ISSN: 2688-9323

Analysis Multiple Regression Prediction Model for Crude Oil Spot Price Driven by Data

Runnan Wang

School of Management, Shanghai University, Shanghai 200444, China

Abstract

Crude oil plays an important role in the global economy. Accurate crude oil price forecast not only provides reference for market investors and regulators to make decisions, but also is conducive to long-term stable development of crude oil market. To make the price of crude oil analysis and prediction is more reliable, this paper not only consider the crude oil price time series of their own law of development, and join in the model in supply, demand, inventory, market, technical indicators such as the price of crude oil influence factors, by data preprocessing and inspection, such as Hodrick Prescott (HP) filter, linear test, correlation analysis, multiple regression analysis method is adopted to establish the crude oil price forecasting model, An empirical analysis of West Texas Intermediate crude oil (WTI) verifies the feasibility and effectiveness of the proposed method.

Keywords

Crude Oil Price Prediction; Data-driven; Multiple Regression.

1. Introduction

Crude oil plays an important role in the global economy [1]. According to world Energy Statistical Review data in 2016, crude oil accounts for 32.9% of global primary energy consumption [2]. As a key energy financial derivative, spot crude oil prices change due to many unstable factors such as overall economic conditions, oil demand and supply, speculative trading and geopolitical conflicts. The fluctuation of oil price will inevitably have a wide impact on the economic development of the whole world, the energy security of each country and the survival and development of oil enterprises. In recent years, the international oil price has fluctuated greatly, and its changing characteristics are extremely complex. In this volatile oil market, in order to better safeguard their own interests, the first task is to accurately grasp the law of international crude oil price changes, explain the reasons for the change of crude oil price, and predict its trend. Only by adjusting oil production planning and grasping oil consumption level under the condition of full understanding of oil price can we reduce the negative impact of oil boom or sudden drop on economy and safeguard the interests of the country and the group. Therefore, accurate price forecast not only provides decision-making reference for market investors and regulators, but also contributes to the long-term stable development of the crude oil market. However, the non-stationary, nonlinear and multi-frequency characteristics of crude oil price series pose certain challenges to accurate price prediction [3].

At present, there are many methods for quantitative analysis and prediction of oil price at home and abroad, which can be divided into two categories: one is to predict future oil price from historical oil price, that is, from oil price to oil price prediction, including ARIMA-GARCH series model, wavelet analysis, empirical mode decomposition, fractal theory, fuzzy pattern matching, etc. [4-6]; The other is to predict future oil prices by influencing factors, including regression model, multiple time series model, artificial neural network, system dynamics model, etc. [7-10].

Based on the advantages and disadvantages of various oil price forecasting methods, this paper not only considers the development law of oil price time series itself, but also adds the main influencing factors of oil price into the model and adopts the prediction method combining Hodrick Prescott (HP) filter and multiple regression. Firstly, the oil price sequence is processed by HP filter to extract the fluctuation term and eliminate the trend term. Meanwhile, the influencing factor sequence is extracted by HP filter. Since the filtered sequence of influencing factors can better reflect the change rule of corresponding oil price sequence, multiple regression is carried out between the sequence of oil price fluctuation and the sequence of influencing factors, and the prediction model of crude oil price pre-multiple regression is established, and the final prediction result of oil price is obtained.

2. Methods

2.1. Multiple Regression Analysis

Regression analysis is the most mature and commonly used statistical tool in statistics. It is a basic quantitative technique to analyze the relationship between variables. In econometrics, if the regression function describes the linear relationship between one explained variable and multiple explanatory variables, the regression function set thus is the multiple linear regression model [11]. Generally, the multiple linear regression function including explained variables Y and k-1 explanatory variables X_1, X_2, ⋯, X_k is in the form of:

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_k X_k + \varepsilon \tag{1}$$

where β_j (j=1,2, ..., k) is the parameter of the model, ε is the random error disturbance term, and k-1 is the number of variables.

In the linear regression function, each regression coefficient is unknown and can only be estimated by using the sample observation value. The main problem to be solved by multiple linear regression analysis is how to estimate the parameters in the regression model according to the observation value of the variable sample and conduct statistical tests on the estimated parameters and the overall characteristics of the regression equation, and finally use the obtained regression model that all the tests are qualified to analyze and predict the research object. Under the condition that the model meets the basic assumptions (zero mean, homoscedasticity, no autocorrelation, no correlation between random perturbation terms and explanatory variables, no multicollinearity, normality), ordinary least square method (OLS) can be used to obtain the unbiased, efficient and consistent estimation of multiple linear regression models.

