ERS

What Drives WTI Futures Returns?

BY: JULIAN MURILLO, LINDSEY BENSON, REHJII MARTIN, & ISABELLA MORROW

Link to Video Presentation: https://www.youtube.com/watch?v=9147NfjKp2A

INTRODUCTION

Goal : Create the best possible model from a set of given researched variables to most effectively predict WTI Crude Oil Futures prices.

Analyzing the crude oil futures returns predictability from 2000 to 2019.

The oil market is very volatile and difficult to forecast.

LITERATURE REVIEW

Key Points:

Steel Production and the price of Oil are very significant.

Fluctuations in oil prices can be explained by oil production, oil inventories and oil demand.

Renewable energy and sustainability are major contributors to volatile oil markets.

Natural gas and crude oil share direct connection with regards to supply and demand. Some found the significant relationship has weakened overtime.

CONCEPTUAL FRAMEWORK

Our model consisted of 27 x-variables upon initial modeling.

Models Utilized:

LASSO (Regularization)

Elastic Net Model (Regularization)

Variance Inflation Factor (VIF)

Ordinary Least Squares (OLS)

Autoregressive Integrated Moving Average (ARIMA)

Logistic Regression Models

DATA DESCRIPTION

First selected data from class drive only 27 variables survived initial selection due to NAs and lack of information. We also gathered our own datasets to bring in more variables.

Data parameters were from Jan 1st 2000 to August 2019.

Data was all checked for stationarity and collinearity and multicollinearity upon first computations.

There was no seasonality: checked using a plot (decomposed function) and an ADF test.

The initial OLS model was heteroskedastic even after filtering a bunch of variables, so then we decided to do K-Fold.

Variables

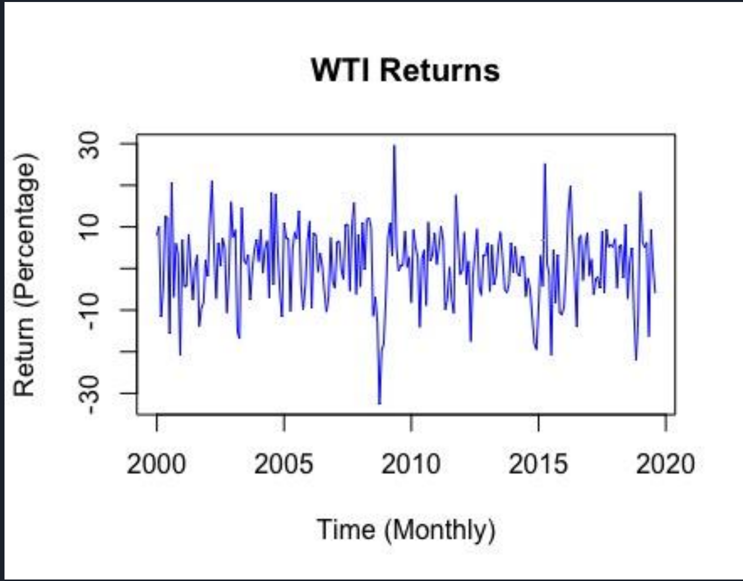
- WTI_return (Target)
- ppi_oil
- cpi_market_basket
- import_opec
- export_opec
- foss_fuel_prod
- foss_fuel_cons
- renew_energy_prod
- steel_prod
- oil_prod
- b_com
- us_rig
- n_gas
- ag_price_index

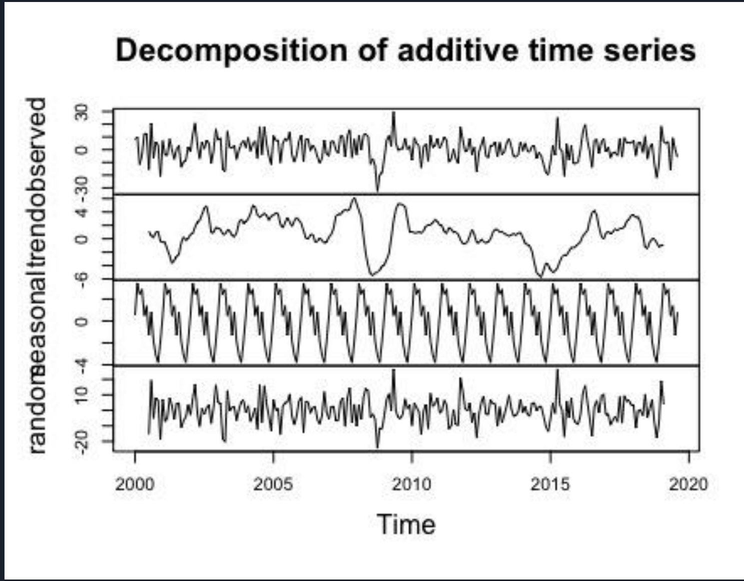
- renew_energy_cons
- primary _energy_prod
- primary_energy_cons
- primary_energy_import_net
- us_unemployment
- price_gold
- tb_rate
- usd_index
- inflation_rate
- fed_funds_rate
- nasdaq_comp
- sp_500
- us_gasoline
- supply_m2

