Investing in the Stock Market:
The Health and Energy Industries vs Macroeconomic Indicators

By: Noah LeFevre, Gabriel Mercado, Valeria Ricaurte, Sid Vanam, Daisy Xue

November 18, 2021

ECON 573

Introduction

The problem statement is which macroeconomic factors affect the Health and Energy Industries?

This objective is to explain the role macroeconomic indicators play in predicting health and energy industry sector volatilities. The findings in this investigation would be advantageous for investors looking to diversify their portfolios and get exposure to these industries. It is helpful for investors because it will allow them to focus their limited time analyzing the macroeconomic factors that historically have affected these industries while ignoring or devoting less time to those with negligible effects. This paper will also determine if these industries are affected similarly or differently by the economy. If they are affected differently, then investing in both of them can become a hedging strategy. Otherwise, investors should consider other hedging strategies.

However, an important question arises: how do analysts explore the relationship between industries and macroeconomic factors? This paper will explore industry-specific Exchange Traded Funds (ETFs). According to Investopedia, an ETF is “a type of security that tracks an index, sector, commodity, or other assets, but which can be purchased or sold on a stock exchange the same way a regular stock can." ETFs are useful because they give a sample size of how many companies belong to the same industry - all by looking at just one number. Additionally, by examining the price history, daily returns can be computed by following the simple formula = (close price - open price)/open price. Open price is the price of the ETF set at the beginning of the day. The close price is the price of the ETF when the market closes at the end of the day. For example, if there are multiple negative daily returns in a row and a rise in inflation, we can conclude that inflation negatively affects the industry. We follow these methodologies to explore the relationship between industry and a set of macroeconomic factors.

Although daily returns are a way to inspect how macroeconomic factors affect the industry-specific ETF, this paper takes the analysis one step further by considering a volatility response variable in place of daily returns. The GARCH model is used to calculate this volatility response. GARCH stands for Generalized AutoRegressive Conditional Heteroskedasticity, but the inner workings of this model are beyond the scope of this paper. The most important thing to note is that this model takes a time series of returns as an input and gives a time series of volatility as an output. This time series is the response variable analyzed.

Data

The team used Yahoo Finance to source the public ETF prices and FRED (Federal Reserve Economic Data) to source the macro indicators. Specifically, we included observations from the beginning of 2006 through the end of 2020. Selecting a wide time frame window provides enough data points to build well-supported models.

The macroeconomic predictors obtained from FRED include the following:

- Crude Oil Prices: The price of crude oil, recorded daily.
- International Trade Weighted Dollar: The weighted average ratio of the foreign value of the U.S. dollar to the value of the currencies in a group of countries that the U.S. frequently trades, recorded daily.
- 2-Year Bond Yield: Daily data regarding the Market Yield in the U.S. Treasury Securities with 2-Year Maturity.
- Initial Jobless Claims: Weekly data regarding the number of people filing for unemployment.
- Chicago Fed National Financial Conditions Index (NFCI): This is an indicator for the U.S. financial conditions in money, debt, and equity markets, and also the traditional or "shadow" banking systems. Specifically, positive values indicate tighter than average financial conditions, while negative values indicate looser than average conditions, recorded weekly.
- Fed Balance Sheet: A weekly data that tracks the Federal Bank's total assets. It generally peaks in times of financial crisis.
- Treasury General Account: Weekly data regarding the primary operational account of the U.S. Treasury at the Federal Reserve. Virtually all U.S. government disbursements are from this account.
- M1: Monthly data consisting of currency outside the U.S. Treasury, Federal Reserve Banks, and the vaults of depository institutions; demand deposits at commercial banks and other checkable deposits.
- CPI: Monthly data regarding consumer prices or a method to track inflation.

- Unemployment Rate: Monthly data of workers in the labor force who are not employed but are actively looking.
- SP500: Market Index

The ETF response variables sourced from Yahoo Finance include:

- Vanguard Health ETF (VHT)
- Vanguard Energy ETF (VDE)

Since the data is a time series, it needs to be stationary. Stationarity in simple terms means that the mean and variance of the series are constant over time. For example, if a macroeconomic indicator increases over time, then the mean increases as well. That means the time series is non-stationary. To use different statistical methods and avoid unreliable method errors, we have to fix this. This paper uses the Augmented Dickey-Fuller Test to verify stationarity. For each variable (independent or dependent), the null hypothesis is that the series is non-stationary. The alternate hypothesizes that the series is stationary. After performing the tests, we rejected the nulls for initial jobless claims, VDE volatility, and VHT volatility.

To fix the non-stationary variables, we compute changes to remove any patterns or “detrend” the data. For example, instead of having a value x for CPI on 01/01/06 and another value on 01/02/06, we calculated the change and plugged it in for 01/02/06. Since we cannot calculate changes without a “before” or an “after”, we dropped the first and last observations. After computing the changes for all the variables, we performed the Augmented Dickey-Fuller Test once again and plotted the variables to confirm the deterrent. We rejected the null hypothesis for all the predictors and can now be certain that all variables are stationary.

Methods

Tree - Based Methods

Trees will be useful in this paper since one of their greatest advantages is interpretability and the purpose of this paper is to interpret which and how macroeconomic factors affect the health and energy industry. Trees involve “segmenting the predictor space into a number of simple regions and then it assigns the mean or mode response from the training observations in that region to a test observation”. In other words, the algorithm follows the following steps:

1. Dividing the predictor space into J distinct regions
2. All the observations that fall into the same regions, receive the same prediction. The mean is used in the case of regression and the mode is used in classification

Specifically, the algorithm tries to minimize the RSS:

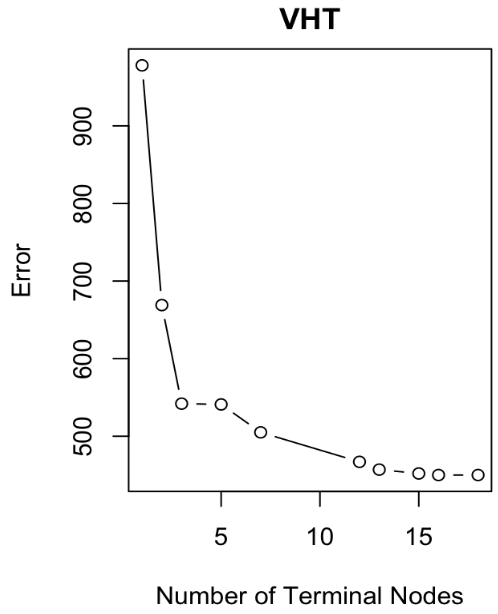
$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2,$$

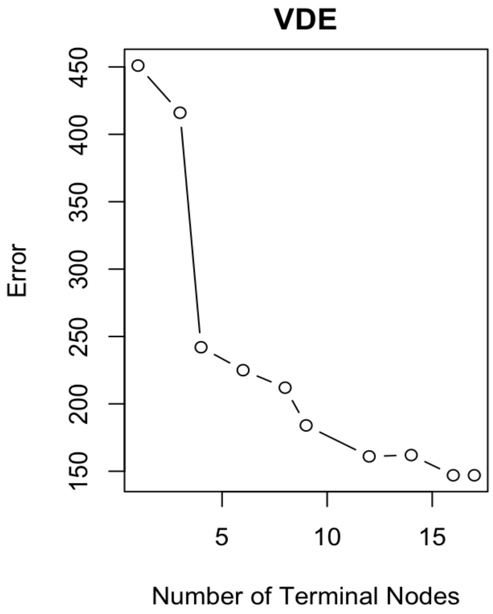
This formula sums the difference between the actual values and the predicted in all the regions and then sums the total differences for all regions. Ideally, a computer would consider all possible splits, but this is impossible. Therefore, beginning with all the data, the data looks for the predictor that reduces RSS the most, and then with these two regions repeats the decision.

