**Case Prediction of Oil Prices with a Large Number of Financial Market Variables**

Susan Wiggins

Cases in Business Analytics BAN 525

Professor: Dr. Cetin Ciner



**Table of Content:**

Introduction…………………………………………………………………………………………………………………………………………3-4

Analysis and Model Comparison………………………………………………………………………………………………………….4-7

Interpretation……………………………………………………………………………………………………………………………………..7-8

Appendix……………………………………………………………………………………………………………………………………………9-16

References……………………………………………………………………………………………………………………………………………17

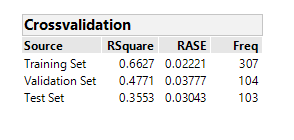
**Introduction**

Oil is considered one of the most important global commodities. This case study of prediction of oil prices will cover a large number of financial market variables based on a pre-generated dataset and address the explanation of oil prices in a financial market as well as any lagged effects. The dataset consists of variables from currency, bond, and stock markets. The data is weekly and consisted in the time series of March 18, 2009 through March 6, 2019. The dataset was prepared based on the following criterion: raw prices were converted into percentages, a difference of a natural log was established, continuous compound returns, returns are noted with R labeling, and L denotes lagged instances. The dependent variable being measured in this study is oil prices (RUSO) and the predicator candidates are exchange rates, interest rates, stock market, market volatility, and inflation. Exchange rate is generally priced based on United States dollars for oil price, if the dollar decreases then a decrease will occur in demand for oil outside of the United States. In addition, stocks and oil prices are correlated together based on the economic conditions of the economy. Also, when economic conditions change bond prices will move in opposite direction of the price of crude oil. The building block for this is due to bond prices and interest rate enforcing the same characteristics of moving in opposite directions. Three methods will be addressed in this case ordinary linear regression, stepwise forward, and stepwise backward. Peter Christie, Jim Georges, Jeff Thompson, Chip Wells and colleagues point out the specifics on these three modeling methods, “In standard linear regression, a prediction estimate for the target variable is formed from a simple linear combination of the inputs. The intercept centers the range of predictions, and the remaining parameter estimates determine the trend strength (or slope) between each input and the target. The simple structure of the model forces changes in predicted values to occur in only a single direction (a vector in the space of inputs with elements equal to the parameter estimates). Intercept and parameter estimates are chosen to minimize the squared error between the predicted and observed target values (least squares estimation). The prediction estimates can be viewed as a linear approximation to the expected (average) value of a target conditioned on observed input values. Forward selection creates a sequence of models of increasing complexity. The sequence starts with the baseline model, a model predicting the overall average target value for all cases. The algorithm searches the set of one-input models and selects the model that most improves on the baseline model. It then searches the set of two-input models that contain the input selected in the previous step and selects the model showing the most significant improvement. By adding a new input to those selected in the previous step, a nested sequence of increasingly complex models is generated. The sequence terminates when no significant improvement can be made. In contrast to forward selection, backward selection creates a sequence of models of decreasing complexity. The sequence starts with a saturated model, which is a model that contains all available inputs, and therefore, has the highest possible fit statistic. Inputs are sequentially removed from the model. At each step, the input chosen for removal least reduces the overall model fit statistic. This is equivalent to removing the input with the highest p-value. The sequence terminates when all remaining inputs have a p-value that is less than the predetermined stay cutoff” (4.4-4-26).

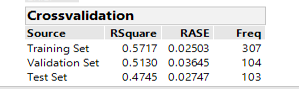
**Analysis and Model Comparison**

The dataset employed in this case study is large and sometimes modeling can be affected when Big data is taken into account. Ordinary Least Squares (OLS) a modeling technique most highly used in prediction methods can cause problems when Big data is being used. For instance, OLS modeling can cause large variances in the data, models random noise, and poor forecasting can exists. Therefore, other machine learning techniques need to be implemented such as forward stepwise and backward stepwise. In order to reach the best results in the analysis one must depend on cross validation which will decrease the ability to interpret random noise. In cross validation, data is held out and the predication criterion is based on this interpretation. Estimates are then built on the data after the completion of each modeling affect occurs. In the case predication of oil prices, the cross validation is established on 60/20/20 split of the data. Therefore, sixty percent of training data is used in estimation to row 309, twenty percent used for validation in order to stop the modeling process between row 310 and 413, and twenty percent used for testing for unbiased analysis in predication ability of the model from row 414 to end. The final interpretation of the model will be based on the results of the test data due to the basis of new observations.

