Daniel Sanchez
BAN 525
Dr. Ciner
Module 1: Assignment 1 -- Risk Factors for Oil Prices: Before and After Covid-19
May 16th, 2021

Introduction

The Covid 19 pandemic greatly disrupted global markets, creating uncertainty in almost every sector. Stock prices plummeted in the spring of 2020, as the world grappled with a never-before-seen virus. Supply chains and travel were stalled, and demand for sanitation products skyrocketed. Many workers were forced to stay at home (with their similarly home-bound kids), and mundane errands to the store became potentially deadly. This fear and uncertainty are sure to have impacted the price of oil, among many other things. If we were to analyze oil price changes and their determinant factors, would a model that is used to predict fluctuations in oil prices pre-Covid 19 be applicable to the Covid 19 pandemic timeframe? Which determinant factors would stand-out between the periods, and how effective would these models be? This report aims to answer these questions.

The datasets for this analysis cover two separate time periods. The first is Pre-Covid 19 data, capturing information from September 2014 to December 2019. The second dataset examines the data from during the Covid 19 pandemic, from January 2020 through April 2021. The datasets both include 66 variables from a wide array of sources. Oil being such a ubiquitous commodity, it is helpful to cast a wide net in examining potential predictors. Stock prices, U.S. Treasury bond prices, as well as exchange rates are present. Variables of interest also include Bitcoin prices, global stock prices, clean energy stocks, and an economic uncertainty index. Out of the total 66 variables, there are 33 contemporaneous variables and 33 that are lagging. The lagged variables provide an opportunity to see any relationship between current prices and historical performance.

The analysis in this report has been performed in JMP Pro 16, using three methods: Standard Linear Regression (OLS), Stepwise Backward Regression, and Stepwise Forward Regression. With OLS, the model attempts to minimize the differences (squared) of the predicted and actual values. The Stepwise Backward and Stepwise Forward Regression methods take opposite approaches to each other. Stepwise Backward begins with all variables in the model and removes one at a time until no improvements can be made to the model's predictive capability. Stepwise Forward starts with no variables, adding one at a time until its predictive capability is optimized.

Analysis and Model Comparison

The models being used in this analysis have advantages and disadvantages. While OLS can be effective in finding linear relationships between variables, it has limitations if the relationships are too complex (non-linear). With a large number of variables being included in the analysis, OLS also tends to be sensitive to outliers in a variable's range. The alternative approaches with Backward and Forward Stepwise Regression are helpful in determining dependent variables. They each take incremental approaches to improving the final model. One concern with these methods is multicollinearity, since individually removing/adding variables might impact the inclusion/exclusion of other variables further down the chain. All three methods were applied to both data sets (Pre-Covid 19 and Mid-Pandemic), in order to determine any differences between the model traits.

The analysis of these datasets included a similar cross-validation procedure. 60% of the data was used to train each model. 20% of the data was used for Validation, and the remaining 20% was set aside for Testing. Since these datasets were time-series in nature, the cross-validated data was ordered chronologically. This helped the model project forward and prevented it from seeking-out causal relationships between more recent data and previous price changes.

The Pre-Covid Data consisted of 1332 observations. Rows 1-799 were used for Training, Rows 800-1065 were used for Validation, and Rows 1066-1332 were used for Testing. The Mid-Pandemic Data was comprised of 326 observations. Rows 1-195 were used for training, Rows 196-260 were used for Validation, and Rows 261-326 were used for Testing.

Hoping to predict the price of oil, RUSO was selected as the response variable. All of the other variables in the dataset were included as predictor variables. Using JMP 16 Pro, this analysis was conducted with all three methods. As predictions from each model were generated, they were saved and stored in new columns in the dataset. Represented below are the model comparisons for all three methods, performed on both datasets. The Holdback values of 0, 1, and 2 represent the Training, Validation, and Test Sets, respectively. Reviewing the individual components of the model comparison, notable metrics included RSquare (better if higher), RASE (better if lower), and AAE (better if lower).

