Lauren Ferguson

Introduction

Oil is the most important commodity for the global economy. Crude oil prices are determined by changes in the economy, which dictate supply and demand. These price changes can be an indicator of what will happen with the rest of the global economy. If oil prices can be predicted by other variables then based off the changes predicted in oil prices, we can forecast what is to be expected with the global economy.

A dataset was compiled including a large number of financial variables by week for a ten-year time span. Exchange rates against the United States dollar for the British pound, Swedish krona, Japanese yen, Australian dollar, Canadian dollar, and Swiss franc are all included. A number of bonds are also included such as high yield United States corporate bonds and 10 to 20-year U.S. Treasury bonds as well as stock market groups like the S&P 500 index and the Technology sector index.

These variables were chosen as potential predictors of oil prices for a variety of reasons. First, exchange rates based off the United States dollar are included because oil benchmarks are usually priced in U.S. dollars. So, if the exchange rate for the U.S. dollar decreases then that represents a devaluation, which decreases the value of oil in other countries. This would then lead to increased demand due to lower pricing, which in turn drives costs higher.

Second, bonds are included as potential predictor variables because people tend to invest their money in bonds when the economy is unstable or on the decline as bonds are safer than the stock market. Therefore, if bonds are increasing then the economy is suffering and when the economy declines so do the profits of businesses who require oil. With less demand on oil, the price of oil goes down. For that reason, bond prices could be an indicator for oil prices.

Third, stock prices are included because they also are an indicator of how the economy is performing. If the economy is doing well, then stocks are also doing well. When the economy improves, company profits improve and due to this, there is an increased need for oil. With the higher demand on oil, the price of oil increases.

Along with the returns for all these variables, or percentage change of week-to-week prices, a lagged return for each one is also included. A lagged return is what happened with that particular variable the previous week. If any lagged variable is able to predict oil prices in the future, then that means what happened last week can predict what will happen next week.

To determine if any of these variables can predict oil prices, I will run a statistical modeling analysis. The models I will use are Ordinary Least Squares (OLS), which is a standard linear regression tool, as well as forward and backward stepwise.

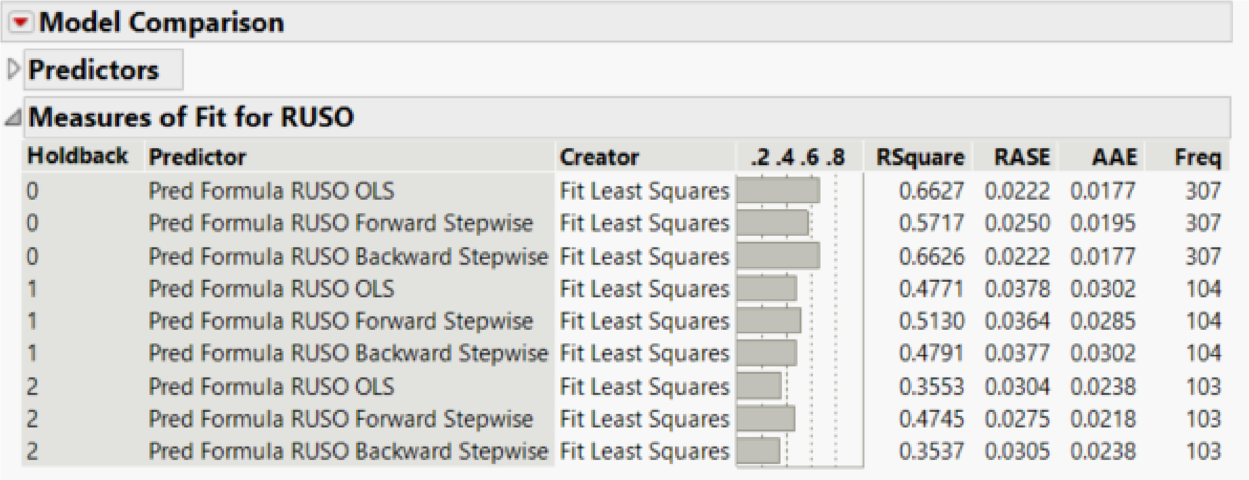
Analysis and Model Comparison

To begin, I will split the dataset into three parts. Sixty percent will be the training data that is used to estimate the model. Twenty percent will be the validation data that is held out from the estimation so the model can be continuously tested. The third portion will be the final 20 percent of data, the testing data, which will be hidden away in order to provide new, never before seen data to test the model. This will ensure an unbiased analysis of the predictive ability of the model created. The model that performs the best on this section of the data will be the one chosen.

First, I will run the OLS model. With big data, OLS can produce estimates with a great deal of variance and has the potential to model random noise and generate poor forecasts. While I know this to be the case, I will still employ this model first as a benchmark to compare other models to help determine which one is best.

Forward stepwise will be the second model examined. This model will add one variable from the dataset at a time while continuously testing it on the validation portion of the data. Once the r squared value stops improving, the model will stop building. Therefore, the model can end up with many of the variables from the dataset or just one. I will then run OLS using the handful of variables chosen through forward stepwise modeling.