As a result, the first predictor to split the data is the most important one in predicting or classifying.

This paper first analyzes volatility using a classification tree. Although this paper focuses on regression problems, we wanted to explore converting a continuous response into a binary one. In order to do this, the volatility of VHT and VDE was divided. After examining the range of both measures and looking at their plots, a threshold of 0.015 was chosen for VHT and 0.02 was chosen for VDE. This means that if the numbers are above 0.015 and 0.02 for VDE and VHT respectively, the response will be considered “High”. After this, we divided the data in a training and testing sample using a 50-50 allocation. We fitted the model to the training data and tested it with the testing sample. After this, we looked into all the tree splits to determine which predictors were the most important.

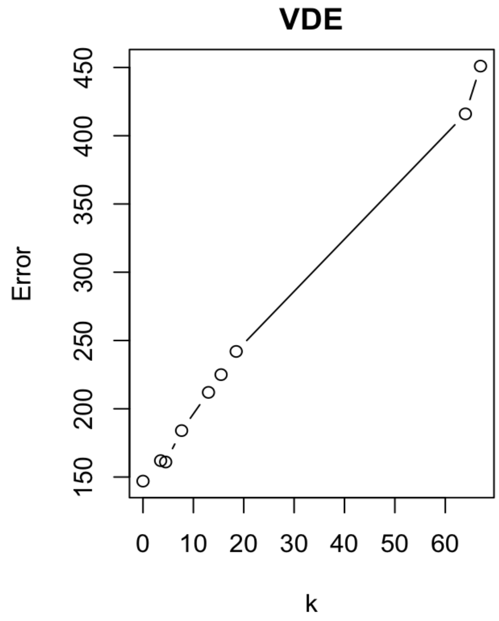
After fitting the tree, concerns about overfitting surfaced. If a tree is too “tall” then it may lead to overfitting. Therefore, we considered pruning the tree. We used cross-validation to see which complexity of tree is best. Here we can see how the error varies with the size:

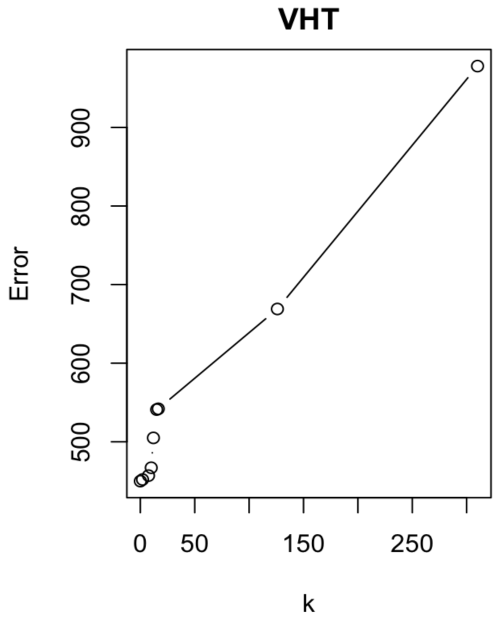




We can see that for both trees, the number of nodes that reduces the error is around 15.

Moreover, when we are pruning the tree, we are using a parameter k that "punishes" large trees. Here we can see how the error varies when changing k:





The error goes up when using large values of k because we are essentially cutting all the tree as the parameter increases. For both ETFs, the value that reduces the error is 0. All the branches are necessary. Also, for exploration purposes, we decided to prune the VDE tree to 6 terminal nodes.

Besides the classification tree, we used regression trees when using Bagging, Random Forest and Boosting. These methods take a series of regression trees to make a better prediction. When these methods make a better prediction, we can trust that the predictors they classify as important are relevant for the response variable.

In bagging, we are bootstrapping. This means we are taking repeated samples from training data to build different regression trees and then we average all of the predictions. Random forests are very similar to bagging with one difference: we only consider a subset of predictors when considering a split. This will improve the randomness of the trees. In bagging, our trees can end up looking similar to each other and therefore, have high correlation. If there is high correlation when we take the average, we would be increasing the variance. Finally, in boosting, we grow regression trees sequentially. One tree learns from the one that grew before it. Anything that one tree couldn't explain, the other one attempts to explain it. These trees also have a parameter that affects the rate at which they learn. When applying these methods we used a training set and a testing set as before with the classification tree.

Finally, we quantified how important the predictors are when applying Random Forest and Boosting. In Random Forest we look at how predictors reduce node impurity and in boosting we look at the influence of the predictors. The greater the reduction in impurity and the greater the influence, the better the predictor.

Linear Regression

There is an easy way to approach our question, which is linear regression. Linear regression is a statistical analysis method that uses regression analysis in mathematical statistics to determine the quantitative relationship between two or more variables. It is widely used. Its expression is

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

ε is a normal distribution whose error follows a mean value of 0.

In regression analysis, only one independent variable and one dependent variable are included, and the relationship between the two can be approximated by a straight line. This kind of regression analysis is called simple linear regression analysis. If the regression analysis includes two or more independent variables, and the relationship between the dependent variable and the independent variable is linear, it is called multiple linear regression analysis.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon,$$

In our case, we are using different indicators to see if they are useful to predict the volatility of the stock markets we chose.

Linear regression does not require complicated calculations. It still runs fast even with a large amount of data and can also give an understanding and explanation of each variable based on the coefficients. Such as, using p value to perform hypothesis tests. The most common hypothesis test involves testing the null hypothesis of

$$H_0 : \beta_1 = 0$$

which means there is no relationship between X and Y. And the alternative hypothesis of

$$H_a : \beta_1 \neq 0,$$

which means there is no relationship between X and Y.

But it has some limitations too. Linear regression is sensitive to outliers and noises, so it is easy to be overfitted, and easy to fall into local optimum. And it can't fit the nonlinear data well, so it is necessary to judge whether the relationship between the variables is linear.

For the accuracy test, we can use RSE to test the average amount that the response will deviate from the true regression line, we can use the formula as follow:

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}.$$

However, R^2 is a more efficient way to measure the fit by measuring the proportion of variability in Y that can be explained using X but independent from the scale of Y. An R^2 statistic close to 1 indicates that most of the variability in the response is explained by regression. When it is close to 0, that means the regression fails to explain most of the variability in the response. This could be because the linear model is wrong, or the error variance σ^2 is high, or both.