Measures of Fit for RUSO (Pre-Covid 19)

Holdback	Predictor	Creator		RSquare	RASE	AAE	Freq
0	Pred Formula RUSO OLS	Fit Least Squares		0.6784	0.0134	0.0107	797
0	Pred Formula RUSO Stepwise Backward	Fit Least Squares		0.6775	0.0135	0.0107	797
0	Pred Formula RUSO Stepwise Forward	Fit Least Squares		0.6552	0.0139	0.0111	797
1	Pred Formula RUSO OLS	Fit Least Squares		0.4246	0.0129	0.0101	266
1	Pred Formula RUSO Stepwise Backward	Fit Least Squares		0.4371	0.0128	0.0100	266
1	Pred Formula RUSO Stepwise Forward	Fit Least Squares		0.4752	0.0123	0.0097	266
2	Pred Formula RUSO OLS	Fit Least Squares		0.4610	0.0151	0.0116	267
2	Pred Formula RUSO Stepwise Backward	Fit Least Squares		0.4628	0.0151	0.0116	267
2	Pred Formula RUSO Stepwise Forward	Fit Least Squares		0.4777	0.0148	0.0114	267

All of the models in the Pre-Covid 19 analysis fared relatively well. Each model had an RSquare value between 0.65-0.66 for the Training Set, between 0.42-0.43 for the Validation Set, and between 0.46-0.47 for the Test Set. RASE and AAE were all relatively low across all models and cross-validation groups, hovering near 0.01. The Stepwise Forward model performed marginally better in this group.

Measures of Fit for RUSO (Mid-Pandemic)

Holdback	Predictor	Creator		RSquare	RASE	AAE	Freq
0	Pred Formula RUSO OLS	Fit Least Squares		0.7064	0.0280	0.0201	193
0	Pred Formula RUSO Stepwise Backward	Fit Least Squares		0.7064	0.0280	0.0200	193
0	Pred Formula RUSO Stepwise Forward	Fit Least Squares		0.3134	0.0427	0.0259	194
1	Pred Formula RUSO OLS	Fit Least Squares		-1.244	0.0295	0.0242	65
1	Pred Formula RUSO Stepwise Backward	Fit Least Squares		-1.234	0.0294	0.0242	65
1	Pred Formula RUSO Stepwise Forward	Fit Least Squares		0.1112	0.0185	0.0148	65
2	Pred Formula RUSO OLS	Fit Least Squares		-0.908	0.0311	0.0240	67
2	Pred Formula RUSO Stepwise Backward	Fit Least Squares		-0.906	0.0311	0.0240	67
2	Pred Formula RUSO Stepwise Forward	Fit Least Squares		0.3316	0.0184	0.0150	67

The three models did not perform as uniformly with the Mid-Pandemic dataset. The OLS and Stepwise Backward models performed quite well on the Training Set, both having an RSquare value of 0.7064. The Stepwise Forward model did not perform as well on the Training Set. None of the models fared as well with the Validation Set. The RASE and AAE metrics revealed a larger error in these predictions, ranging from 0.01 to 0.03. In the end, Stepwise Forward Regression proved to be the best model in this case, performing significantly better than the other two on the Mid-Pandemic Test Set.

<u>Interpretation</u>

After choosing Stepwise Forward Regression as the model for both datasets, variable importance analysis provided additional insights. In the Pre-Covid dataset, a number of variables were shown to have an impact on the price of oil. RXLE, RFXC, and RSPY were the top three predictor variables, with parameter estimates of 1.26, 0.99, and -1.59, respectively.

The variable importance analysis of the Mid-Pandemic dataset showed, however, that RXLE was the only predictor variable still in-effect. Its parameter estimate in this dataset only came to 0.71. RXLE is an exchange-traded fund, focused on certain elements of the Energy Sector (https://finance.yahoo.com/quote/XLE). This makes sense, due to the energy sector's dependence on oil as a resource.

It is striking to see the comparison between the Pre-Covid 19 and Mid-Pandemic dataset profilers. Adjusting any of the predictor variables of the Pre-Covid 19 dataset shows an impact to other variables in the model, including the RUSO response variable. Having only one predictor variable in the Mid-Pandemic model results in any variation to RXLE simply impacting RUSO.