Lastly, I will complete a third model called backward stepwise. Backward stepwise is the opposite of the previous model, forward stepwise. This model will add in all of the variables from the dataset and remove them one-by-one until it reaches the highest r squared value. The potential disadvantage to these two stepwise models is if there are too many highly correlated variables in the dataset, then multicollinearity could be an issue with the model. The model could kick weaker variables out and we would not end up with a clear picture of all the predictors. This is definitely possible in this situation because exchange rates, stock prices, bond prices, and inflation all work hand-in-hand based off economic situations.



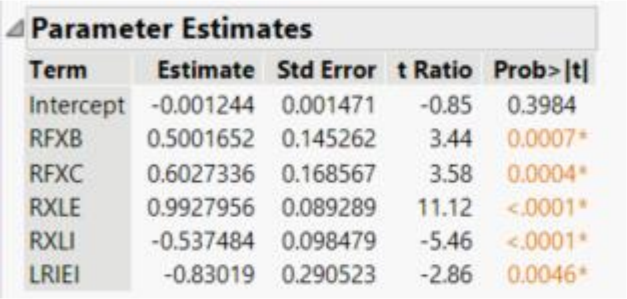
Holdback	Predictor	Creator	.2 .4 .6 .8	RSquare	RASE	AAE	Freq
0	Pred Formula RUSO OLS	Fit Least Squares		0.6627	0.0222	0.0177	307
0	Pred Formula RUSO Forward Stepwise	Fit Least Squares		0.5717	0.0250	0.0195	307
0	Pred Formula RUSO Backward Stepwise	Fit Least Squares		0.6626	0.0222	0.0177	307
1	Pred Formula RUSO OLS	Fit Least Squares		0.4771	0.0378	0.0302	104
1	Pred Formula RUSO Forward Stepwise	Fit Least Squares		0.5130	0.0364	0.0285	104
1	Pred Formula RUSO Backward Stepwise	Fit Least Squares		0.4791	0.0377	0.0302	104
2	Pred Formula RUSO OLS	Fit Least Squares		0.3553	0.0304	0.0238	103
2	Pred Formula RUSO Forward Stepwise	Fit Least Squares		0.4745	0.0275	0.0218	103
2	Pred Formula RUSO Backward Stepwise	Fit Least Squares		0.3537	0.0305	0.0238	103

Once all three of these models were run using all the variables on the validation, training, and test data, the chosen model is forward stepwise for a number of reasons. One, forward stepwise performed the best out of all three on the brand new, never before seen test data. Its 0.47 r squared value on the test data was much higher than the 0.35 r squared value for the OLS and backward stepwise models. While OLS and backward stepwise had a high 0.66 r squared value on the training data, the value dramatically decreased with the validation data and again with the test data. These two models had their r squared value almost cut in half from the training data to the test data.

This is not the same case for the forward stepwise model and another reason why it is the chosen model. The r squared values stay relatively the same with each set of data in the forward stepwise model, which leads me to believe there is no overfitting happening with this model. The third and final reason forward stepwise is the selected model is due to the root average squared error (RASE) and the average absolute error (AAE). The lower these numbers are for the model, the less error there is in the model. Both of these are lower for the forward stepwise model with the validation data, and more importantly, the testing data.

Interpretation

The forward stepwise model uses five variables from the dataset to predict oil prices. RFXB and RFXC are the exchange rates for British pounds and Canadian dollars, respectively. RXLE and RXLI are, in order, the energy and industrials sector index. Lastly, LRIEI is last week's returns on three to seven-year U.S. Treasury bonds.



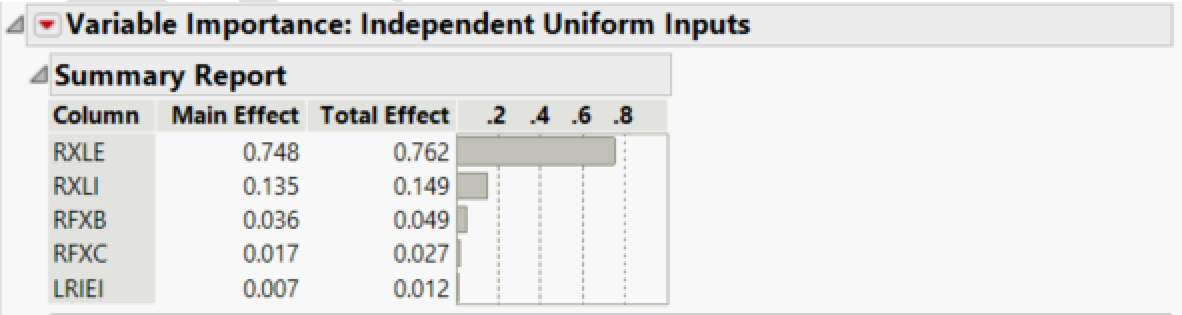
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.001244	0.001471	-0.85	0.3984
RFXB	0.5001652	0.145262	3.44	0.0007*
RFXC	0.6027336	0.168567	3.58	0.0004*
RXLE	0.9927956	0.089289	11.12	<.0001*
RXLI	-0.537484	0.098479	-5.46	<.0001*
LRIEI	-0.83019	0.290523	-2.86	0.0046*

As the exchange rate based in U.S. dollars go up for British pounds by a half percent and Canadian dollars by over a half percent, oil prices also increase the same amount. This increase in the exchange rate makes the U.S. dollar more valuable and because oil prices are usually priced in U.S. dollars then when the dollar appreciates, oil prices outside of the United States increase.