Lasso

Another useful method when trying to understand the impact of certain predictors on a response variable is the Lasso. The Lasso is a shrinkage statistical method utilized to find the most optimal subset of predictors. It relies on a penalizing variable, called lambda (λ) by convention. This lambda parameter penalizes large coefficients; hence it can be used to shrink the value of the model's covariates. Differently from the ridge method, the Lasso utilizes a l_1 penalty which essentially means that the optimal lambda parameter may force some coefficients to zero. Mathematically, the Lasso minimizes the following expression:

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j| \;=\mathrm{RSS}+\lambda\sum_{j=1}^{p}|\beta_j|.$$

From the formula one can see that lambda plays a key role in determining how large the coefficients will be, and it can even define what predictors might not be relevant by zero them out when lambda's value is large enough. Since this parameter is so essential to the model, its optimal value must be found before generating our Lasso model. Several approaches can be taken to obtain the best lambda value for one’s model such as Cross-Validation, AIC information criterion, and BIC information criterion. While the most common of this three is Cross-Validation, AIC will also be used to find the optimal lambda for this model as time series can sometimes produce inaccurate cross validation results.

The Cross-Validation approach obtains an optimal lambda by k-folding the data and running the regression on each fold. After the minimum optimal lambda is found, the value is used to obtain the lasso regression and its respective coefficients. The issue with time-series data and cross validation is that cross-validation partitioning and subsequent iterative modeling could alter the sequential order of one’s data.

RIDGE

Another regression tool of interest when hoping to gain insight into the factors that contribute to the volatility of our chosen ETFs is the ridge regression, which generates a linear regression model using L2 regularization, which involves minimizing the sum of squared residuals, subject to a tuning parameter lambda, which is equal to the squared magnitude of coefficients on predictors. Mathematically, the Ridge regression minimizes the following equation:

$$\sum_{i=1}^{M} (y_i - \hat{y_i})^2 = \sum_{i=1}^{M} \left(y_i - \sum_{j=0}^{p} w_j \times x_{ij} \right)^2 + \lambda \sum_{j=0}^{p} w_j^2$$

The result of this tuning parameter and regularization is a model that penalizes extremely high coefficient values, shrinking them closer to 0. This is particularly useful for data which suffers from multicollinearity, as is often common with asset prices and macroeconomic indicators. Regularization strategies like the ridge regression can improve models in these cases by bringing down variances by shrinking coefficient values closer to 0. The key to any regression strategy like this is a carefully selected tuning parameter, which can usually be selected via cross validation. In this case, which involves time series, minimum AIC selection is a better method for tuning our lambda value, so models were built around both methods.

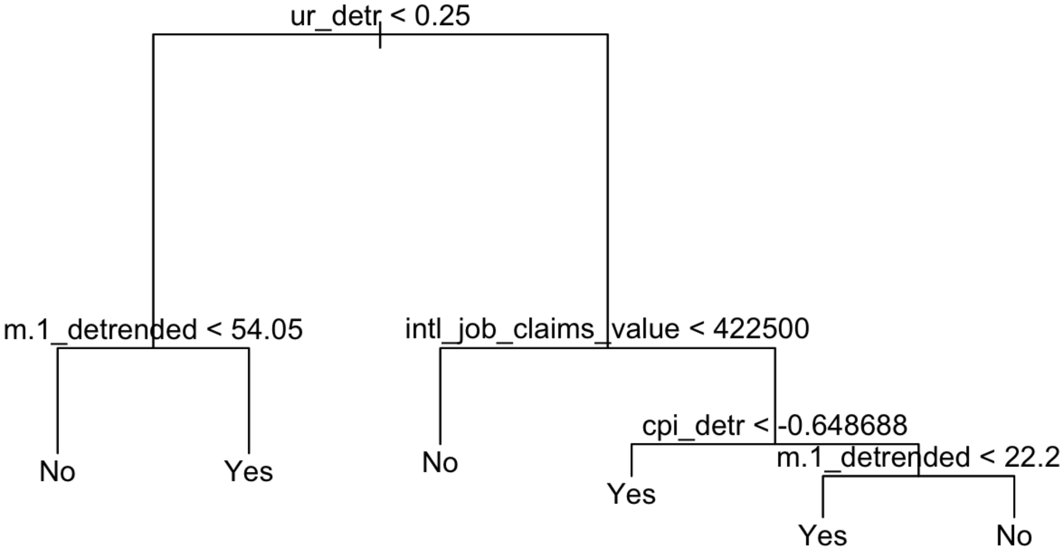
Results

Tree - Based Methods

After applying the classification tree, similar and different results were obtained for VDE and VHT. For both of them the most important predictor was the unemployment rate. However, after that, the most important predictor for VDE was the monetary base while the most important one for VHT was the Chicago Fed National Financial Conditions Index. Then, VDE moves on to the Chicago Fed National Financial Conditions Index and VHT moves to the Initial Jobless Claims. Although prediction is not our main goal, it is of interest to mention that our misclassification rate was 89% for VDE and 85% for VHT.

Another important highlight from the classification tree was found while pruning. When we cut the VDE classification tree at 6 terminal nodes, we found that although the overall accuracy of the model decreased from 89% to 86%, the model predicted better high volatility cases.

Increasing from 68% to 74%. This means that investors interested in predicting high volatility better, would be advised to prune the tree. Here is a figure of the 6 node tree:



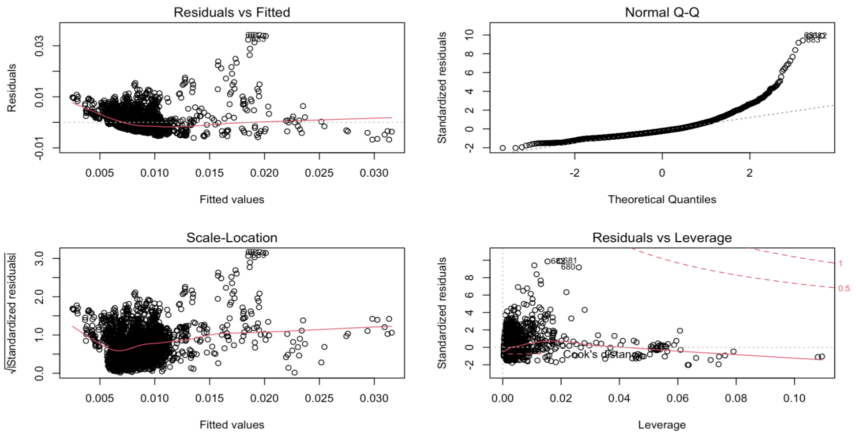
When applying Random Forests, we had extremely small MSEs for both ETFs and the most important predictors for VDE were the CPI, unemployment rate, Initial Jobless claims and the Chicago Fed National Financial Conditions Index. For VHT, the most important ones were CPI, Initial Jobless Claims and the Chicago Fed National Financial Conditions Index.

Finally, when applying boosting, the most important predictors for VDE were CPI, unemployment rate, the Chicago Fed National Financial Conditions Index and the Initial Jobless

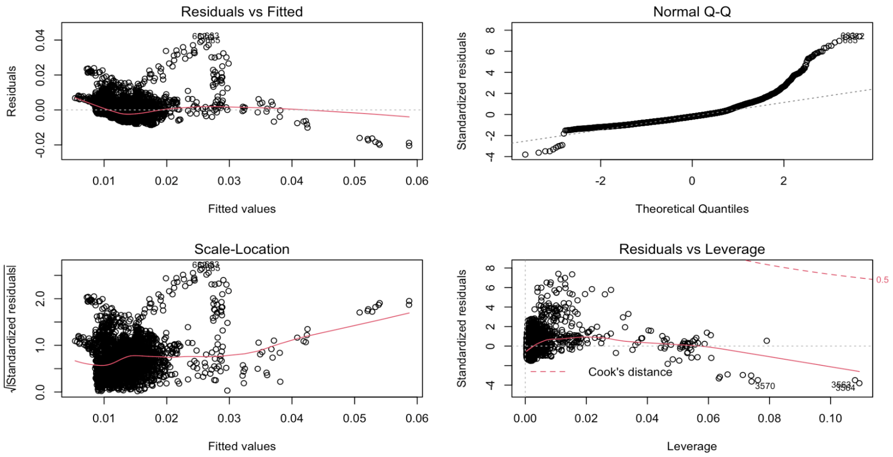
Claims. For VHT, the most important predictors were the same, but among the top for instead of including unemployment rate, it included the monetary base.

