sum_{i=1}^{n} ui = 0 \tag{7}$$

$$\frac{\delta SSR}{\delta \beta 1} = 2\sum_{i=1}^{n} x1ui = 0 \tag{8}$$

$$\frac{\delta SSR}{\delta \beta 2} = 2\sum_{i=1}^{n} x2ui = 0 \tag{9}$$

By solving these three differential equations we can get $\beta 0, \beta 1$ and $\beta 2$ Other coefficient B can be calculated in the same way.

2.5 R-square test

When deciding goodness-of-fit, total sum of squares must be considered and thus:

$$SST = SSE + SSR \tag{10}$$

Where SSE and SSR are explained and residual sum of squares.

R^2 test decides how good the data fits the model and R-square is defined as:

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST} \tag{11}$$

3 RESULTS & DISCUSSION

3.1 Results

Based on the multi-linear regression model, this paper investigates the correlations and impacts of the relevant factors and features on the oil price. Using data from Fred’s economic data and performed multiple linear regression with 11-year monthly data between 2011.1.1-2022.1.1 to get coefficients that illustrate the correlation between variables and thus predict future price. R2 method was used to grade the model. The results are given in Table. I and Table. II, where β1 is the coefficient of time, β2 is the coefficient of US CPI, β3 is the coefficient of demand, β4 is the coefficient of supply and β5 is the coefficient of US GDP

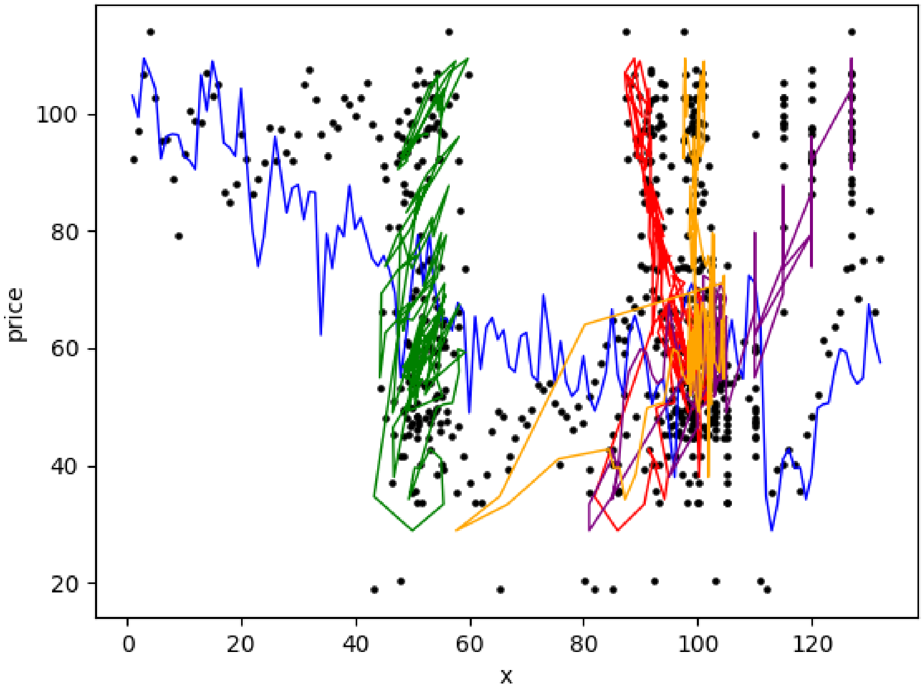

Figure 2. The relation between price of crude oil and independent variables Linear relation of y to mean value of x.

Table 1 Coefficient of Multiple Linear Regression model

	β1	β2	β3	β4	β5	u
crude oil	0.2759	15.60	-0.590	187.9	1.728	-175.8

Table 2 Ranking based on R2 test and price prediction

	R^2
crude oil	63.57%

3.2 Discussion

As demonstrated in the Fig/ 2, the factors are normalized to 100 as the base except for time and US CPI for a clearer view. GreenLine (US CPI) plays a significant and positively correlated role in the price of crude oil and has a coefficient of 15.60. The Red line(demand) shows that the demand for crude oil is inelastic. When the price per gallon changes from 30 to over 100, the demand only slightly decreases. That is reasonable. People who have a car will drive to markets and work regardless of the price of gasoline, which is the same for everyone, and other conditions of using oil, including the manufacturing side. The coefficient of -0.590 proves the elasticity of demand for crude oil and shows the negative correlation between price and demand.