The test of the multiple linear regression model with estimated parameters is mainly about the goodness of fit of the estimated model, the significance of each parameter in the model and the significance of the whole regression equation, in addition to the test of whether the assumed conditions are met. The sample determination coefficient R^2 can be used to calculate the goodness of fit of the sample regression equation. The larger R^2 ($0 \le R^2 \le 1$) is, the better the fit of the regression equation and the sample value is. Otherwise, the fitting of regression equation and sample is poor. In addition, the significance test of the regression equation (F test) and the significance test of the estimators of each parameter in the model (t test) are also required. In addition, it is also necessary to conduct ADF unit root test on the residual sequence of the regression equation. If the residual is non-stationary, it indicates that the change of the dependent variable except the part that can be explained by the independent variable is still irregular and cannot be used to predict future information, which is called pseudo regression.

2.2. Hodrick Prescott (HP) Filter

Let Y_t be an economic time series containing trend components and fluctuation components, where $\{Y_t^T\}$ is the trend component, $\{Y_t^C\}$ is the fluctuation component. The calculation of HP filter is to separate Y_t^T from Y_t and extract the wave component Y_t^C. The output gap formula is: $c=\frac{(Y_t-Y_t^T)}{Y_t^T}$, where the potential output Y_t^T is obtained by minimizing the loss function, that is, $\min\left\{\sum_{t-1}^{T}(Y_t-Y_t^T)^2+\lambda\sum\left[(Y_{t-1}^T-Y_t^T)-(Y_t^T-Y_{t-1}^T)\right]^2\right\}$. After the potential output is obtained by HP filtering method, the output gap is the difference between actual output and potential output divided by potential output. HP filtering depends on parameter λ, and different values of this parameter can obtain trend sequences of different smoothness degrees, where λ=25. HP filtering analysis is performed on empirical data respectively.

3. Analysis of Influencing Factors of Crude Oil Price

Table 1. Internal and external variables influencing the crude oil market

First class index	Second class index	Variables	Unit	Symbols	Data source
/	/	WTI spot price	Dollar per Barrel	WTI	EIA
Supply	Production	Crude Oil Production, World	Million Barrels per Day	X_1	EIA
Demand	Developed country	Petroleum Consumption, OECD	Thousand Barrels per Day	X_2	EIA
	Developing country	China oil import	Million tons	X_3	Wind
	Global economic development	US: CPI index	/	X_4	Wind
		US: CPI: energy	/	X_5	Wind
Inventory	The U.S.	Crude Oil Stocks, Total	Million Barrels	X_6	EIA
External market	Monetary	Exchange rate of EUR against USD	/	X_7	Wind
	Stock market	NASDAQ index S	/	X_8	Wind
	Commodity market	COMEX: Gold: Future closing price	Dollar per oz	X_9	Wind
Technical index	Spread	Crude oil spot price spread: WTI-Brent	Dollar	X_{10}	EIA

There are many factors causing the fluctuation of international crude oil price [12]. Changes in supply and demand will directly affect the price of crude oil, such as international crude oil production and the original demand and consumption of each country, including the existing crude oil stocks of each country. Crude oil commercial inventory is an important reflection of supply and demand. In addition, the fluctuation of crude oil price is influenced by market situation and market trading, including money market, stock market and metal market. For a long time, most international oil trade is priced and settled in US dollars, so the fluctuation of US dollar exchange rate has become one of the important factors leading to the fluctuation of international oil prices. The fluctuation of dollar exchange rate leads to the change of purchasing power of petrodollars. Therefore, considering purely the valuation factor, the depreciation of dollar is bound to push up the oil price, and the appreciation of dollar may bring down the oil price to a certain extent. Speculative capital is an important participant in the global crude oil market and has a significant influence on the fluctuation of crude oil price. As a technical index, the price difference of different types of crude oil also has a certain effect on the fluctuation of crude oil price. Based on the above analysis, the main influencing factors of international oil price selected in this paper and their correlation coefficients with oil price are shown in Table 1.

4. Empirical Study

4.1. Data Description and Preprocessing

4.1.1. Data Description

As an empirical example, monthly data of WTI price and its main influencing factors from January 1997 to December 2016 are used as training data of the model, and data from January 2017 to December 2019 are used as test data of the established model. Data were obtained from the US Energy Administration (EIA), the Federal Reserve website and Wind database.