Linear Regression

I first built a model with all the predictors. After I built the model, I checked the plots of each database to check the linearity. For both Vanguard Energy ETF and Vanguard Health ETF, we can see the residuals vs fitted plots have a slight fan shape. This means our data is not spreading constantly and has heteroskedasticity. There are several outliers around 0.04 for VDE and 0.03 for VHT. Outliers from both models are pretty visible in the plots too. The data of the right-hand side are above the fitted line. This means the value x such that P(X<=x) = 0.99 is larger under the empirical CDF for the standardized residuals than it is under a normal distribution. However, for the scale-location plots, even though both have relatively horizontal red lines, the spread around red lines is not optimal. We still can see a strong homoscedasticity. Lastly, both VDE and VHT have decreasing trends of standardized residuals vs leverage, indicating heteroskedasticity. Also there are no dots outside the cook's distance.

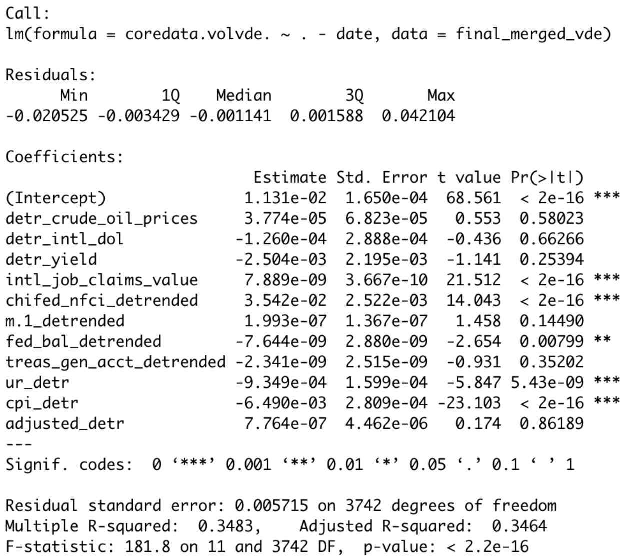


VHT

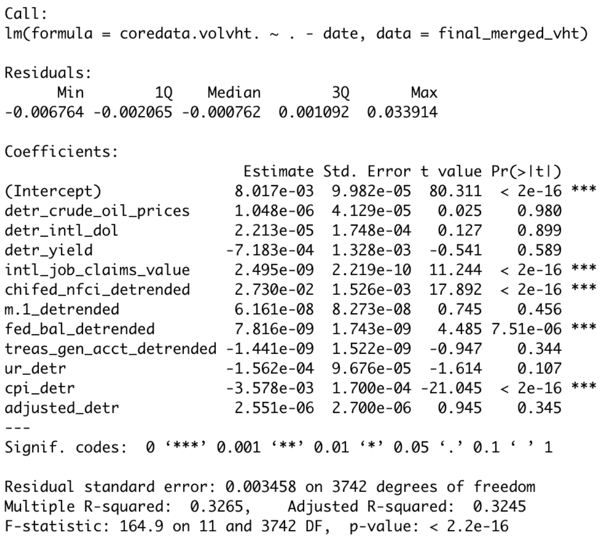


VDE

By looking at the summaries of the two models, we can see that the model of vht has a r square value as 0.3245, and the model of VDE has a R square of 0.3464, which are similar. As we discussed before, R-Squared measures how much variation of a dependent variable is explained by the independent variable(s) in a regression model. So our result is not optimal. It is better to do some modification of the data, or the variables. For VHT, only initial jobless claims, NFCI, federal balance and CPI have a p-value that are less than 0.05, which are able to reject the null hypothesis. This also means only them are the indicators that are important. Same as VDE, only initial jobless claims, NFCI, federal balance, unemployment rate and CPI have influences to the response variable volatility.

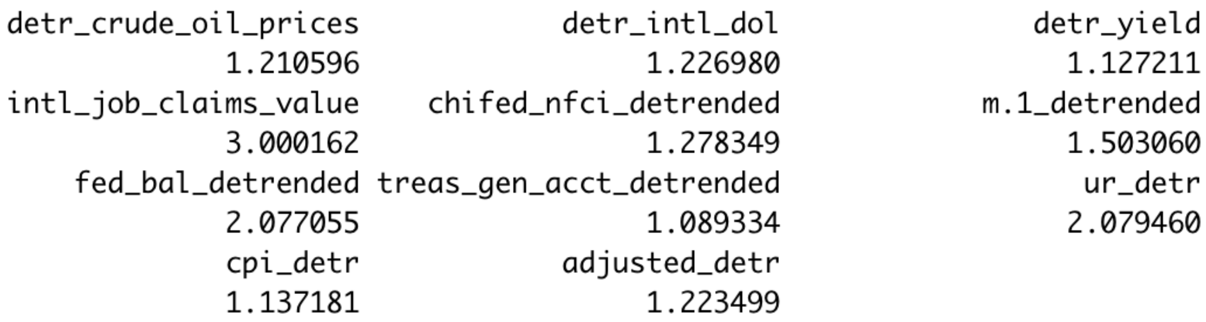


```
Call:
lm(formula = coredata.volvde. ~ . - date, data = final_merged_vde)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.020525 -0.003429 -0.001141  0.001588  0.042104 

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)    
(Intercept)               1.131e-02  1.650e-04  68.561  < 2e-16 ***
detr_crude_oil_prices     3.774e-05  6.823e-05   0.553  0.58023    
detr_intl_dol            -1.260e-04  2.888e-04  -0.436  0.66266    
detr_yield               -2.504e-03  2.195e-03  -1.141  0.25394    
intl_job_claims_value     7.889e-09  3.667e-10  21.512  < 2e-16 ***
chifed_nfci_detrended     3.542e-02  2.522e-03  14.043  < 2e-16 ***
m.1_detrended             1.993e-07  1.367e-07   1.458  0.14490    
fed_bal_detrended        -7.644e-09  2.880e-09  -2.654  0.00799 ** 
treas_gen_acct_detrended -2.341e-09  2.515e-09  -0.931  0.35202    
ur_detr                  -9.349e-04  1.599e-04  -5.847 5.43e-09 ***
cpi_detr                 -6.490e-03  2.809e-04 -23.103  < 2e-16 ***
adjusted_detr             7.764e-07  4.462e-06   0.174  0.86189    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.005715 on 3742 degrees of freedom
Multiple R-squared:  0.3483,	Adjusted R-squared:  0.3464 
F-statistic: 181.8 on 11 and 3742 DF,  p-value: < 2.2e-16
```