Data Key:

WTI_return: Average (monthly) Return
ppi_oil: Producer Price Index Oil Manufacturing
cpi_market_basket: Consumer Price Index (USD) Standard Market Basket of Goods
import_opec: Imports from OPEC countries
export_opec: Exports to OPEC countries
foss_feul_prod: Total Fossil Fuel Production (US)
foss_feul_cons: Total Fossil Fuel Consumption (US)
renew_energy_prod:Total Renewable Energy Production (US)
steel_prod:Total Renewable Energy Consumption (US)
oil_prod:Total Primary Energy Production (US)
primary_energy_cons:Total Primary Energy Consumption (US)
primary_energy_import_net:Net Imports from the world to the US (Primary energy)
us_umeployment:Unemployment Rate
price_gold:Price of Gold Per Troy Ounce
tb_rate:3-Month Treasury Bond yield rate
usd_index:US Dollar Weighted Index: Exchange Rate Index
inflation_rate:US Inflation Rate
fed_funds_rate:US Federal Funds Rate
nasdaq_comp:Index of nasdaq securities (Prices)
sp_500:Index of S&P 500 stock prices
steel_prod:American Steel Production (Million Tonnes Crude Steel)
oil_prod:American Oil Production
b_com:Bloomberg commodity index (Prices USD)
us_rig: Number of US Oil Rigs in Operation
n_gas: Price of Natural Gas USD per million metric BTU
ag_price_index: Global Agricultural Price Index (USD)
us_gasoline:US stock of gasoline (in thousands of barrels)
supply_m2: Money Supply (Cash and Cash equivalents)

