THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

SPRING 2020

DATS 6450: MULTIVARIATE MODELLING

TERM PROJECT

SECTION 15

INSTRUCTOR:

DR. REZA JAFARI

ARSHIFUL ISLAM SHADMAN

GWID: G36335759

DUE: APRIL 22, 2020

1. ABSTRACT

In the context of time series data analysis several different techniques can be applied to find the best model that fits a given time series dataset. Among them a mainstream technique is multiple linear regression. However techniques such as the Holt winter method and ARMA can serve the purpose even better in some cases. In this project we will determine which model performs the best to predict time series data.

2. INTRODUCTION

The process of time series analysis involves several steps such as,

1. Understanding the dataset
2. Model Selection
3. Order Determination
4. Parameter estimation
5. Diagnostic Testing
6. Forecasting and Survival Analysis

In between the first 2 steps several factors such as stationarity, seasonality, autocorrelation and etc. There are many ways to model a time series in order to make predictions. Here, I will present,

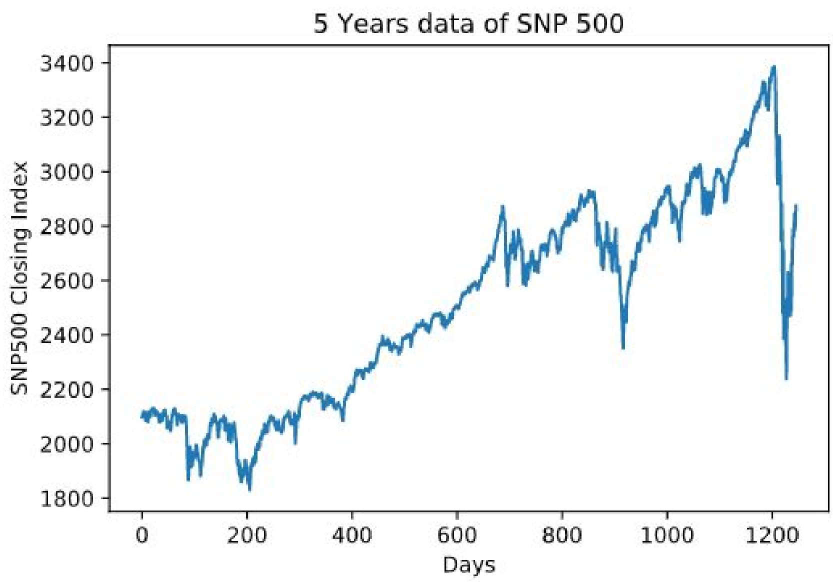
1. Holt Winter
2. Multiple Linear Regression
3. ARMA

The daily movement in the world's equity markets is influenced by a multitude of factors, ranging from large institutional block trades and program trading to earnings and economic reports. One factor that makes a splash is the influence of commodity prices. In fact, fluctuating commodity prices can have a tremendous impact on the earnings of public companies and, by extension, the markets. Hence I have gathered the prices of some commodities such as,

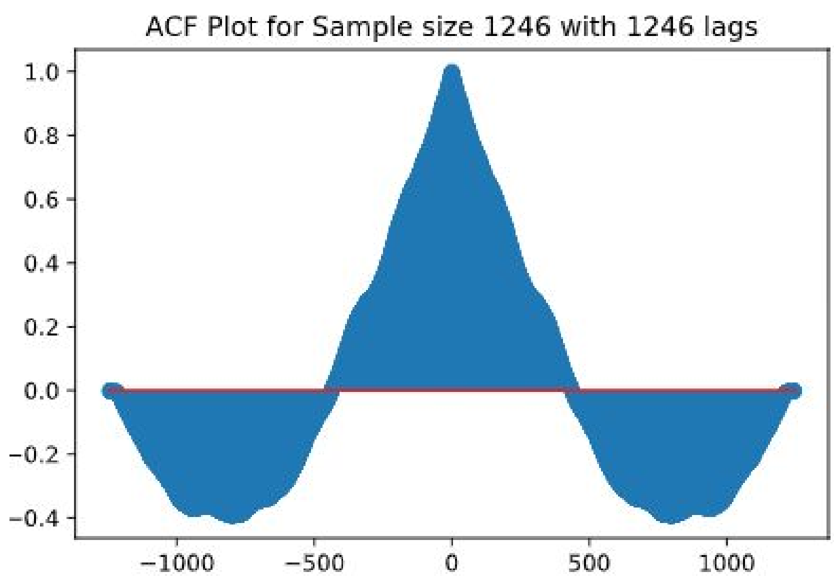
1. Lumber
2. Oil
3. Cotton
4. Wheat
5. Corn
6. Coffee and
7. Gold