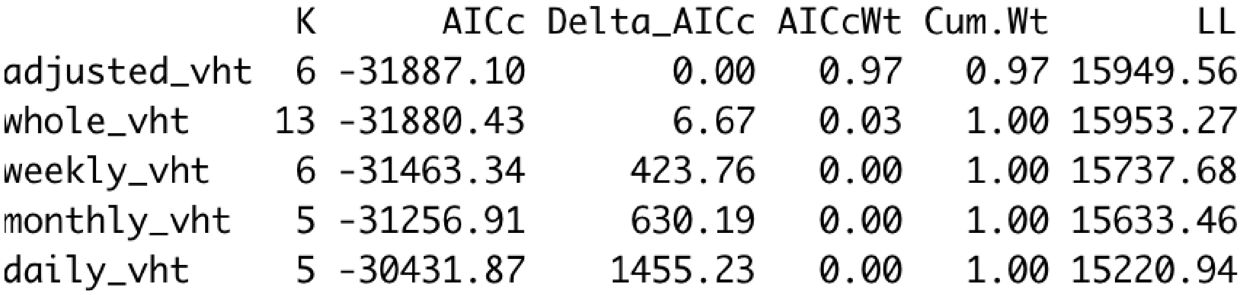
```
Call:
lm(formula = coredata.volvht. ~ . - date, data = final_merged_vht)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.006764 -0.002065 -0.000762  0.001092  0.033914 

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)    
(Intercept)               8.017e-03  9.982e-05  80.311  < 2e-16 ***
detr_crude_oil_prices     1.048e-06  4.129e-05   0.025    0.980    
detr_intl_dol             2.213e-05  1.748e-04   0.127    0.899    
detr_yield               -7.183e-04  1.328e-03  -0.541    0.589    
intl_job_claims_value     2.495e-09  2.219e-10  11.244  < 2e-16 ***
chifed_nfci_detrended     2.730e-02  1.526e-03  17.892  < 2e-16 ***
m.1_detrended             6.161e-08  8.273e-08   0.745    0.456    
fed_bal_detrended         7.816e-09  1.743e-09   4.485 7.51e-06 ***
treas_gen_acct_detrended -1.441e-09  1.522e-09  -0.947    0.344    
ur_detr                  -1.562e-04  9.676e-05  -1.614    0.107    
cpi_detr                 -3.578e-03  1.700e-04 -21.045  < 2e-16 ***
adjusted_detr             2.551e-06  2.700e-06   0.945    0.345    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.003458 on 3742 degrees of freedom
Multiple R-squared:  0.3265,	Adjusted R-squared:  0.3245 
F-statistic: 164.9 on 11 and 3742 DF,  p-value: < 2.2e-16
```

We can also test the correlation between the independent variable and the other variables by VIF. VIF starts at 1 and has no upper limit. When VIF = 1, no correlation between the independent variable and the other variables. From VIF of VHT, we have the similar result as the p value. Initial jobless claims, federal balance and unemployment rate have a higher VIF value, which means they have more correlation to the response variable. For VIF of VDE, we have the same result.



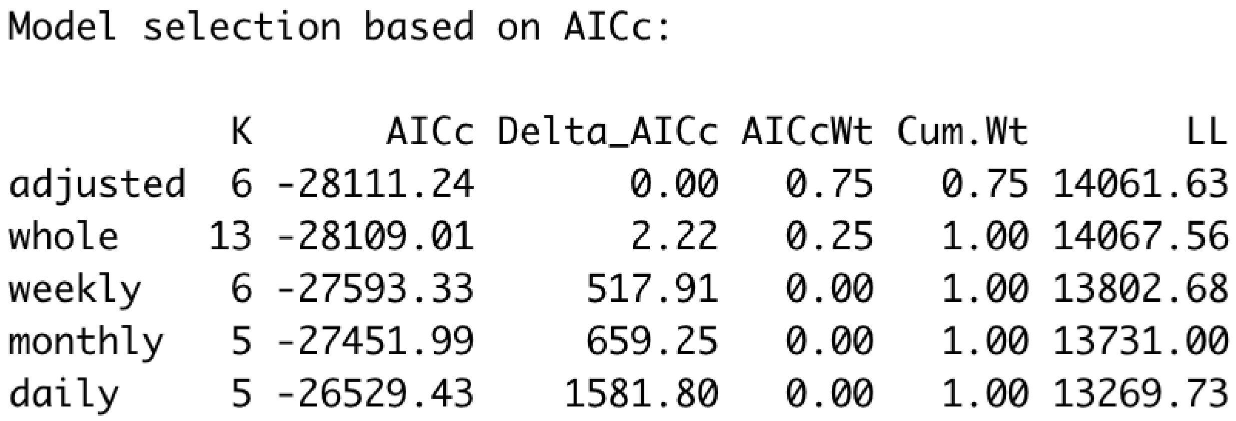
```
   detr_crude_oil_prices            detr_intl_dol               detr_yield
                1.210596                 1.226980                 1.127211
   intl_job_claims_value    chifed_nfci_detrended           m.1_detrended
                3.000162                 1.278349                 1.503060
      fed_bal_detrended treas_gen_acct_detrended                 ur_detr
                2.077055                 1.089334                 2.079460
                cpi_detr             adjusted_detr
                1.137181                 1.223499
```

With this knowledge, I created some new models aiming to do a better prediction. First I only use the indicators that have p-values that are less than 0.05 to do the prediction(adjust). Then I use daily variables(daily), weekly variables(weekly) and monthly variables(monthly) as my predictors. I use AIC to determine the best fit model for the data. The Akaike information criterion (AIC) is a metric that is used to compare the fit of several regression models. It is calculated as: AIC = 2K – 2ln(L). Not surprisingly, both VDE and VHT have the highest AIC value in “adjusted” as we can see in the tables below.



```
              K      AICc Delta_AICc AICcWt Cum.Wt       LL
adjusted_vht  6 -31887.10       0.00   0.97   0.97 15949.56
whole_vht    13 -31880.43       6.67   0.03   1.00 15953.27
weekly_vht    6 -31463.34     423.76   0.00   1.00 15737.68
monthly_vht   5 -31256.91     630.19   0.00   1.00 15633.46
daily_vht     5 -30431.87    1455.23   0.00   1.00 15220.94
```

VHT



```
Model selection based on AICc:

          K      AICc Delta_AICc AICcWt Cum.Wt       LL
adjusted  6 -28111.24       0.00   0.75   0.75 14061.63
whole    13 -28109.01       2.22   0.25   1.00 14067.56
weekly    6 -27593.33     517.91   0.00   1.00 13802.68
monthly   5 -27451.99     659.25   0.00   1.00 13731.00
daily     5 -26529.43    1581.80   0.00   1.00 13269.73
```

VDE

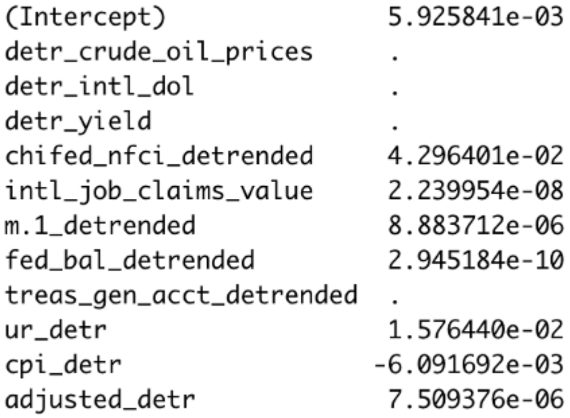
To be summarized, even though we could find a best fit model using linear regression, it is still not an ideal method to analyze the problem. Since the data does not completely follow the rules of linearity, it is better to do some transformation for the data itself.