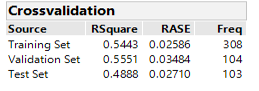
Since the oil case dealt with time series data, the predictions are included at the end of the analysis making validation and test dead last in the column. JMP allows the user to set up the training, validation, and testing based on the labeling technique of 0, 1, 2. Zero is associated with training, ones are associated with validation, and twos are indicated by testing data. Next, modeling techniques of OLS, forward stepwise, and backward stepwise are initiated in the process to discover oil prices. First, ordinary linear regression was performed on variables listed from RFXB to LRIYR while removing the RUSO dependent variable from the list. The results for the OLS are represented in the following chart.



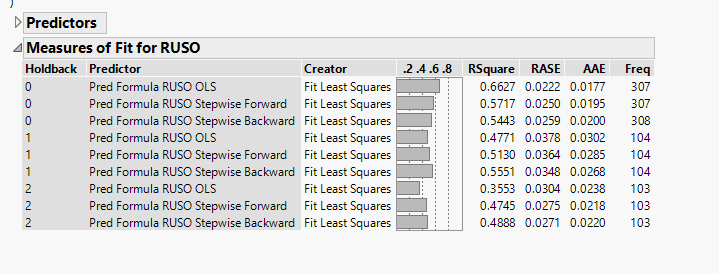
The testing data reveals that 35.53% of the data can be interpreted from the model. The root average square error (RASE) score is also low at .03043. Second, forward stepwise indicated better results based on test set of 47.45% and an even lower RASE score at .02747 shown in the following table.



The final model backward stepwise resulted in an even higher yield of Rsquare found to be 48.88% and a low RASE score at .02710. The following chart list the cross validation for the backward stepwise modeling technique.



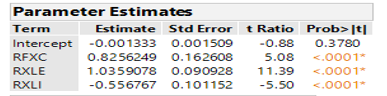
Overall, when studying each of the cross validation the backward stepwise shows the greatest value for Rsquare and the lowest value for root average square error (RASE). A comparison of the modeling techniques was performed, and the results discovered are in the table below.



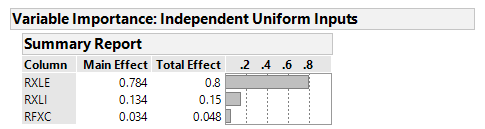
Furthermore, Stepwise Backward is the modeling technique that reveals the best unbiased results for testing. RSquare was the highest yield at 48.88% and the root average square error was the lowest value also. In addition, the average absolute error (AAE) was also second lowest in comparison to the models indicating a good fit for the model. Therefore, Stepwise Backward explains 48.88% of variations in oil prices and one can see RASE scores are much less than other models as well as low scores for AAE.

**Interpretation**

The parameter estimates for the backward stepwise model are listed as RFXC, RXLE, and RXLI. RFXC is the Canadian dollar exchange rate against the United States dollar. RXLE is Energy sector in the stock market group and RXLI is the Industrial sector in the stock market group. The exact numerical interpretations are listed in the following table.



The Canadian dollar is a general importance towards commodities. Canada is a big commodity exporter so then it would show developments in commodities. The stock market groups of industrial sector and energy sector are of definite interest due to these two categories having possible increases or decreases in the amount of oil demand. All value parameter estimates are statistically significant because they are less than 5 percent. Next, the variable importance is included in the predication of oil prices and the following table gives us the results below.