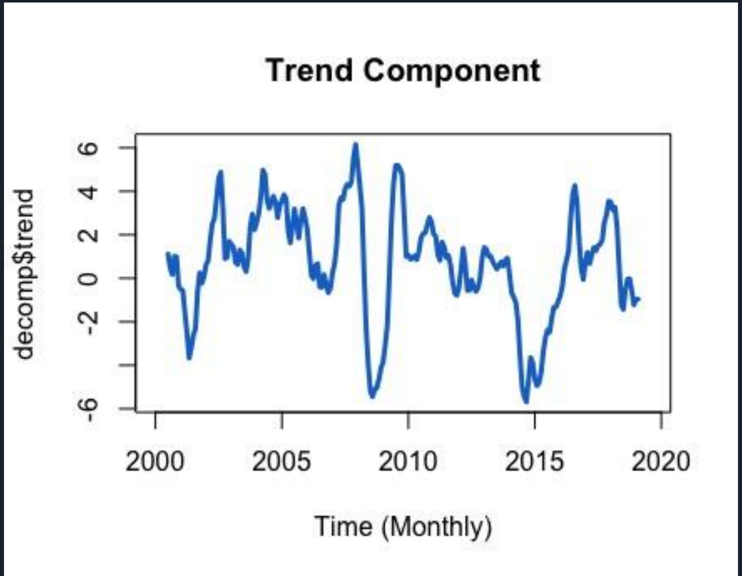
Checking For Stationarity: Visuals



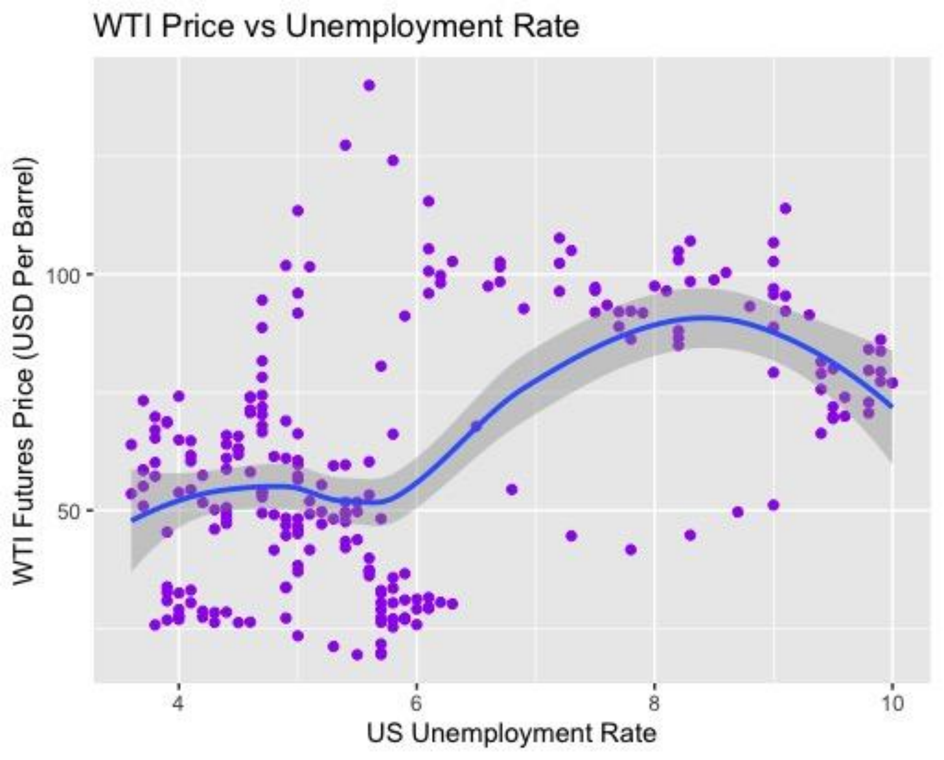


Checking For Stationarity: Augmented Dickey Fuller Test

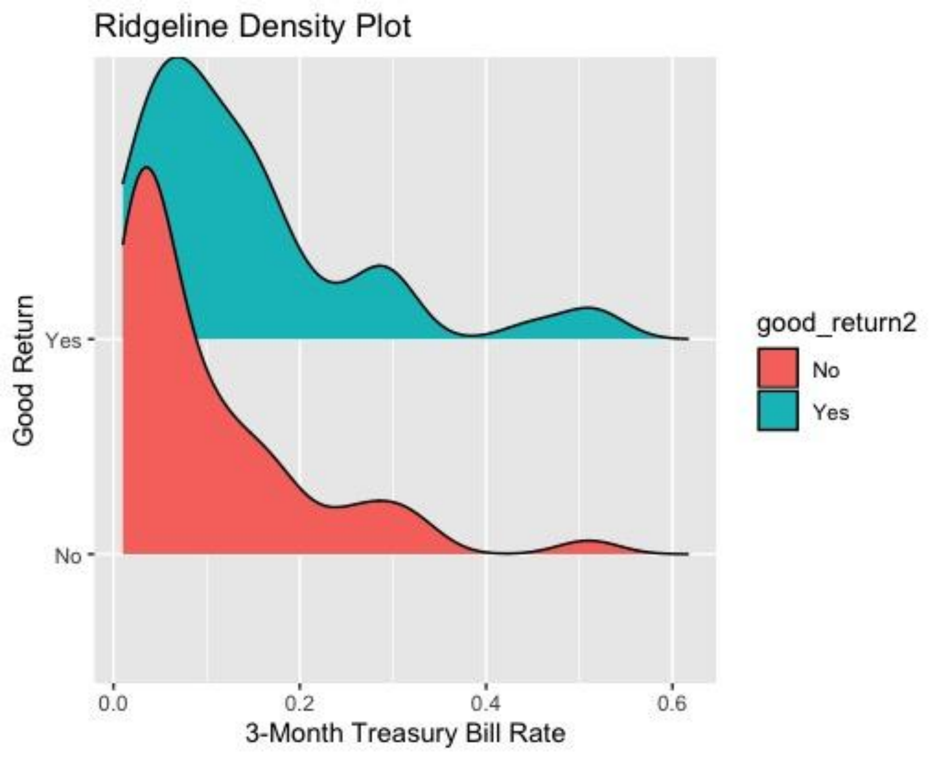
```
        Augmented Dickey-Fuller Test

data:  data.ts
Dickey-Fuller = -4.836, Lag order = 12, p-value = 0.01
```