4.1.2. Data Preprocessing

To eliminate the influence of multicollinearity, HP filtering is applied to the data in this paper to remove the trend term and extract the fluctuation term. After data processing, correlation coefficients of all variables were analyzed, as shown in Table 2. Compared with before data processing, correlation coefficients were generally lower, and the influence of multicollinearity was reduced to a large extent. To further test the multicollinearity problem, OLS estimation was used to conduct variance inflation factor (VIF) test to further judge whether the regression result was still affected by multicollinearity. The results show that the mean value of variance inflation factor is 1.54, and the maximum value is 2.29, which can basically exclude the influence of multicollinearity. ADF stationarity test was conducted on the variables filtered by HP, and it was found that the sequences in this paper were all stationarity sequences

Table 2. Correlation coefficient between variables after HP filtering

	WTI	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
WTI	1										
X_1	0.133	1									
X_2	-0.013	0.2559	1								
X_3	0.001	-0.128	0.001	1							
X_4	0.800	0.1487	-0.044	0.0116	1						
X_5	0.867	0.1833	0.009	0.0228	0.949	1					
X_6	-0.417	-0.192	-0.278	-0.074	-0.494	-0.492	1				
X_7	-0.628	-0.014	-0.118	-0.066	-0.388	-0.513	0.394	1			
X_8	0.374	0.2027	0.151	0.0851	0.188	0.3330	-0.297	-0.522	1		
X_9	0.619	0.0948	0.053	0.0971	0.450	0.5933	-0.215	-0.530	0.396	1	
X_{10}	0.088	0.1153	0.002	-0.045	-0.141	-0.119	0.028	0.0494	-0.008	-0.169	1

4.2. Modeling and Testing

4.2.1. Modeling

A multiple regression model was established for the crude oil price series and other variables after HP filtering. The equation to be estimated is:

$$WTI_i=\alpha(1)+\alpha(2)*X_1+\alpha(3)*X_2+\cdots+\alpha(11)*X_{10}+\alpha(12)*WTI_i(-1) \tag{2}$$

Through OLS estimation, the parameter estimation results of this equation are shown in Table 3. The R^2 statistic measuring goodness of fit is 0.9360, and the adjusted sample determination coefficient is 0.9250, indicating that the regression equation fits well with the sample value. The F statistic for testing the significance of variance is 971.226, and its P value is close to zero. The null hypothesis is rejected at the significance level of 1% and the equation is significantly valid.

In addition, the t-statistic for testing the significance of variables also meets the requirements overall, so it can be concluded that the parameter estimation results of the regression equation are basically reasonable.

Table 3. Regression equation parameter estimation results

Variable	Coefficient	Standard error	t-statistic	P-value
C	58.2787	14.2860	4.0794	0.0001
X_1	-0.0007	0.0003	-2.6813	0.0078
X_2	-0.000385	0.000258	-1.4954	0.1360
X_3	-0.003942	0.003675	-1.0725	0.2845
X_4	-0.0928	0.0849	-1.0923	0.2757
X_5	0.2465	0.0532	4.6294	0.0000
X_6	-0.0000165	0.0000982	-1.6769	0.0947
X_7	-0.2486	0.0816	-3.0477	0.0025
X_8	0.000509	0.000381	1.3351	0.1830
X_9	2.2778	0.8410	2.7085	0.0072
X_{10}	0.2353	0.0713	3.3016	0.0011
WTI(-1)	0.5778	0.0565	10.2281	0.0000
R-squared	0.9360	Mean dependent var		56.1769
Adjusted R-squared	0.9250	S.D. dependent var		28.5022
S.E. of regression	4.5094	Akaike info criterion		5.8928
Sum squared resid	5347.95	Schwarz criterion		6.0507
Log likelihood	-798.2664	F-statistic 971.2260		971.2260
Durbin-Watson stat	1.0151	Prob(F-statistic)		0.0000

The estimated result of the equation is:

$$WTI=58.2787-0.0007X_1-0.000385X_2-0.003942X_3-0.0928X_4+0.2465X_5-0.0000165X_6 -0.2486X_7+0.000509X_8+2.2778X_9+0.2353X_{10}+0.5778WTI \quad (3)$$

4.2.2. Testing

ADF unit root test was performed on the residual sequence of the equation, and the results were shown in Table 4. The T statistic was -9.7127 and its P value was 0.0000, so the residual of the equation rejected the null hypothesis at the significance level of 1% and accepted the conclusion that there was no unit root. The residual sequence was relatively stable, and the setting of the regression equation was reasonable.

