Gil Graybill – BAN 525
Assignment 1 – Risk Factors for Oil Prices
Professor Cetin Ciner
May 16, 2020

Introduction

Oil is everywhere in the world economy, and the thirst for this commodity is seemingly endless. It is used extensively in transportation for automobiles, jet fuel, and diesel trucks. Kerosene and other refined products are used for heating homes and cooking. It is also used to make plastics, which are everywhere these days.

Would it be possible to create a statistical model that could predict or describe changes in the price of oil? Are there variables from currency, bond, and stock markets that can help determine the price of oil? Could the overall economy of a given country affect the price of oil? A model may not be able to directly determine the results of a unique event like a pandemic, but can the changes in price of other variables as a result of that event indicate what might happen to the price of oil? Can any of these variables help predict what the future price of oil will be? Knowing what affects the price of oil, either in a positive or negative way, would be useful for businesses and industries who depend on it and would like stability in budgeting. This is what we'll be investigating.

The dataset we are working with consists of 10 years of weekly prices for 6 different currency exchange rates, 7 different interest rates for bonds of various terms, 12 stock market indices, and 3 miscellaneous measures (See Appendix A). Each of these variables has been normalized with a percentage change from week to week. In addition, we have created a "lag" variable so we can compare last week's changes for a given variable against this week's change in oil prices. In all, there are 56 different variables we will be testing to see if they have an influence on the price of oil.

We will be using 3 types of analysis on this dataset:

(1) OLS (Ordinary Least Squares): This type of linear regression model calculates the best fit through the data points that minimizes the squared differences between the observed values and the fitted values.
(2) Forward Stepwise Regression: Starting with no variables and only a constant, a different variable is considered for addition to a linear formula. Variables are added until there are no more significant contributions made by adding them.
(3) Backward Stepwise Regression: Starting with all variables, variables that do not contribute to the effectiveness of the model are removed one-by-one based on a set criteria

Analysis and Model Comparison

If the relationship between the response variable and the potential dependent variables is linear, OLS will help find those relationships. While the resulting formula can involve many dependent variables, it's easy to understand how these variables affect the response variable. It does assume that the data is normally distributed. One disadvantage is that since OLS tries to fit the best model using all values of a given variable, outliers can bias the results. Another disadvantage is that a novice user may find OLS easy to use and consequently will use it when a different model might be more appropriate.

Stepwise Regression, both forward and backward, are intuitive approaches to finding dependent variables. With forward regression, the most statistically significant variable is added each step, and then the model recalculated. When there are no more statistically significant variables to add the process stops. With backward regression, a model is created with all dependent variables, and one-by-one they are removed based on statistical insignificance. The process stops when removing a variable hurts the accuracy of the model. One disadvantage of this modelling is that sometimes a variable that in reality has a casual effect on the model will be included merely by coincidence. Also, when a variable is included (or removed), that action is permanent for the rest of the processing of the model. There may be two variables that work well together to determine the response variable, but they may not end up together in the final model.

For cross-validation we employed a 60-20-20 split for training-validation-test split. This gave us 309 rows of training data, 103 rows of validation data, and 103 rows of test data. Because this is time series data, we did not split up the data randomly, but instead used the first 60% of the data (time-wise) for training, the next 20% for validation, and the final 20% for testing. This makes sense since we are trying to build a model that could potentially be used to predict future oil prices. With time series data there is always the risk that an event or phenomena will alter the data from a certain date forward. For instance, if demand dropped precipitously because a technological breakthrough could provide inexpensive tabletop fusion, past performance may not be able to predict future results.

The price of oil was chosen as the response variable, and all of the normalized variables and their lags were tested as dependent variables using JMP Pro 15. The calculated value for the response variable, the change in the price of oil, was determined and recorded and the resulting formulas determined using all three methods. It is noteworthy that both Forward Stepwise Regression and Backward Stepwise Regression found the same significant dependent variables.