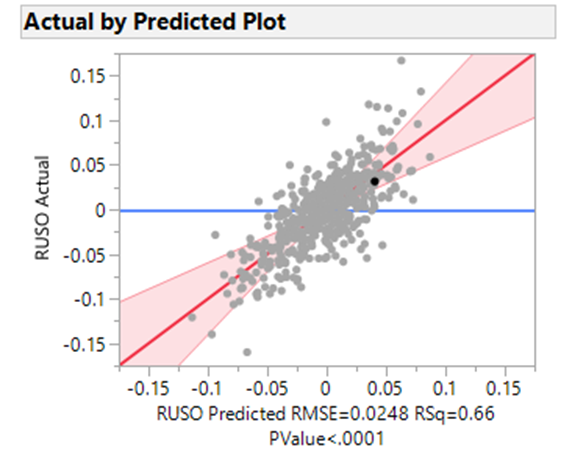


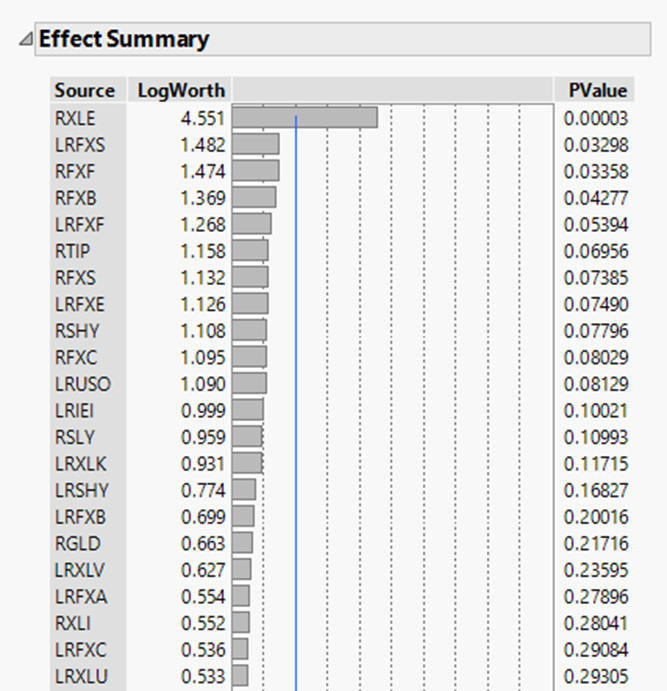
The total effect takes into account the variations in the data and the RXLE energy sector explains 80% of the variations in the oil price. The most important variable to look at is the energy sector. The second variable to look at pertaining to oil prices is the industrial sector at 15% and the final variable of interest is the Canadian dollar with total effect at 4.8%. The model prediction profiler reveals that the energy sector and the Canadian dollar have a positive relationship whereas the industrial sector has a negative relationship. As you increase the average of both the energy sector and the Canadian dollar the price of oil increases whereas when you increase the industrial sector average a decrease in oil price exists within the prediction profiler.

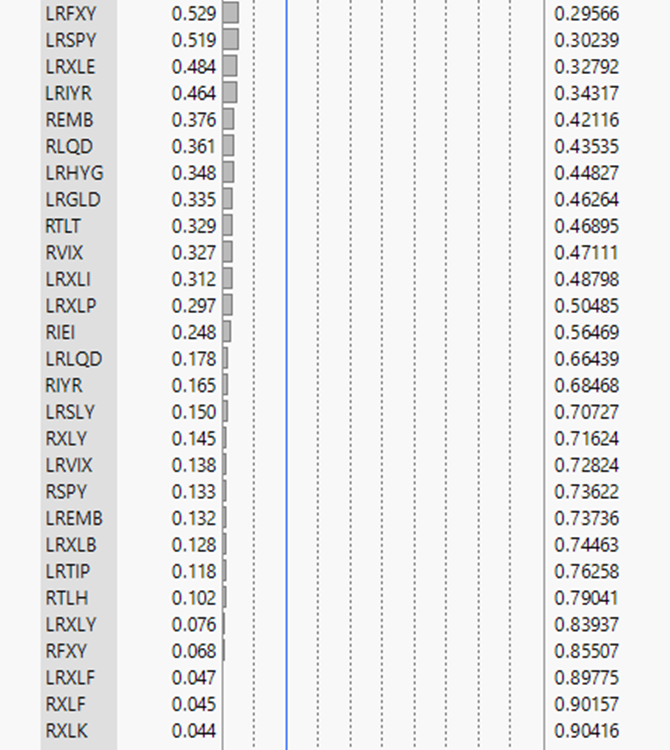
In conclusion, the selected model to predict oil prices was Stepwise Backward. The highest variable for fluctuations in oil price is related to the energy sector of stock market group at 80 percent and has a positive relationship. The second most important variable associated with oil price is the industrial sector and has a negative relationship on oil prices. The final variable studied in the Stepwise Backward is the Canadian dollar which only takes into account a small percentage at 4.8% and has a positive relationship in account that when the Canadian dollar rises so does the oil price. Supply and demand in oil at an energy standpoint as well as industrial can significantly change the price of oil for consumers. Canadian dollar plays a role in fluctuation of oil price due to significantly large commodity exporter. In addition, if one was to export oil from Canada the oil would cost more than oil that is from the Middle East.