A one percent growth in the energy sector index leads to a one percent rise in oil prices. As the energy economy does better, energy sources have higher demand and as a result, have an increase in prices. A half percent decrease in the industrials sector index leads to a half percent increase in oil prices. If the industrial sector is doing poorly then that means there is less mining occurring, which leads to less existing oil. With less oil availability, the price for oil is driven up.

An almost one percent decrease in the lagged returns for three to seven-year U.S Treasury bonds causes an almost one percent increase in oil prices. When people run to bonds, it is an indicator of a bad economy. They want to invest their money in a safer place than the stock market. When the economy is good, profits are high for businesses, which means a greater need for raw materials like oil. This increased demand generates higher oil prices. Due to this reasoning, these three to seven-year U.S. Treasury bonds and oil prices move in opposite directions.



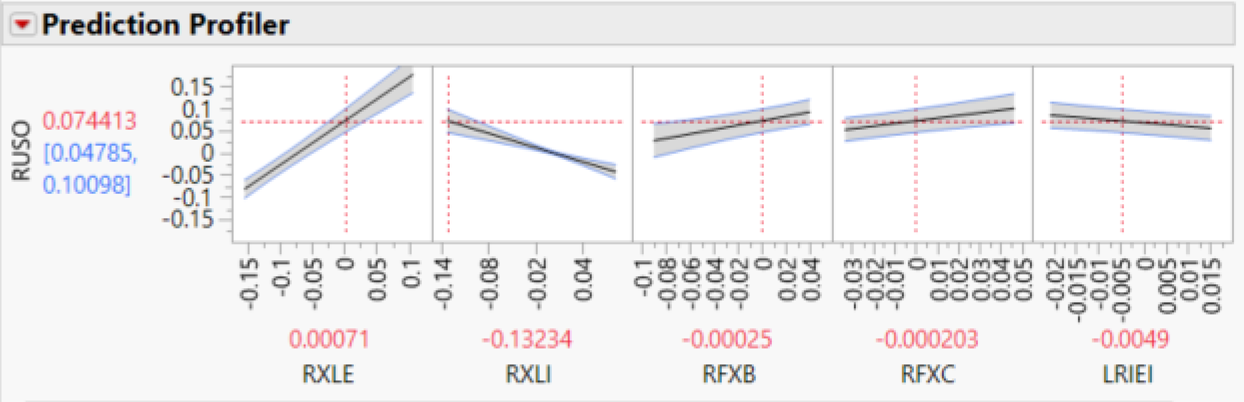
Variable Importance: Independent Uniform Inputs

Summary Report

Column	Main Effect	Total Effect
RXLE	0.748	0.762
RXLI	0.135	0.149
RFXB	0.036	0.049
RFXC	0.017	0.027
LRIEI	0.007	0.012

Which of these five variables in the forward stepwise model impact oil price predictions the most? The energy sector index explains about 76 percent of the variations in the oil market. It makes sense that the energy economy would be the most important variable to predict oil prices. As the energy sector does better, there is a higher demand on oil and prices will increase.

The second most important variable is a distant second place. Fluctuations in the industrials sector index explain about 15 percent of changes to oil prices. British pound changes explain five percent, Canadian dollars two percent, and the lagged return on three to seven-year U.S. Treasury bonds has the lowest with one percent.



Oil prices have a positive correlation with three out of the five predictor variables and a negative correlation with two. The strongest positive correlation with oil prices is RXLE, returns on the energy sector index. These two variables have an almost perfect correlation, as one goes up, the other also goes up. The second strongest correlation with oil prices is a negative correlation. As the lagged returns on three to seven-year U.S. Treasury bonds go down, oil prices go up.

Next, fluctuation in the Canadian dollar, RFXC, has a positive correlation with oil prices. Canada has one of the largest oil reserves in the world. The fourth strongest correlation with oil prices is a negative correlation with RXLI, returns on the industrial sector index. As the industrial sector index goes down, oil prices go up. The British pound has almost the exact same correlation with oil prices as RXLI, but in the opposite direction. As returns on the British pound goes up, so do oil prices. The United States became the top supplier of oil to Britain in early 2019.

Using a forward stepwise model with a 0.47 r squared value on new, test data, 0.51 on the validation data, and 0.57 on the training data, oil prices can be predicted by five variables. These five variables are returns on the British pound and Canadian dollar exchange rate, returns on the energy and industrials sector index, and lagged returns on three to seven-year U.S. Treasury bonds. When either of the two exchange rates go up or the energy sector index goes up, so do oil prices. When the industrial sector index or the three to seven-year U.S. Treasury bonds go down, oil prices go up.