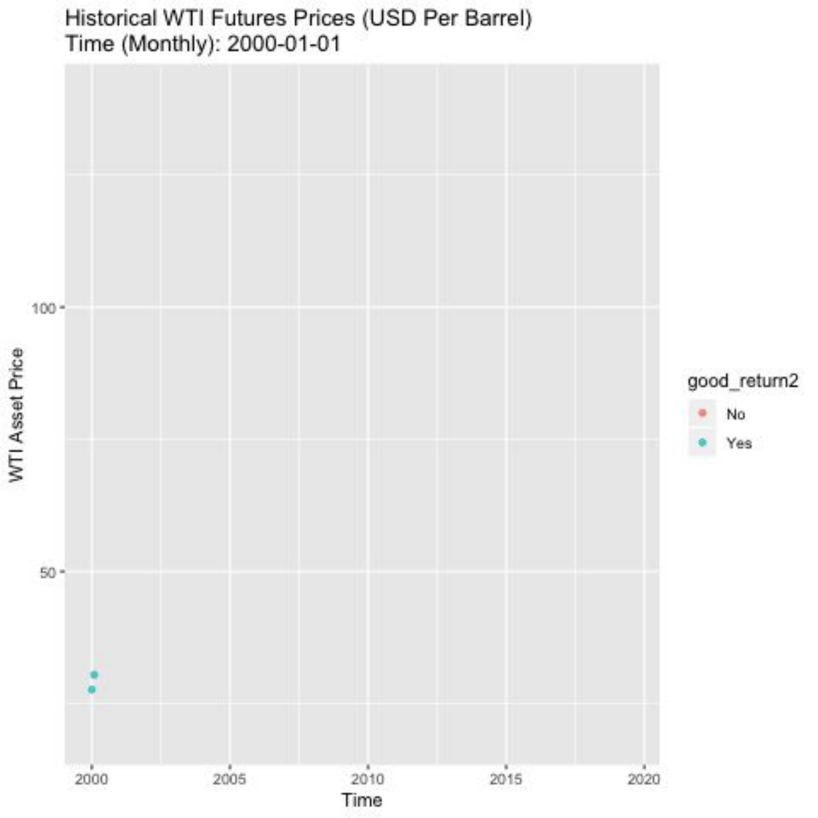
Exploratory Plots



Exploratory Plots



Exploratory Plots:

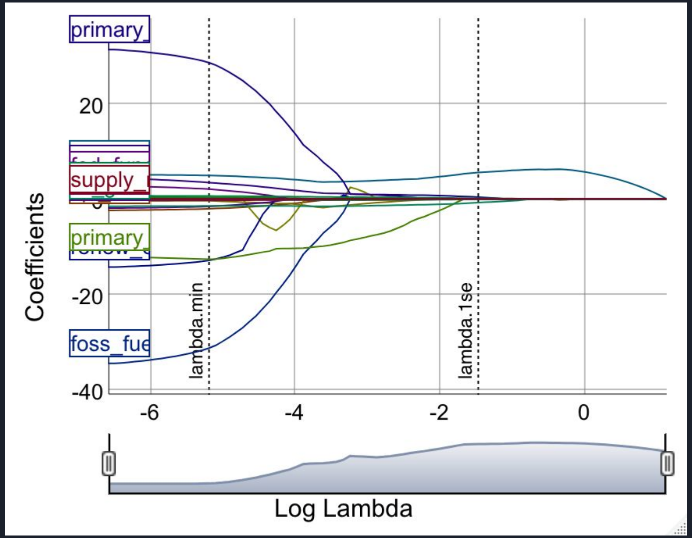


Variable Selection: Lasso Model (Shrinkage)

Lasso: Most Severity on Penalty Term (Lambda)