Lasso

After applying the Lasso method on both the volatility measures of Vanguard Energy ETF and Vanguard Health Care ETF several interesting results were found. However, before going into the quantitative results, it is necessary to remark that two different model findings will be presented. First, we will show the results from applying Lasso with an optimal lambda obtained from a Cross-Validation process. Following this, we will show a very different set of numbers that came from performing the Lasso with an optimal lambda according to model comparisons based on AIC criterion.

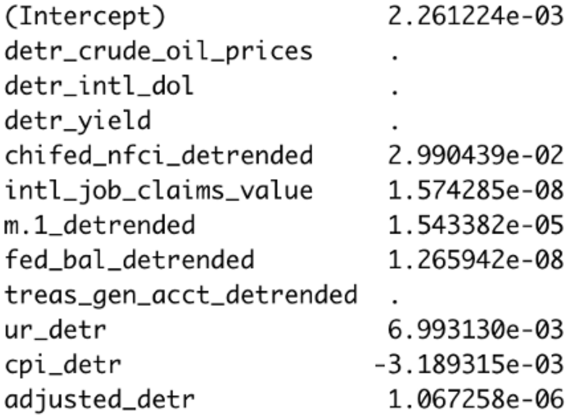
The lambda obtained from Cross-Validation approach was 0.0001221047 and 0.0001221047 for the Vanguard Energy ETF and Vanguard Health Care ETF respectively. Using these values of lambdas, the following coefficients where obtained:

Vanguard Energy



```
(Intercept)                5.925841e-03
detr_crude_oil_prices      .
detr_intl_dol              .
detr_yield                 .
chifed_nfci_detrended      4.296401e-02
intl_job_claims_value      2.239954e-08
m.1_detrended              8.883712e-06
fed_bal_detrended          2.945184e-10
treas_gen_acct_detrended   .
ur_detr                    1.576440e-02
cpi_detr                  -6.091692e-03
adjusted_detr              7.509376e-06
```

Vanguard Health Care



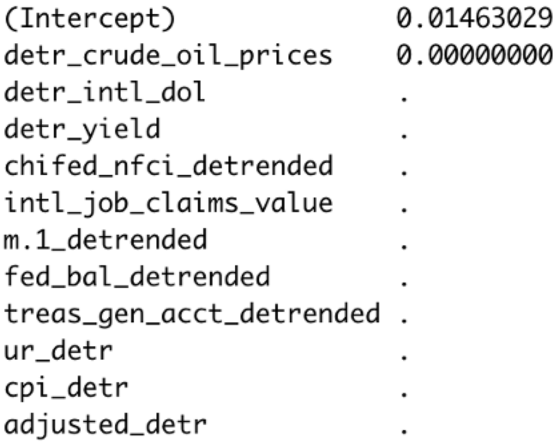
```
(Intercept)                2.261224e-03
detr_crude_oil_prices      .
detr_intl_dol              .
detr_yield                 .
chifed_nfci_detrended      2.990439e-02
intl_job_claims_value      1.574285e-08
m.1_detrended              1.543382e-05
fed_bal_detrended          1.265942e-08
treas_gen_acct_detrended   .
ur_detr                    6.993130e-03
cpi_detr                  -3.189315e-03
adjusted_detr              1.067258e-06
```

For both ETFs, crude oil prices, international traded dollars, the 2-year treasury yield, and the U.S. Treasury General Account, all zero out of the regression. This means that with the lambdas shown above, these predictors do not significantly explain anything about the variation on the volatility measure for the ETFs in question. We can also conclude from these numbers that all of our remaining predictors have a positive effect on volatility except for the CPI indicator. In other words, as all those covariates increase the volatility measure also increases. Finally, a table showing some model statistics is shown below:

Statistics	MSE	In Sample R^2	AIC	Optimal Lambda
VHT	1.097067e-05	0.5556351	14.02059	0.0001221047
VDE	3.07041e-05	0.5380359	14.05763	0.0001221047

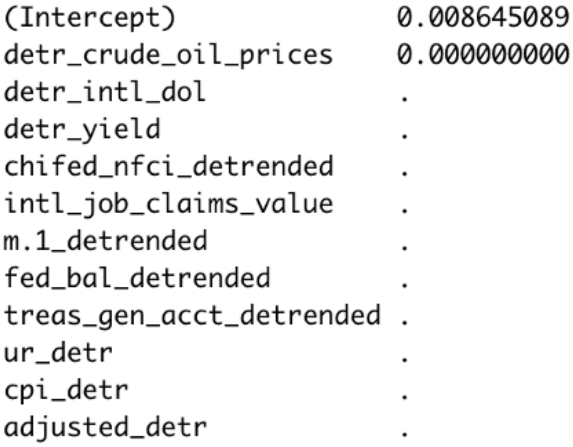
On the other hand, when we calculated lambda by iteratively comparing AIC criterions for regressions with different lambda values, we found that the lowest AIC for Vanguard Health Care ETF, 0.04634018, was reached for the first time at a lambda value of approximately 0.01001. Similarly, for the Vanguard Energy ETF the lowest AIC, 0.1247534, was obtained when performing a model with lambda equal to 0.00501. Subsequently, a final model was created using these optimal lambda values according to our AIC method. The following coefficients where obtained:

Vanguard Energy



```
(Intercept)               0.01463029
detr_crude_oil_prices     0.00000000
detr_intl_dol             .
detr_yield                .
chifed_nfci_detrended     .
intl_job_claims_value     .
m.1_detrended             .
fed_bal_detrended         .
treas_gen_acct_detrended .
ur_detr                   .
cpi_detr                  .
adjusted_detr             .
```

Vanguard Health Care



```
(Intercept)               0.008645089
detr_crude_oil_prices     0.000000000
detr_intl_dol             .
detr_yield                .
chifed_nfci_detrended     .
intl_job_claims_value     .
m.1_detrended             .
fed_bal_detrended         .
treas_gen_acct_detrended .
ur_detr                   .
cpi_detr                  .
adjusted_detr             .
```

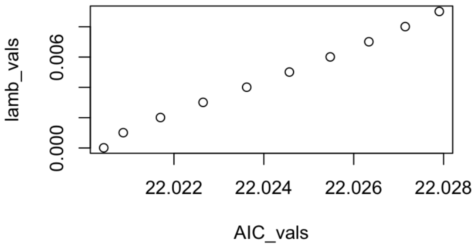
Clearly, this approach is telling us that the best model based on Akaike's Criterion is one where all predictors are zero out. This means that, following this method, none of our macroeconomic indicators seem to have any real effect on the volatility of either of the ETFs studied in this project. A plausible explanation to these results could be the fact that volatility measures are very hard to predict. Since many factors can affect the returns of equities, it might be the case that the combination of these macro-indicators to predict volatility does a worse job than having none of them.

In summary, the lasso method gave us two different behaviors depending on the approach taken to calculate the lambda parameter. However, taking into consideration that we worked with timed series data the results from the AIC method are probably more accurate than those obtained from cross-validation.