Along with the stock market index of S&P 500. Price data from April 20th 2015 to April 20th 2020 was collected from https://finance.yahoo.com/ and https://www.investing.com/. The dependent variable for the project is the S&P500 market index and all the commodities mentioned above are the independent variables.

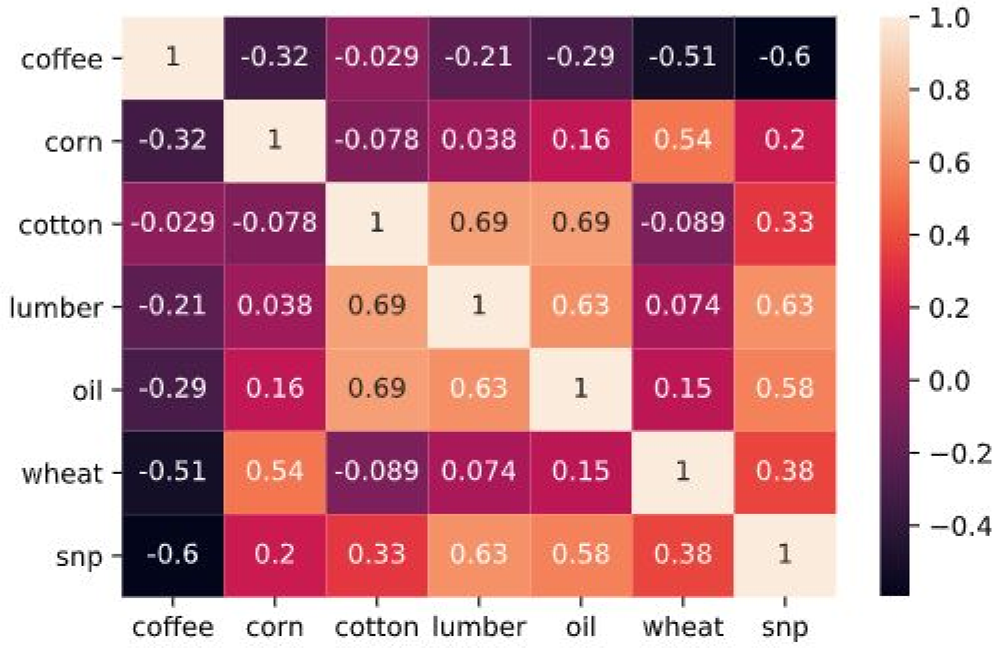
Let's look at a plot of the dependent variable against time:



And the ACF of the dependent variable:



The correlation Matrix with seaborn heatmap and Pearson's correlation coefficient:



The heat map above suggests that the dependent variable is positively correlated with lumber, oil, wheat, cotton and corn in a decreasing manner. The following pairs of independent variables show collinearity to some extent:

1. Wheat and corn
2. Oil with cotton and lumber
3. Lumber and cotton

Preprocessing of the data includes datetime transformation. I wrote the following two functions in **helper.py** file to help with the pre processing:

1. datetime_transformer(df, datetime_vars) - breaks down a date into year, month and days and returns a new dataframe with year, month and days as new columns
2. nan_checker(df) - check for any nans in a data frame and returns a dataframe with variables containing nans and the proportion of nan in the variable

```
helpme.nan_checker(all_data)

   var  proportion  dtype
```

The datetime_transformer() assisted in joining all the different variables with common dates to form a combined data frame to work with. Fortunately there was no nans in the combined data frame.