- Lambda Penalizes Variables/Coefficients that have high variances

(Intercept)	1.53370
date	.
ppi_oil	.
cpi_market_basket	.
inflation_rate	5.53093
import_opec	-0.00002
export_opec	.
foss_fuel_prod	.
foss_fuel_cons	.
renew_energy_prod	.
renew_energy_cons	.
primary_energy_prod	.
primary_energy_cons	0.37939
primary_energy_import_net	.
us_unemployment	.
price_gold	.
tb_rate	.
usd_index	.
fed_funds_rate	.
nasdaq_comp	0.00000
sp_500	.
steel_prod	-0.07096
oil_prod	.
b_com	0.14288
us_rig	-0.00737
n_gas	-0.77157
ag_price_index	.
us_gasoline	.
supply_m2	.

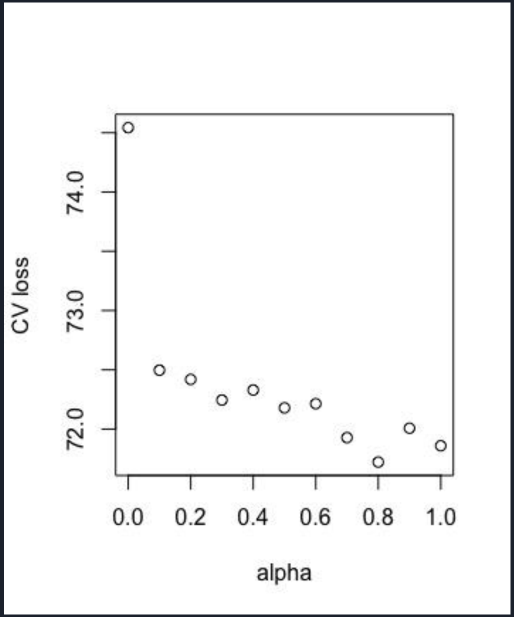


Variable Selection: Elastic Net Model (Shrinkage)

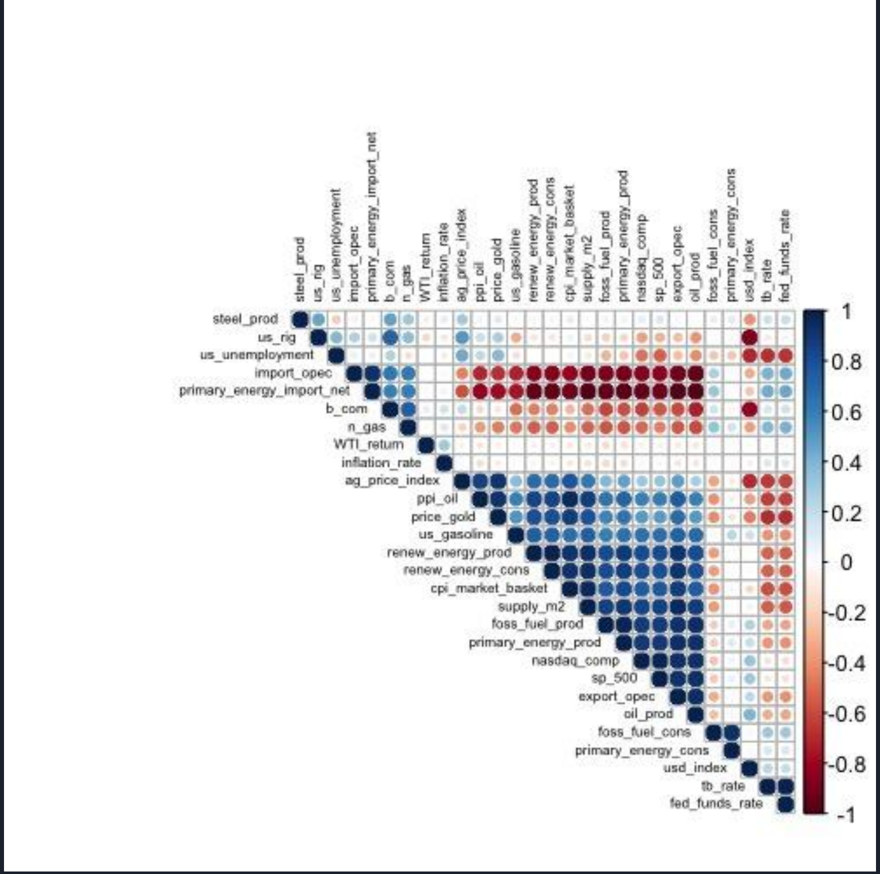
Elastic Net: Hybrid Between Ridge and Lasso Models