To compare results between the models, we calculate the difference between each observed value for the change in the price of oil and the calculated value for the change in the price of oil. These differences can be compared in different ways. "RSquare" is a percentage that tells how much of the variability is explained by the model. The higher RSquare is the better. "RASE" is "Root Average Standard Error", and because this describes values of error, smaller is better. "AAE" is "Absolute Average Error" is an error term that uses the absolute value of the error to compare the size of the error without regard for direction (positive or negative). While all are fine measure of model effectiveness, we'll be using determining the best model using RSquare. It is worth noting that RSquare was used as the stopping criteria for Stepwise Regression. The resulting equation is the maximum value for RSquare that could be found.

When comparing the models we will be using our cross-validation split. Because we are looking for a predictor of future oil prices we will put more emphasis on the RSquare values for the test data (later chronologically) than the results for the training and validation. The results are as follows:

Cross-validation step	Model	RSquare
Training	OLS	0.6536
Training	Stepwise	0.5403
Validation	OLS	0.5243
Validation	Stepwise	0.5619
Test	OLS	0.3606
Test	Stepwise	0.4895

While the training data yielded an amazing 65% RSquare value for the OLS model, the prediction power did not hold up for the validation and testing data. The RSquare values for Stepwise Regression (both Forward and Backward) were 54%, 56%, and 49%, which is a pretty close cluster. These consistent values give us confidence that using Stepwise Regression against our dataset yields a better predictive model for oil prices than OLS, which got worse as times went on.

Interpretation

The formula that was determined through Stepwise Regression is

Predicted oil price = (-0.00115157029784242) +
0.835955791205807 * RFXC +
1.04788643316131 *RXLE +
-0.583251491137531 * RXLI

RFXC represents changes in the Canadian dollar vs. American dollars. Canada is the world's 7th largest oil producer and is a big employer when compared per capita with other countries (https://en.wikipedia.org/wiki/List_of_countries_by_oil_production). A rise in oil prices is good news for the Canadian economy and the Canadian dollar. No other countries whose exchange rates were included in this dataset are in the top 20 producers.

RXLE is a stock market group that represents the energy sector. Since oil is part of the energy industry, it's reasonable to assume that as the fortunes of the oil market change, the stocks of the energy sector will do likewise. Looking at the makeup of the sector shows that the companies related to oil refining, exploration, and transportation are a major part of the index (https://www.etf.com/XLE#overview)

RXLI is a stock market group that represents the Industrials sector. Note that the coefficient is negative, meaning that as Industrial stocks rise or fall, the price of oil does the opposite. This makes sense since oil is an expense that bears heavily on industry. Looking at the makeup of the companies in this index (https://www.etf.com/XLI#overview), it represents a wide cross-section of the worldwide industry.

When running variable importance with these three variables against this model, we see the following:

Variable	Main Effect	Total Effect
RXLE	0.765	0.78
RXLI	0.164	0.18
RFXC	0.029	0.041

RXLE is responsible for 78% of the total effect of this model, with RXLI coming in at 18% total effect. It makes sense that oil prices have a large effect on the Energy stock market index because oil-related stocks make up so much of index (https://www.etf.com/XLE#overview).

Appendix A – Normalized Variables used in analysis

FXB: British pound
FXS: Swedish krona
FXY: Japanese yen
FXA: Australian dollar
FXC: Canadian dollar
FXF: Swiss franc
LQD- Investment grade corporate bonds
SHY- 1 to 3 year US treasury bonds
EMB- International (Emerging market) bonds
TLT- Long term (20+) US treasury bonds
IEI- 3 to 7 year US treasury bonds
TLH- 10 to 20 year US treasury bonds
HYG- High yield (risky) US corporate bonds
SPY: S&P 500 (large company) index
SLY: S&P 600 (small company) index
XLB: Material sector index
XLE: Energy sector index
XLF: Financial services sector index
XLI: Industrials sector index
XLK: Technology sector index
XLP: Consumer staples sector index
XLU: Utilities sector index
XLV: Healthcare sector index
XLY: Consumer discretionary sector index
IYR: Real estate sector index
USO: oil prices
VIX: stock market volatility
TIP: inflation measure
GLD: Gold prices