The data frame is further divided into X_df and Y_df to form the training and testing set with 80:20 split

X_df

	coffee	corn	cotton	gold	lumber	oil	wheat
0	140.30	373.00	63.03	1,202.68	242.9	56.61	500.50
1	142.40	372.50	63.09	1,187.25	252.9	56.29	499.13
2	140.45	370.75	65.42	1,194.05	252.7	57.50	498.13
3	141.15	364.50	66.31	1,179.65	258.5	57.20	486.38
4	136.10	360.75	66.14	1,202.10	256.2	56.81	473.88
...	...	...	...	...	...	...	...
1241	119.75	331.50	52.94	1,714.30	333.7	22.41	554.88
1242	117.20	326.00	52.43	1,730.45	331.0	20.11	546.75
1243	120.20	319.25	52.74	1,716.90	324.0	19.87	539.88
1244	118.60	319.75	52.90	1,717.70	332.6	19.87	528.50
1245	116.05	322.25	52.59	1,683.85	341.7	18.27	535.38

1246 rows × 7 columns

Y_df

	snp
0	2097.290039
1	2107.959961
2	2112.929932
3	2117.689941
4	2108.919922
...	...
1241	2761.629883
1242	2846.060059
1243	2783.360107
1244	2799.550049
1245	2874.560059

1246 rows × 1 columns

x_train

	coffee	corn	cotton	gold	lumber	oil	wheat
0	140.30	373.00	63.03	1,202.68	242.9	56.61	500.50
1	142.40	372.50	63.09	1,187.25	252.9	56.29	499.13
2	140.45	370.75	65.42	1,194.05	252.7	57.50	498.13
3	141.15	364.50	66.31	1,179.65	258.5	57.20	486.38
4	136.10	360.75	66.14	1,202.10	256.2	56.81	473.88
...	...	...	...	...	...	...	...
991	91.40	362.75	76.66	1,273.15	348.0	63.40	459.50
992	90.50	359.00	77.93	1,272.95	335.8	64.05	445.00
993	87.05	358.25	78.17	1,274.13	338.9	63.76	447.50
994	90.20	358.50	77.34	1,274.10	335.7	64.00	445.25
995	91.10	354.75	78.42	1,273.25	323.6	65.70	435.75

996 rows × 7 columns

x_test

	coffee	corn	cotton	gold	lumber	oil	wheat
996	91.50	351.25	77.84	1,265.65	314.9	66.30	437.75
997	90.45	346.75	77.15	1,275.38	328.5	65.89	431.62
998	92.05	347.50	78.33	1,277.35	342.8	65.21	434.00
999	92.70	351.25	77.66	1,286.25	349.5	63.30	434.88
1000	91.35	352.00	76.97	1,288.05	340.0	63.50	434.00
...	...	...	...	...	...	...	...
1241	119.75	331.50	52.94	1,714.30	333.7	22.41	554.88
1242	117.20	326.00	52.43	1,730.45	331.0	20.11	546.75
1243	120.20	319.25	52.74	1,716.90	324.0	19.87	539.88
1244	118.60	319.75	52.90	1,717.70	332.6	19.87	528.50
1245	116.05	322.25	52.59	1,683.85	341.7	18.27	535.38

250 rows × 7 columns

y_train

	snp
0	2097.290039
1	2107.959961
2	2112.929932
3	2117.689941
4	2108.919922
...	...
991	2905.580078
992	2907.060059
993	2900.449951
994	2905.030029
995	2907.969971

996 rows × 1 columns

y_test

	snp
996	2933.679932
997	2927.250000
998	2926.169922
999	2939.879883
1000	2943.030029
...	...
1241	2761.629883
1242	2846.060059
1243	2783.360107
1244	2799.550049
1245	2874.560059