- Optimal Alpha Value: 0.8
  - Least Amount of Cross-Validated Loss
  - Variables Selected: 11

varname	alpha08
(Intercept)	6.4347973561
date	0.0000000000
ppi_oil	-0.0170715406
cpi_market_basket	0.0000000000
inflation_rate	5.1130520302
import_opec	-0.0000252660
export_opec	0.0000000000
foss_fuel_prod	0.0000000000
foss_fuel_cons	0.0000000000
renew_energy_prod	0.0000000000
renew_energy_cons	0.0000000000
primary_energy_prod	0.0000000000
primary_energy_cons	0.5725467091
primary_energy_import_net	-0.8433099527
us_unemployment	0.0000000000
price_gold	0.0000000000
tb_rate	0.0576247357
usd_index	0.0000000000
fed_funds_rate	0.0000000000
nasdaq_comp	0.0002503974
sp_500	0.0000000000
steel_prod	-0.1264871966
oil_prod	0.0000000000
b_com	0.1904120826
us_rig	-0.0080770793
n_gas	-1.0000482993
ag_price_index	0.0000000000
us_gasoline	0.0000000000
supply_m2	0.0000000000



Checking For Multicollinearity



Variance Inflation Factor Test (VIF):

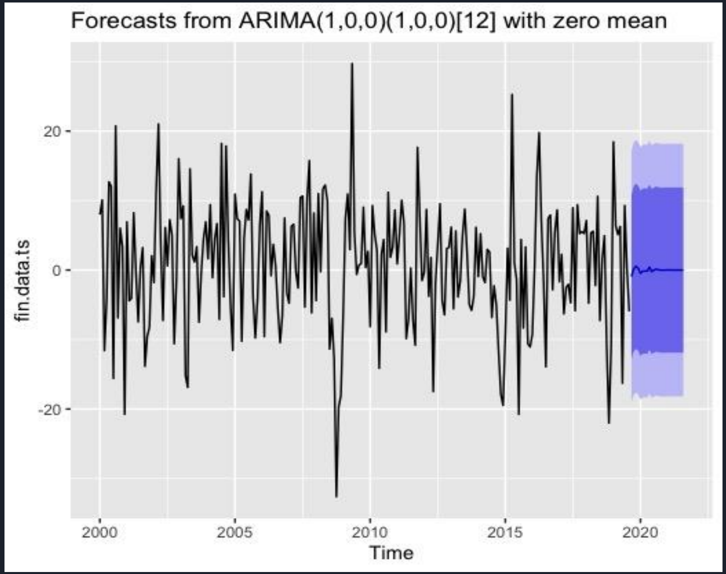
- Anything above 10 is potentially problematic
- Variables with Highest VIF scores:
  - primary_energy_prod
  - foss_fuel_prod
  - foss_fuel_cons
  - renew_energy_prod
  - renew_energy_cons
  - cpi_market_basket
  - primary_energy_cons
  - supply_m2
  - Oil_prod
  - Fed_funds_rate
- For Matrix: Used cut-off of .9 (Absolutely collinear: Extract from data set)

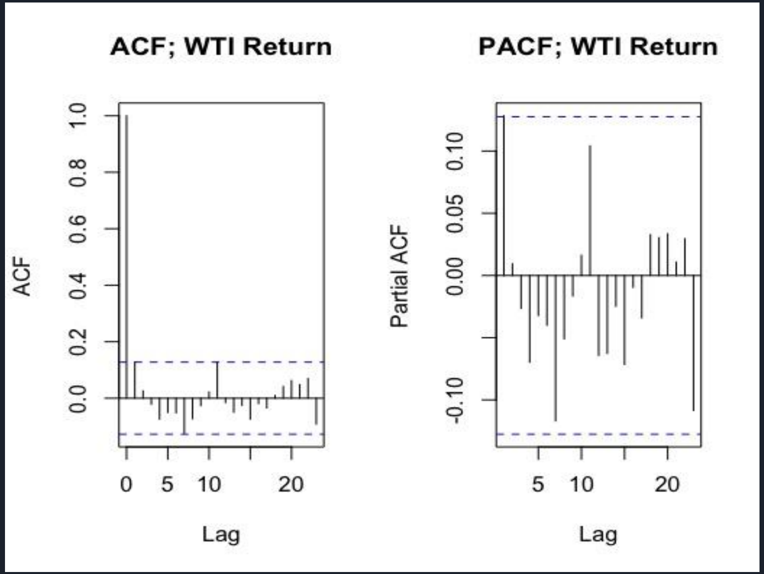
FINAL VARIABLES:

- WTI_return (TARGET)
- Date
- Inflation_rate
- Import_opec
- Primary_energy_cons
- Primary_energy_import_net
- Tb_rate
- Nasdaq_comp
- Steel_prod

- B_com
- Us_rig
- N_gas

ECONOMETRIC MODEL: ARIMA Results





ECONOMETRIC MODEL : OLS Results

- Initial OLS Model with all (potentially confounding) variables :

```
Multiple R-squared:  0.382,	Adjusted R-squared:  0.2984
```

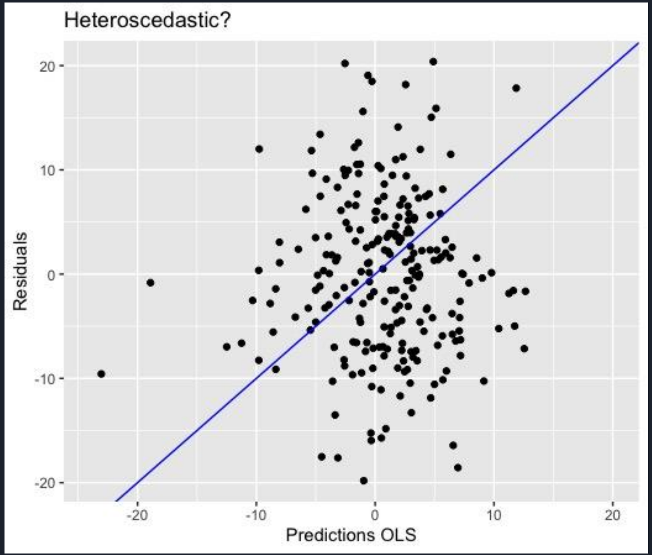
- OLS Model Post Best Feature Selection:

```
Multiple R-squared:  0.2904,	Adjusted R-squared:  0.2522
```

- After having removed 16 potential predictor variables from our initial data set our Adjusted R-Squared only went down by a mere 3.7%
- We have achieved some level of parsimony in our model: much less complex and similar predictive power

ROBUSTNESS CHECKS: OLS

- Breusch-Pagan Test




```
p-value = 0.6211
```

```
library(caret)
RMSE <- function(t,p) {
  sqrt(sum(((t - p)^2)) * (1/nrow(t)))
}
```

```
RMSE_ols <- RMSE(final_data %>% select(preds_ols), final_data %>% select(WTI_return))
RMSE_ols
```

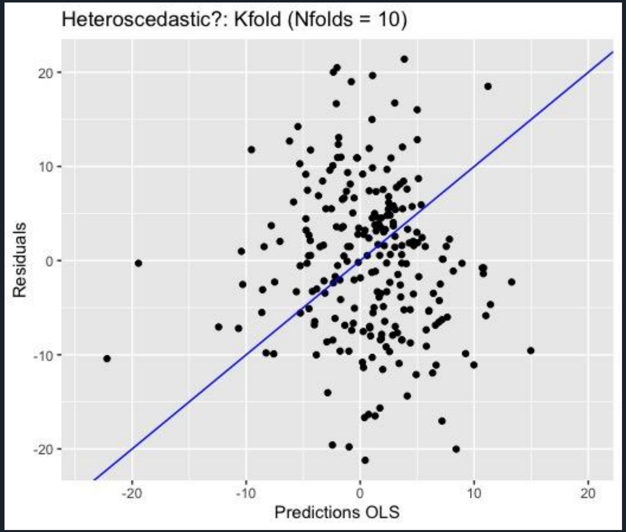
```
> RMSE_ols
[1] 7.751968
```

- Given the distribution of the WTI_return (max = 29.71) data this RMSE is relatively high and would be placed above the 3rd quartile (7.0700)
- Can't take the natural log of a negative number (WTI_return)