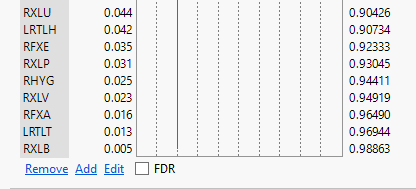
**Appendix:**

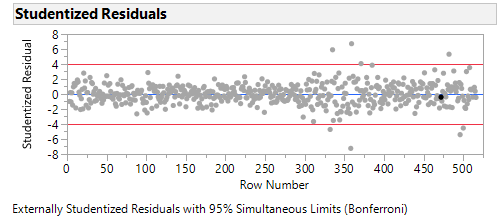
**Ordinary Linear Regression**

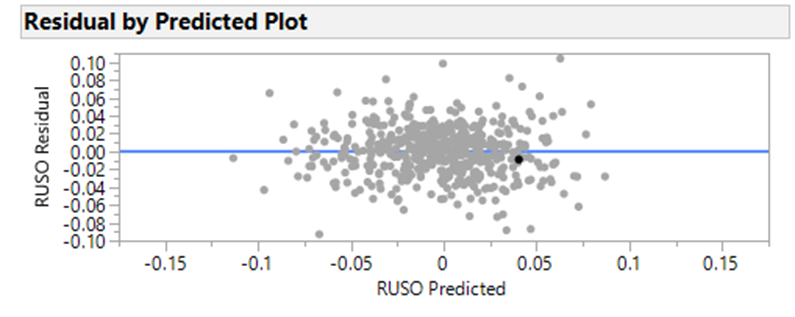
****

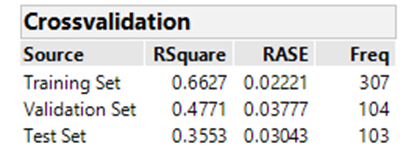
****

****

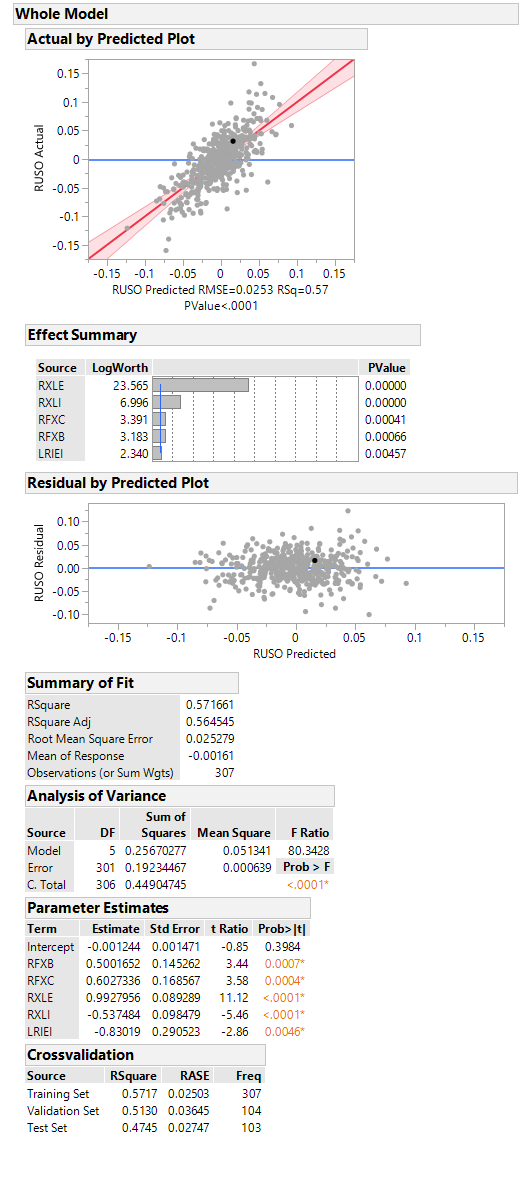




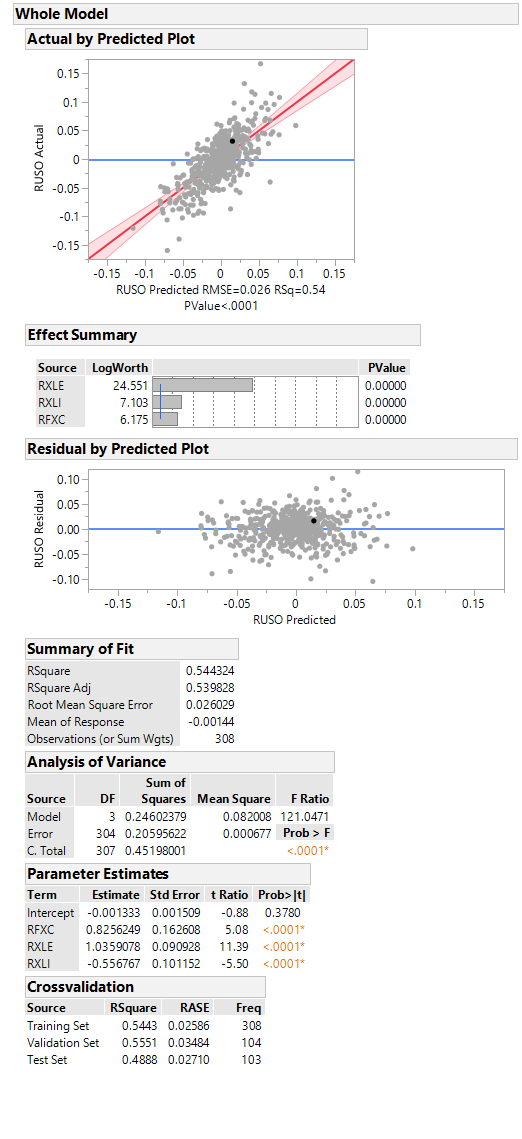
****

****

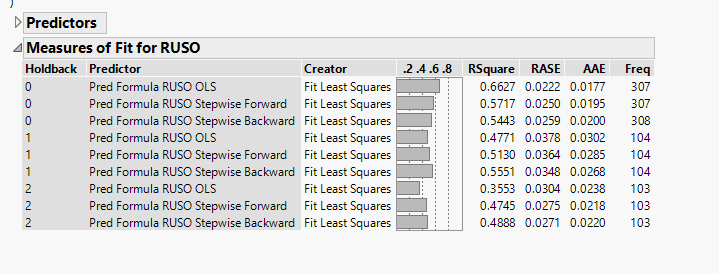
**Forward Stepwise**



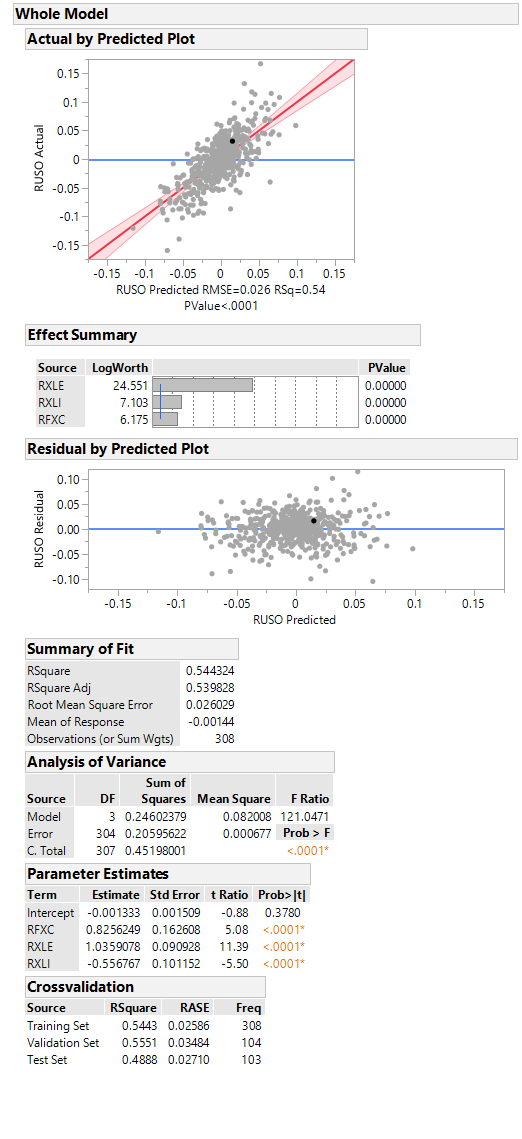
**Backward Stepwise**

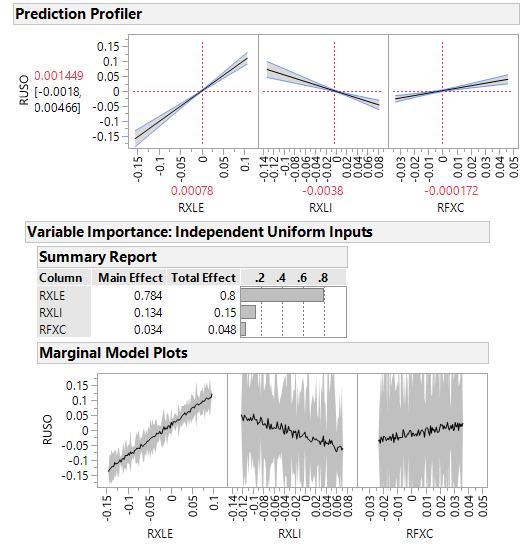