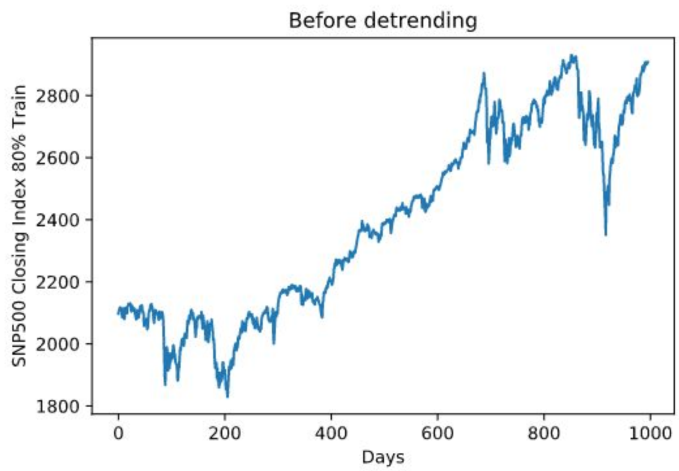
250 rows × 1 columns

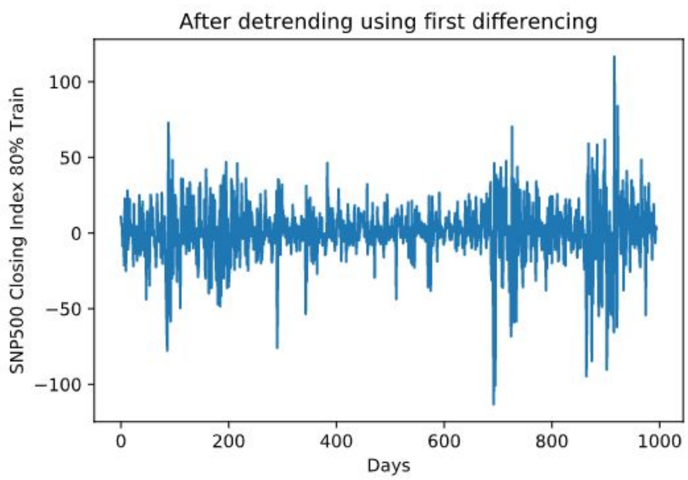
3. STATIONARITY

It is highly essential for time series data to be stationary for building an ARMA model. The dependent variable was found to be non-stationary with p-value found from the ADF test to be greater than 0.05.

```
ADF for dependent variable:
ADF Statistic: -1.232266
p-value: 0.659564
Critical Values:
        1%: -3.436
        5%: -2.864
        10%: -2.568
```

Therefore we fail to reject the null hypothesis (H0), and the data does have a unit root and is definitely not stationary. Detrending the dependent variable will make it stationary. The plot of the variable before detrending and after detrending using first differencing method is given below:





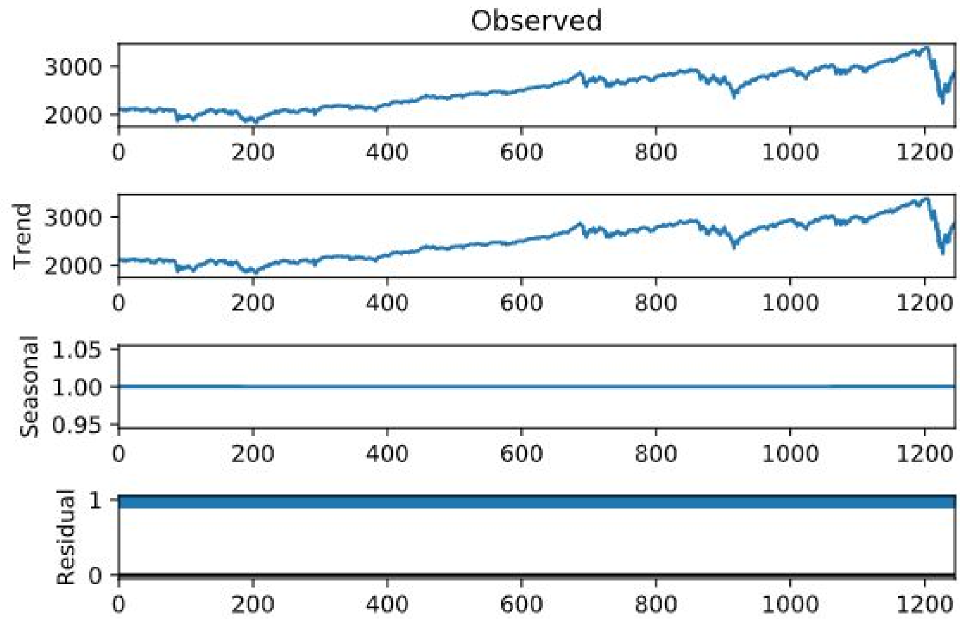
Performing the ADF test over the first differenced values of the dependent variable:

```
ADF for differenced traiing dependent variable:
ADF Statistic: -12.176675
p-value: 0.000000
Critical Values:
        1%: -3.437
        5%: -2.864
        10%: -2.568
```

