Christopher Dillard 9/3/2023

MIS 525

Module 1: Assignment 1 – Risk Factors for Oil Prices: Before and After Covid-19

In this assignment, I was tasked with Identifying the determinants of oil price changes in two datasets: one comprising data from 4 years leading up to the Covid-19 pandemic, and the other containing data from 16 months following the advent of the pandemic. These datasets each contain the same 100 variables, where the response variable is RUSO (the US Oil Fund), representing oil prices, and my goal is to run three different models on each dataset to see whether the determinants have shifted between the two periods. Rs and Ls are used in front of stock symbols to differentiate between contemporaneous effects (where R precedes a stocks name in the variables), and Ls are placed in front of the names to denote lagged effects. If a lagged effect is identified, this could explain how long it takes for market changes to take effect between changes in certain stock prices and the oil stock, particularly where market sentiment plays a role in prices if there is an anticipated drop/increase in market activity in relation to certain current events. In the context of the pandemic, barriers to trade and production losses where quarantine took effect, resulting in lower consumption, along with general uncertainty in the demand of oil given these assumptions could drive closely related stocks down. That being said, my intent was to see which determinants have a direct effect on the oil prices, as opposed to the other way around.

The three models that I employed are Standard Linear Regression, Stepwise Forward, and Stepwise Backward. In the case of a standard linear regression (listed as an ordinary least squares, or OLS, model in JMP), there are several assumptions made by the software that could result in less-than-ideal results if these are not appropriately accounted for. Primarily, OLS models assume linearity where it may not necessarily exist. Another potential issue is where multicollinearity exists between variables in the dataset. In other words, some predictor variables may be extremely correlated to each other which hinders the model's ability to accurately compute the individual impact of each predictor variable on the response variable. OLS also could tend to include irrelevant predictor variables since it includes all variables by default, something that the Stepwise models will be much better at avoiding. With the Stepwise forward model, predictor variables are added one-by-one and then ranked by their significance. This is extremely helpful when looking at variable selection, with the unfortunate disadvantage that the model fails to completely address multicollinearity. Unlike the previous model, in a Stepwise Backward model, JMP starts with all the predictor variables and then selects the best variables given their contribution to the model by iteratively eliminating less relevant variables until the model fails to improve. So, by nature, Stepwise Backward models might leave a model with more variables than the alternative Stepwise Forward model where the analyst must discern how relevant each variable really is since the potential for overfitting may lead a Backward model to create the illusion that it's better than the generally simpler Forward model. The difference means that an analyst must be knowledgeable on the data at hand, in order to determine whether multicollinearity is an issue.

All three of these models yield results that identify variables of importance, otherwise considered to be determinants of oil prices. By comparing pre-Covid data with post-Covid data, I could identify possible changes in the determinants between the two time periods. A positive relationship between the variables and prices entails an increase across the variables and a decrease across both the predictor and the response variable (oil prices). In other words, both will go in the same direction.

As a first step in preparing the data for analysis, I needed to create a cross validation column and split my data into training, validation, and testing sets, represented by 0,1, and 2, respectively. I used a 60-20-20 split on the data and measured the performance of the three models on how well they predicted the testing set, given that the models were not trained on this last 20% of the time series data, which would be a rational way to measure the forecasting capabilities of the models as true predictive models. Each model selects the variables that it determines to be the most significant in predicting fluctuations in oil prices.

Once all three models were created, I then compared them against each other by measuring how well they predicted the oil prices in the testing set. Specifically, I was interested in seeing which model yielded the highest R square value and the lowest values for both the Root Average Standard Error (RASE) and the Absolute Average Error (AAE). The idea here is that variables selected by the best performing models should be logical candidates as predictive variables in oil prices. Given the list of variables across both the Pre-Covid and Covid dataset, conclusions could be drawn to see how the determinants in oil prices might have shifted between the two time frames.

To determine which models were the best fit for the data, I looked at the model comparison feature to examine how well each one fared at making accurate predictions on the testing splits of their datasets. For both Pre-Covid and Covid data, the stepwise forward model performed the best at predicting the test data. In the Pre-Covid data this was demonstrated by the higher R Square (at 0.49) than the other models on the last split of our data, encoded by a 2 in the cross-validation column. In the Covid data, the R Square value was 0.31. These models greatly outperformed their counterparts, especially since the others yielded negative values for their R squares on the test data. Additionally, both Forward Stepwise models yielded smaller RASE and AAE values than the other models, which should ideally be kept smaller since they both measure error.

Given this information, I went back to the stepwise backward formulas to see which variables were selected as the most significant. I found that prior to Covid, several different variables were selected by JMP to be determinants of oil prices, with 11 in total.

To see which has the true highest effect on oil prices, I reran the stepwise formula and looked into the prediction profiler. The next step was exploring the option of assessing variable importance by their independent uniform inputs. This additional step yielded a list of the variables by their total effects on oil prices, which can be interpreted as how powerful each variable is as a determinant of oil prices. In the dataset for the few years prior to the Covid-19 pandemic, the two most impactful variables were RXLE and RSPY, with RXLE having a positive impact and RSPY, a negative one. By analyzing the independent uniform inputs on the best Pre-Covid model, I saw that RXLE accounted for 46% of the change in the oil prices and RSPY was not too far behind at 40%. All other variables in the model had effects less than 2%. When running a backward stepwise model on the Covid data, the only variable that JMP kept in the model was RXLE. This model produced an R Square of 0.33 on the test set, which is comparatively poorer than the best model on the Pre-Covid data, where the R Square was 0.4861.

To understand why RXLE and RSPY were identified as being determinants in oil prices, it is important to understand what each of them are. XLE is the Energy Select Sector SPDR Fund. This is easy to explain since oil is currently a large source of energy across the globe, where an increase in demand/profits of oil would logically lead to an increase in the valuation of a fund that directly relates to the performance of this commodity, at least in part. SPY (SPDR S&P 500 ETF Trust) represents the 500 best performing businesses. One potential reason behind a negative relationship here could be the increase in operating costs when the price of gasoline goes up. Intuitively, this increase would result in lower profits, with all other factors held constant.

I think that both models placing RXLE as the most impactful determinant on oil prices makes perfect sense. One thing to note, however, is that the vast difference in variables selected (going from 11 to only 1 between the two time periods), might have something to do with the difference in the quantity of data that the Pre-Covid model was trained on (799/1332 rows), compared to the Covid one (195/326 observations). So, the exclusion of RSPY from the Covid model might just be the result of there being far less data to train on and then to test on.