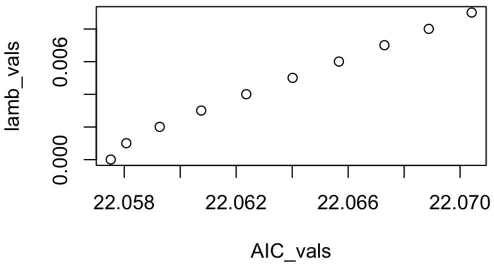
RIDGE

With our lambda value selected via AIC as shown below, coefficients were extracted for each ridge regression, with the results being displayed in the tables below.

Vanguard Energy

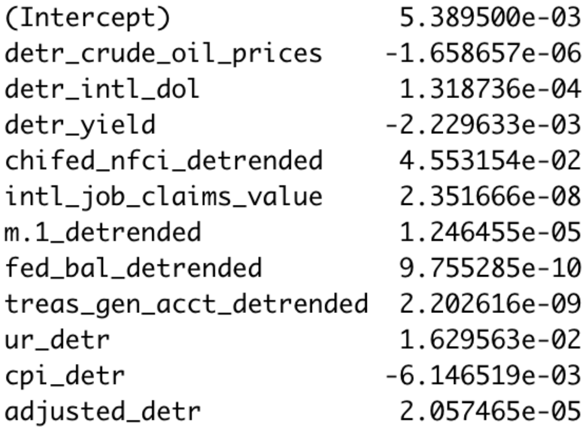


Vanguard Health Care



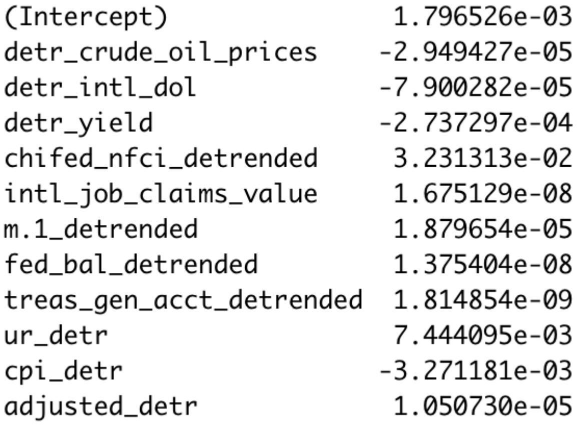
Ridge Lambda selection for VDE and Ridge Lambda selection for VTE

Vanguard Energy



```
(Intercept)                  5.389500e-03
detr_crude_oil_prices       -1.658657e-06
detr_intl_dol                1.318736e-04
detr_yield                  -2.229633e-03
chifed_nfci_detrended        4.553154e-02
intl_job_claims_value        2.351666e-08
m.1_detrended                1.246455e-05
fed_bal_detrended            9.755285e-10
treas_gen_acct_detrended     2.202616e-09
ur_detr                      1.629563e-02
cpi_detr                    -6.146519e-03
adjusted_detr                2.057465e-05
```

Vanguard Health Care



```
(Intercept)                  1.796526e-03
detr_crude_oil_prices       -2.949427e-05
detr_intl_dol               -7.900282e-05
detr_yield                  -2.737297e-04
chifed_nfci_detrended        3.231313e-02
intl_job_claims_value        1.675129e-08
m.1_detrended                1.879654e-05
fed_bal_detrended            1.375404e-08
treas_gen_acct_detrended     1.814854e-09
ur_detr                      7.444095e-03
cpi_detr                    -3.271181e-03
adjusted_detr                1.050730e-05
```

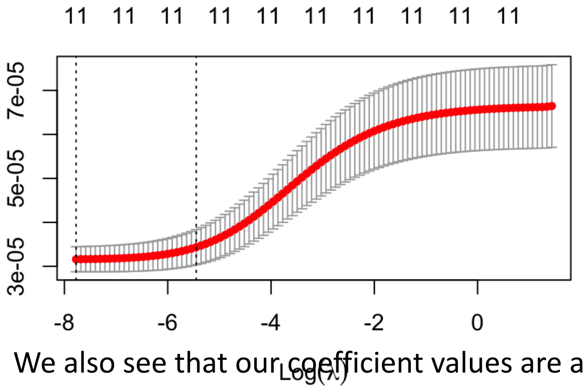
While the coefficients between the two ETFs are similar, they do differ in magnitude for each predictor. It is also interesting to see that while the volatility of the two ETFs seem to move in the same direction for most predictors, the sign on the variable for stationary international dollar value is different between the two, suggesting the effect is different between the two ETFS. Overall, we can see that the volatility of the two ETFs can be explained by similar coefficients. Notable coefficients are that of crude oil prices, the magnitude of which is higher for VDE, an ETF tracking the energy market. This is consistent with expectations, which would suggest that the volatility of an ETF composed of energy assets would respond more to changes in oil prices than on composed of assets in the healthcare industry. Plots for the lambda model are also included, showing how the MSE varies as the lambda changes over a range of values. The signs for many of our coefficient values are consistent with economic theory, with the NFCI value being a good example, with a positive sign. This indicator is set to have a mean 0 and standard deviation of 1, and is designed to reflect economic market conditions, with higher values being associated with more stringent economic conditions. This positive sign indicates that as the value rises, so does volatility in both indexes. Below are the summary statistics for the Ridge regression outputs, reflecting the lambda value chosen via AIC. The R^2 value for our regressions suggests that the two equations explain around 56% and 54% of variation in volatility in VHT and VDE respectively.

Statistics	MSE	In Sample R^2	Optimal Lambda	AIC
VHT	0.02043762	0.5594422	1e-05	22.02044

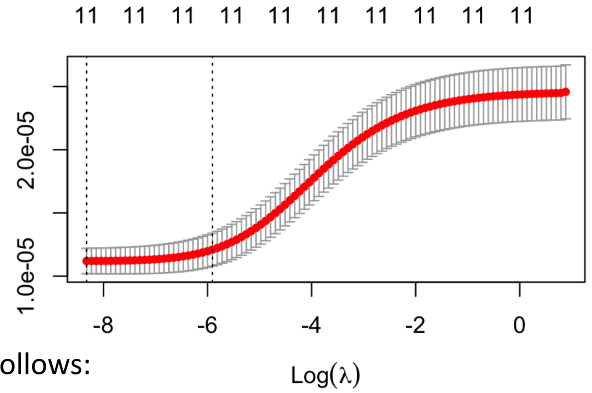
VDE	0.05751898	0.5408823	1e-05	22.05752

When our tuning parameter lambda is selected via cross-validation, we find the lambda values to be 0.0004205551 and 0.0002421746 for VDE and VHT respectively.