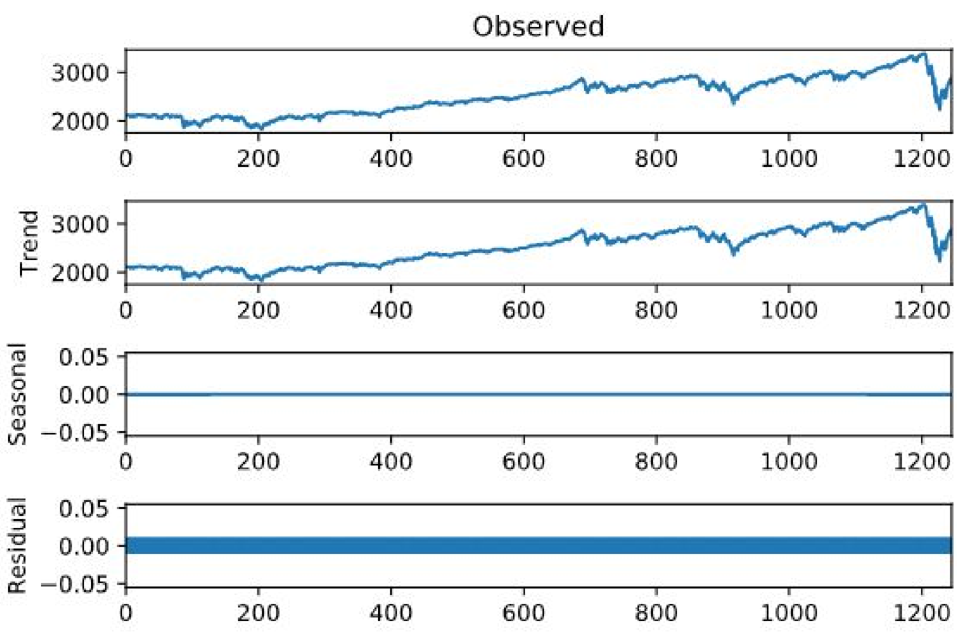
The p-value is less than 0.05, Therefore we can reject the null hypothesis (H0), and we can observe that first differencing has made the data stationary.

4. TIME SERIES DECOMPOSITION

Multiplicative Decomposition



Additive Decomposition



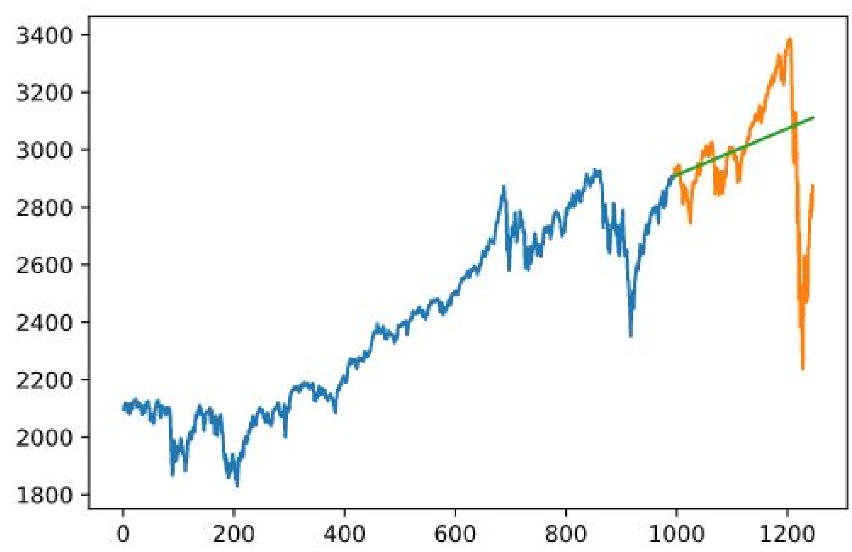
The amplitude of the seasonal component in the multiplicative decomposition is constant and does not change. The best decomposition here is additive decomposition.