**Model Comparison:**



**Final Analysis of Project**





**References :**

Christie, Peter. Georges, Jim. Thompson, Jeff. Wells, Chip. Blanchard, Robert. Bohannon, Tom. Hardin, Mike. Kelly, Dan. Laramore, Jay. Lucas, Bob. De Waal, Andre. Walsh, Sue. “Applied Analytics Using SAS Enterprise Miner.” SAS Institute Inc. 14 Jan. 2016. 1-639. Print.

Ciner, Cetin. August 2019. Module 1. “Cross Validation and Predicting Gold Prices.” Slides 1-53. Retrieved from: <https://uncw.instructure.com/courses/24252/pages/module-1-lecture-materials-cross-validation-stepwise-regression-and-risk-factors-for-gold-prices?module_item_id=524495>

Lioudis, Nick. “What Causes Oil Prices to Fluctuate.” Investopedia. 20 April. 2018. Web. <https://www.investopedia.com/ask/answers/012715/what-causes-oil-prices-fluctuate.asp>

Kosakowski, Paul. “What Determines Oil Prices?” Investopedia. 25 June. 2019. Web. <https://www.investopedia.com/articles/economics/08/determining-oil-prices.asp>

Web. Picture. <https://www.bizjournals.com/houston/morning_call/2014/08/oxy-building-600-employee-office-to-serve-permian.html>