Vanguard Energy



Vanguard Health Care



We also see that our coefficient values are as follows:

Vanguard Energy

```
(Intercept)                 5.741436e-03
detr_crude_oil_prices      -9.474560e-06
detr_intl_dol               1.310978e-04
detr_yield                 -2.037072e-03
chifed_nfci_detrended       4.226466e-02
intl_job_claims_value       2.252168e-08
m.1_detrended               1.410936e-05
fed_bal_detrended           3.601176e-09
treas_gen_acct_detrended    1.908549e-09
ur_detr                     1.563695e-02
cpi_detr                   -5.947963e-03
adjusted_detr               1.825861e-05
```

Vanguard Health Care

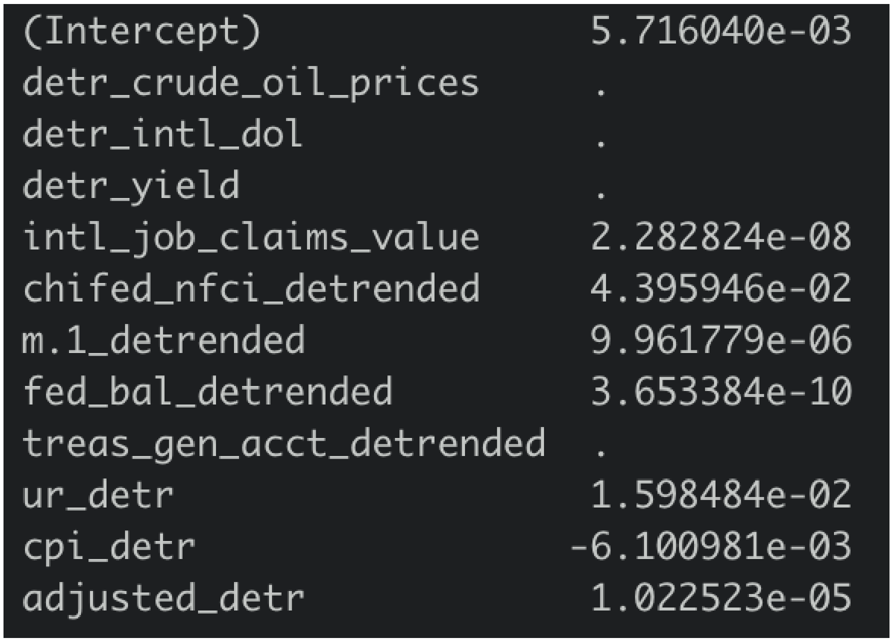
```
(Intercept)                 2.114481e-03
detr_crude_oil_prices      -3.190046e-05
detr_intl_dol              -6.481427e-05
detr_yield                 -2.598439e-04
chifed_nfci_detrended       3.030211e-02
intl_job_claims_value       1.592126e-08
m.1_detrended               1.943932e-05
fed_bal_detrended           1.461273e-08
treas_gen_acct_detrended    1.702155e-09
ur_detr                     7.223037e-03
cpi_detr                   -3.199567e-03
adjusted_detr               9.431666e-06
```

From This we can see that while the magnitude may change, the sign on the coefficients is the same whether estimating the optimal lambda by cross-validation or by AIC. In terms of model evaluation, the R^2 value for our cross-validated lambda model for VDE was 0.53998, and for VHT was 0.55781, relatively similar to the values found using AIC criterion.

Elastic Net

The Elastic Net method is a unique method that blends both the lasso and ridge regression. Specifically, it has the lambda and alpha parameters. Alpha specifies the balance between ridge and lasso regressions where when alpha is zero it mimics the ridge regression. Contrastly, when the alpha is one it mimics the lasso regression. Lambda specifies the level of regularization that the method employs. The Elastic Net method provides lots of flexibility but can be costly as far as parameter optimization is concerned.

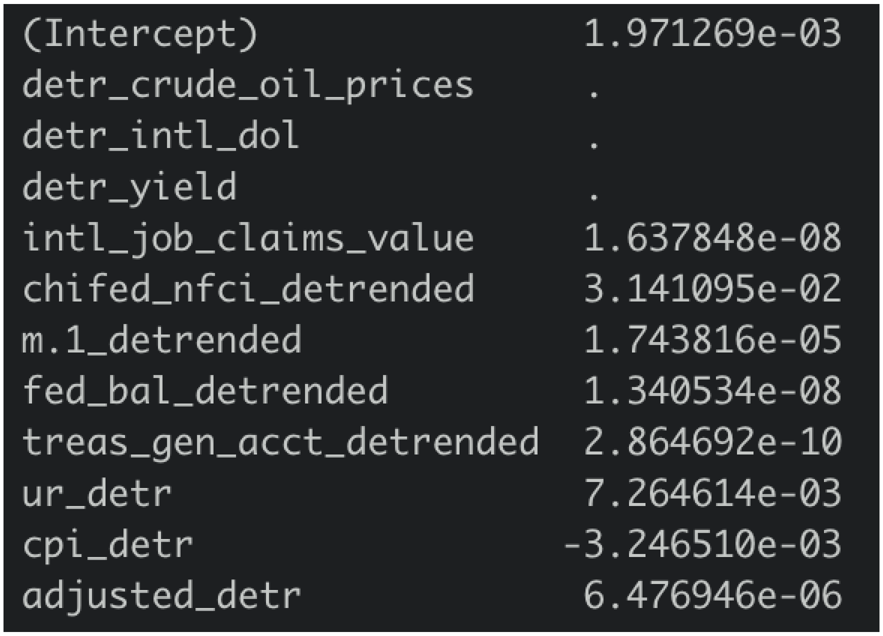
When predicting the volatility for the VDE ETF we see the following output:



```
(Intercept)                  5.716040e-03
detr_crude_oil_prices        .
detr_intl_dol                .
detr_yield                   .
intl_job_claims_value        2.282824e-08
chifed_nfci_detrended        4.395946e-02
m.1_detrended                9.961779e-06
fed_bal_detrended            3.653384e-10
treas_gen_acct_detrended     .
ur_detr                      1.598484e-02
cpi_detr                    -6.100981e-03
adjusted_detr                1.022523e-05
```

From the coefficients above we see that Chifed NFCI predictor, unemployment rate, and CPI have the biggest impact in predicting volatility as per the magnitude.

When predicting the volatility for the VHT ETF we see the following output:



```
(Intercept)                  1.971269e-03
detr_crude_oil_prices        .
detr_intl_dol                .
detr_yield                   .
intl_job_claims_value        1.637848e-08
chifed_nfci_detrended        3.141095e-02
m.1_detrended                1.743816e-05
fed_bal_detrended            1.340534e-08
treas_gen_acct_detrended     2.864692e-10
ur_detr                      7.264614e-03
cpi_detr                    -3.246510e-03
adjusted_detr                6.476946e-06
```

After testing the model with cross-validation to find the optimal alpha and lambda parameters we use the best parameters to find the best model and therefore the most reliable predictors. From the coefficients above we see that Chifed NFCI predictor, unemployment rate, and CPI have the biggest impact in predicting volatility as per the magnitude. When comparing the two ETFs we see that three most impactful market indicators are the same. Furthermore, since these indicators are the same, we can note that they are both valuable in predicting volatility for both industries. It's also important to note that both Elastic Net models had an alpha parameter of 1 which follows the ridge regression.

Conclusion

While all regression methods produced different models, it is important to acknowledge that the coefficients between the two ETFs were relatively consistent within each method. While magnitude of coefficients varied slightly, the sign on each coefficient was almost entirely the same, indicating how the volatility of these ETFs may react differently to different factors, they move in similar directions. From this we can determine that these ETFs have similar risks, meaning that neither is particularly useful for investors looking to diversify assets in a portfolio as a way of minimizing exposure to risk.

References

https://www.investopedia.com/terms/e/etf.asp

https://web.stanford.edu/~hastie/ISLR2/ISLRv2_website.pdf

https://towardsdatascience.com/ridge-and-lasso-regression-a-complete-guide-with-python-scikit-learn-e20e34bcbf0b

https://finance.yahoo.com/