In general, supply is more inelastic than demand, which is also true in this case. The purple line shows the supply curve, and it has a relatively low gradient indicating that it is more inelastic than other goods. That is mainly because the price of machines that extract oil is so high [13]. Therefore, it takes years for oil companies to develop and purchase new equipment and thus inelastic crude oil prices.

US GDP has a coefficient of 1.728. It is positively correlated and more prominent than 1, meaning a change in oil price will result in a more significant change in US GDP. According to American Petroleum Institute, America’s gas and oil contribute nearly 8% of US GDP.

Last but not least, the blue line, which has a coefficient of 0.2759, shows that the price trend over time is dramatic. As shown in Fig/2, it does not increase linearly due to less correlation between supply and demand. Instead, it was mainly caused by changes in government policies of OPEC countries. Government policy changes often, and a slight policy change can cause a tremendous change in price.

3.3 Limitations

Nevertheless, it should be noted that the results in this paper have some defects and shortcoming. The rest factors affecting the price of crude oil (u) are significant in this case, and R2 is only 63.57%, meaning that there is still 38% of variation unexplained in this paper. That is because regulations and other factors cannot be quantized and, therefore, not included in the multiple linear regression model.

Negative u means Some of the five selected factors are correlated and double counts some factors, which means an even more unexplained part of the price of crude oil using this multiple linear regression model. That might be solved by building another linear regression model to illustrate their relations.

Another limitation is that in R-square test falls unless the coefficient of the new variable is identically zero. Therefore, the R-square test may still consider good fitness for some non-related or even negatively related independent variable x since R2 never decreases.

4 CONCLUSION

In summary, this paper investigates the effects of five factors on the crude oil price, including time, supply, demand, price level (CPI), and GDP, based on a multiple linear regression model. They explicitly predict the future price using a multiple linear regression model and grading it using the R-square method. According to the analysis, US GPD and CPI were primarily correlated to the price of crude oil. Moreover, supply and demand are critical factors in the price of crude oil. Nevertheless, the residual part u is negative, meaning these four factors are correlated themselves. With more data used, the correlation between these four factors can be calculated in the future. Thus, an even more accurate coefficient will be calculated, and more possible price predictions will be generated. Overall, these results offer a guideline for forecasts of future prices and provide information for mainly people working in the manufacturing and transporting sectors.

REFERENCES

[1] A. Y. Mikhaylov, "Pricing in oil market and using probit model for analysis of stock market effects.", Diss of ZBW university, 2018.

[2] United states oil, available at: https://www.worldometers.info/oil/us-oil/ Accessed May. 24, 2022

[3] N. D. Uri, Natural Resources and Environment Division, Economic Research Service, US Department of Agriculture, 1301 New York Avenue, NW, Washington, DC 20005, USA

[4] D. A. Freedman, Statistical Models: Theory and Practice. Cambridge University Press, 2009.

[5] H. Lu, et al. "Energy price prediction using data-driven models: A decade review." Computer Science Review vol. 39, 2021, 100356.

[6] E. Chen, and X. James. "Crude oil price prediction with decision tree based regression approach." Journal of International Technology and Information Management vol. 27.4, 2019, pp. 2-16.

[7] X. Hao, Y. Zhao, and Y. Wang. "Forecasting the real prices of crude oil using robust regression models with regularization constraints." Energy Economics vol. 86, 2020, 104683.

[8] B. N, Huang, C. W. Yang, and M. J. Hwang. "The dynamics of a nonlinear relationship between crude oil spot and futures prices: A multivariate threshold regression approach." Energy economics vol. 31.1, 2009, pp. 91-98.

[9] Nalini Gupta, Shobhit Nigam, Crude Oil Price Prediction using Artificial Neural Network, Procedia Computer Science, Volume 170, 2020, Pages 642-647, ISSN 1877-0509,

[10] Energy Insights for WTI, available at: https://www.mckinseyenergyinsights.com/resources/refinery-reference-desk/wti/, "WTI." Accessed Dec. 20, 2020.

[11] M. Su, et al. "Data-driven natural gas spot price forecasting with least squares regression boosting algorithm." Energies vol. 12.6, 2019, 1094.

[12] Admin. U.S. crude oil production forecast to rise in 2022 and 2023 to record high levels. 2022.

[13] The Low Elasticity of Supply, available at: https://www.investopedia.com/ask/answers/040915/how-does-law-supply-and-demand-affect-oil-industry.asp#:~:text=Oil%20has%20a%20low%20elasticity,initially%20set%20up%20oil%20extraction.