Table 4. Stationarity test of residual sequence of equations

Null Hypothesis:Resid has a unit root			
Exogenous:Constant			
Lag Length:0(Automatic-based on AIC,maxlag=8)			
		t-Statistic	Prob.*
ADF test statistic		-9.712724	0.0000
Test critical values	1%level	-3.454085	
	5%level	-2.871883	
	10%level	-2.572354	
*MacKinnon(1996)one-sided p-values.			

4.3. Model Prediction and Result Analysis

Fitting prediction was made according to the regression equation obtained above, and the results are shown in Figure 1. The mean absolute error (MAE), square absolute percentage error (MAPE) and root mean square error (RMSE) were used to determine the prediction error.

$$MAE=\frac{1}{m}\sum_{i=1}^{m}\left|\left(y_i-\hat{y}_i\right)\right| \tag{4}$$

$$MAPE=\sum_{i=1}^{m}\left|\frac{y_i-\hat{y}_i}{y_i}\right|\times\frac{100}{m} \tag{5}$$

$$RMSE=\sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(y_i-\hat{y}_i\right)^2} \tag{6}$$

According to the calculation, the regression equation predicts that THE WTI crude oil price MAE, MAPE and RMSE from January 2017 to December 2019 are 3.23, 5.64% and 3.82, respectively. The prediction accuracy is high, indicating that the multiple regression model combining the fluctuation term extracted by HP filter and the influencing factors can objectively and clearly reflect the change rule of oil price.

5. Conclusion

In this paper, based on HP filter and multiple regression prediction model of crude oil price, the fluctuation components of crude oil time series are extracted, and the influencing factors are introduced into the model, which can reflect the change trend of crude oil price more objectively. At the same time, spot price of WTI crude oil is predicted. The results show that the model has high prediction accuracy, and the predicted value can approach the actual value better, thus demonstrating the feasibility and effectiveness of this method. The trend of international oil price is full of various uncertainties. The prediction model can obtain the general fluctuation of oil price through quantitative analysis, which can provide reference for decision-making to a certain extent but cannot achieve accurate prediction of oil price. Decision-makers should adjust the prediction results of the oil price model according to the actual situation of oil market changes, to overcome the shortcomings of the model's insensitive response to emergencies and further improve the prediction accuracy. There are many and complex factors affecting the price of crude oil, including wars, political conflicts, natural disasters and other unexpected times, as well as psychological expectations of oil market participants. Further analysis and research are needed to analyze the quantitative impact of crude oil price.

References

[1] Wei W. X., Lin B. Q. International and domestic oil price volatilities and their interrelationship [J]. Economic Research Journal, 2007, 042(012):130-141.

[2] Zhao L. T., Guo S. Q., Wang B., et al. International crude oil price analysis and prediction in 2019 [J]. Journal of Beijing Institute of Technology (Social Sciences Edition), 2018, 21(02):26-30.

[3] Xu P., Liu Q. Drivers of international crude oil price: demand, supply, or finance-An analysis based on historical decomposition[J]. Macroeconomics, 2019, (07):84-97.

[4] An H. Z., Gao X. Y., Huang S. P., et al. Mutiscale impacts of oil price fluctuations driven by the demand and supply on the stock market [J]. Chinese Journal of Management Science, 2018,26(11):62-73.

[5] Shen D. P. The research of international crude oil price forecasting model[D].University of International Business and Economics, 2015.

[6] Lin L., Jiang Y., Xiao H. L., Zhou Z. B. Crude oil price forecasting based on a novel hybrid long memory GARCH-M and wavelet analysis model[J]. Physica A: Statistical Mechanics and its Applications, 2020, 543: 123532.

[7] Lin S., Ye X., Liu J. P. Analysis and prediction of international oil price based on multi-factor wavelet and regression method[J]. Journal of Xidian University (Social Science Edition), 2011, (06):55-62.

[8] Cui J. X., Zou H. W. Oil futures price forecasting model named CEEMDAN-PSO-ELM [J]. Computer Systems & Applications,2020,29(02):28-39.

[9] Zhang J. L., Li D. Z., Tan Z. F. International crude oil price forecasting based on a hybrid model [J]. Journal of Beijing Institute of Technology (Social Sciences Edition),2019,21(01):59-64.

[10] Liu Y. Q. Research in world crude oil price and its factors [D]. Tianjin University, 2010.

[11] He X. Q. Modern statistical analysis methods and applications[M]. Renmin University of China Press, 1998.

[12] Lu Q., Li Y., Chai J., et al. Crude oil price analysis and forecasting: A perspective of “new triangle” [J]. Energy Economics, 2020,87:104721.