ROBUSTNESS CHECKS: OLS (Using K-Fold CV)

```
Multiple R-squared:  0.2842,    Adjusted R-squared:  0.2372
```

- Breusch-Pagan Test




```
p-value = 0.4102
```

```r
library(caret)
RMSE <- function(t,p) {
  sqrt(sum(((t - p)^2)) * (1/nrow(t)))
}
```

```
> RMSE_kfold
[1] 8.161091
```

- Given the distribution of the WTI_return (max = 29.71) data this RMSE is also relatively high and would be placed above the 3rd quartile (7.0700)
- Can't take the natural log of a negative number (WTI_return)

ECONOMETRIC MODEL : Logistic Regression Results

- Split good_return binary variable using Median Return Value:

Median
1.7150

```
logit_data <- logit_data %>% mutate(good_return = ifelse(WTI_return > 1.7150,1,0))
```

- Create Logistic Model Using glm and Exponentiate Coefficients:

n_gas	0.82
primary_energy_cons	1.36
tb_rate	1.14
b_com	1.08
inflation_rate	1.23

- Intuitively: exp(coeff)-1
    - High n_gas = 18% less likely to be a good_return
    - High primary_energy_cons = 36% more likely to be a good_return
    - Etc.

ECONOMETRIC MODEL : Logistic Regression Results

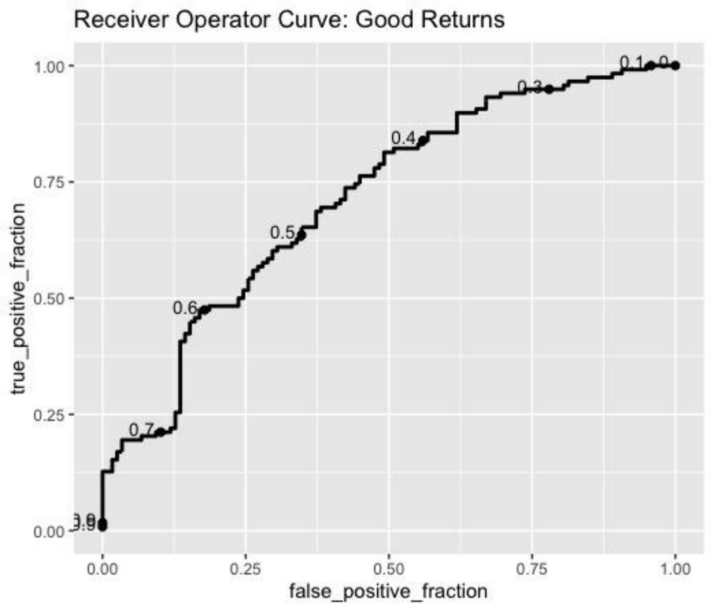
- Using Probability Cut-Off of .5 (50%)
- Confusion Matrix:
  - 0 & 1 determines if they were actually good returns
  - Predicted No and Predicted Yes is whether our model predicted a good/bad return

	0	1
Predicted No	77	43
Predicted Yes	41	75

ROBUSTNESS CHECKS: Logistic Regression

- R.O.C Plot:
  - Displays how changing the probability cut-off will affect True Positive and False Positive Rates

- Area Under Curve:
  - Displays how much better your model is than mere chance




```
calc_auc(logit_ROC)
```

sensi_TPrate	speci_TNrate	FP_rate	accuracy
0.6355932	0.6525424	0.3474576	0.6440678

AUC
0.712439

- Rule of Thumb for AUC:
  - .7-.8 - Acceptable
  - .85-.9 - Good
  - .9-1 - Excellent

Conclusion:

- After shrinkage methods were applied (Lasso, Elastic Net): Final Data set contained 11 predictor variables
- Ran: ARIMA, OLS, OLS (K-Fold), Logistic Regression
- The Logistic Regression and OLS (non-kfold) yielded the best results:
  - R2: .2522
  - AUC: .71

- If we had to apply a model in real world applications, we would use Logistic Regression
- We recommend purchasing futures during a high inflationary period (economic expansion), when the 3 month T-bill yield is high, when the price of natural gas is low, when energy consumption is high, and when the bloomberg commodity index is high as well.