Joseph Whiting

Introduction to Analytics - Econ – 494

Project 2: Predictive Analytics Project – FA20

Dr. Levkoff

18th November 2020

## Executive summary:

This dataset was created for the explanatory analysis project 1, it contains 34 observations of quarterly changes in ETF returns for VOO, which is the Vanguard 500 index Fund ETF, this ETF tracks the movement of the S&P 500, VGT, Vanguard Information Technology, this ETF tracks the performance of a benchmark index that measures the investment return of stocks in the information technology sector, VXUS, which is Vanguard Total International Stock Index, this ETF tracks the performance of returns of stocks issued by companies located outside of the United States, Gross Domestic Product and Unemployment data was taken over the same period, as well as a categorical variable which contains the name of the President that was in office for that observation.

I will be looking at running regression models to see how well the variable of GDP can be used to predict how the ETF VOO will react and track in a similar pattern. I will use the out of sample error to check on how my models are working with this prediction by adding a linear variable, being GDP, and then squaring the GDP variable, then cubing it. I was unable to use a logarithm transformation of the variable as it contains negative numbers and a log of a negative number is undefined.

* Chart, line chart

  Description automatically generatedHere is a scatter plot showing the relationship between VOO and the GDP variable.
* The relationship is very positive showing a good correlation between the two variables until the last two observations where it turns negative. This was due to the COVID-19 pandemic and its effect on the GDP of the United States.

The first step after identifying my variables that will be used will be to partition my data into testing and training.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

## Proposal:

I will be conducting three linear regression models, using the GDP variable in three different processes to predict the same y-variable being VOO, which is my market tracking ETF. Having negative numbers in my variables doesn’t allow me to use a fourth regression using a logarithm transformation. Regression model two and three will be multiple linear regressions incorporating the quadratic and cubic transformations.

1. This will be the basic linear regression model consisting of using GDP to predict my y-variable of VOO.
   1. Code being used to run regression:
      1. M1 <- lm(VOO ~ GDP, Training) using the training data.
2. This will be using GDP and then incorporating a quadratic transformation of this variable to see if this creates a better prediction of the y-variable.
   1. Code being used to transform the variable and store it:
      1. Project\_Dataset$GDP2<-Project\_Dataset$GDP^2 #QUADRATIC TRANSFORMATION (2nd ORDER)
   2. Code being used to run regression:
      1. M2 <- lm(VOO ~ GDP + GDP2, Training)
3. This is be using GDP, GDP^2, and then incorporating a cubic transformation of the GDP variable to see if this helps the prediction of the y-variable.
   1. Code being used to transform the variable and store it:
      1. Project\_Dataset$GDP3<-Project\_Dataset$GDP^3 #CUBIC TRANSFORMATION (3rd ORDER)
   2. Code being used to run regression:
      1. M3 <- lm(VOO ~ GDP + GDP2 + GDP3, Training)

## Diagnostic Results:

1. Basic linear Regression model with only linear term: M1 <- lm(VOO ~ GDP, Training)
   1. A picture containing table

      Description automatically generatedRegression Summary

* Here we can see the summary results from the first regression.
* We can see that the GDP variable shows us that it is significant between the 0 and 0.001 level.
* The R-squared is 44.56% which will be our benchmark to compare the other regression models against.
* The adjusted R-squared is 41.92% which will be our benchmark.
  1. A picture containing text

     Description automatically generatedAssessing in and out of sample errors
* This will be the benchmark to compare the next models’ errors against.
* In sample error = 5.304518
* Our of sample error = 29.59982

1. First multiple linear regression model with linear and quadratic terms: M2 <- lm(VOO ~ GDP + GDP2, Training)
   1. Table

      Description automatically generatedRegression summary

* Here we can see the summary results from the second regression.
* The GDP variable is still significant between the 0 and 0.001 level.
* The quadratic version of the GDP variable is significant between 0,05 and 0.01.
* The R-squared is 52.85%, which is showing us that the fit of the model is increasing and getting better.
* The adjusted R-squared is 48.13% which has increase from the first model.
  1. A picture containing text

     Description automatically generatedAssessing the in and out of sample errors
* In sample error has dropped to 4.891847
* Out of sample error has increase to 109.8342
* Which is showing us that adding this quadratic variable has had a negative effect on the prediction.

1. Second multiple linear regression model with linear, quadratic and cubic terms: M3 <- lm(VOO ~ GDP + GDP2 + GDP3, Training)
   1. Table

      Description automatically generatedRegression summary

* Here we can see the summary results from the third regression.
* The only variable we see that has any notable significance is the first basic GDP variable and this is between 0.05 and 0.1.
* The R-squared is 53.02%, which is showing us that the fit of the model is increasing and getting better.
* The adjusted R-squared is 45.61% which has decreased from our second model, but above our first.
  1. Assessing the in and out of sample errors
* A picture containing text

  Description automatically generatedIn sample error has dropped to 4.882741.
* The out of sample error has decreased from model two to 51.42666 but is still above the out of sample error from model one.
* This shows us that adding the cubic transformation of the variable has help with the predictions from model two, but not compared to model one.

Chart, scatter chart

Description automatically generatedVisual Display of Regression Models:

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

## Error Comparison:

In-Sample Model Performance by Root Mean Square Error:

1. Basic linear Regression model with only linear term: M1 <- lm(VOO ~ GDP, Training)
   1. 5.304518
2. First multiple linear regression model with linear and quadratic terms: M2 <- lm(VOO ~ GDP + GDP2, Training)
   1. 4.891847
3. Second multiple linear regression model with linear, quadratic and cubic terms: M3 <- l,(VOO ~ GDP + GDP2 + GDP3, Training)
   1. 4.882741

Out-Of-Sample Model Performance by Root Mean Square Error:

1. Basic linear Regression model with only linear term: M1 <- lm(VOO ~ GDP, Training)
   1. 29.59982
2. First multiple linear regression model with linear and quadratic terms: M2 <- lm(VOO ~ GDP + GDP2, Training)
   1. 109.8342
3. Second multiple linear regression model with linear, quadratic and cubic terms: M3 <- lm(VOO ~ GDP + GDP2 + GDP3, Training)
   1. 51.42666

Here we can see that by looking at the out-of-sample RMSE that our first model is the best to use in order to predict our y-variable as it has the lowest error of 29.59982. While one would assume that adding more terms to your model would decrease this error in the case of this dataset it has had the opposite effect. We do see that when adding the cubic term, the error decreased, but not enough for it to show that the third model is better at predicting than the first. For the in-sample RMSE we have the lowest error for model three of 4.882741. What we see when comparing the out-of-sample and the in-sample errors is that the model that predicts the best for out-of-sample is model one, but with regards to the in-sample we see it would be model three.

# Conclusion:

In conclusion for this project and for the prediction we would use the first model as it has the lowest out-of-sample error which is the most important. This means that for this model our prediction is off by 29.59982% when predicting the return of the VOO ETF. We use out-of-sample error as it provides the same unit of measurement. This prediction might have shown a lower RMSE error if there were more observations in the dataset as it may increase the viability of the model and return a more accurate prediction.