5. HOLT WINTERS METHOD

The implementation of the Holt Winters Exponential Smoothing method is displayed below:

```
#hotwinter model: <start>
days=range(1,len(all_data)+1)
holt_Y={'day': days,
            'snp': all_data['snp']}
holt_df=pd.DataFrame(data=holt_Y)
holt_df.set_index('day',inplace=True)
holt_df.index.freq='M'
holt_train, holt_test = holt_df.iloc[:996,0], holt_df.iloc[995:,0]
holt_model = ExponentialSmoothing(holt_train, trend='add', damped=False,
seasonal='add', seasonal_periods=2).fit()
holt_predictions = holt_model.predict(start=holt_test.index[0],
end=holt_test.index[-1])
plt.plot(holt_train.index, holt_train, label="train")
plt.plot(holt_test.index, holt_test, label="test")
plt.plot(holt_predictions.index, holt_predictions, label="predictions")
plt.show()
#hotwinter model: <end>
```

The green line in the plot below shows the prediction and the orange line is the actual value. The method could not capture seasonality in the data as it consists of none.



6. ACCURACY OF HOLT WINTERS METHOD

The Holt **Root Mean Square of Forecast Error** is 211.02252647826106

The Holts **Mean of Forecast Error** is -23.206037709241826

The **standard error** using holt is 211.8683115374463

The **R^2** using holt is 0.015194450359197887

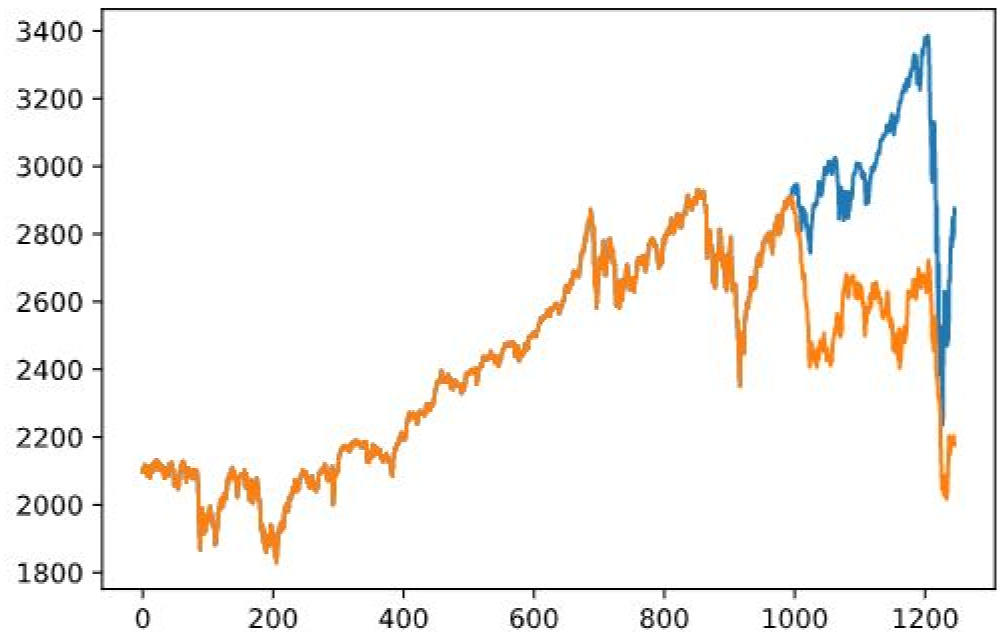
The **Adj R^2** using holt is 0.01123940799116252

7. MULTIPLE LINEAR REGRESSION

The price of gold was in string type while the others were all float type, therefore the combined data was exported to form the **all_data.csv** file. Using excel the data type for gold was converted to numeric and saved back to form the **all_data_numeric.csv** file. This file was then imported to perform the building of the following multiple linear regression model:

```
OLS Regression Results
==============================================================================
Dep. Variable:                    snp   R-squared:                       0.874
Model:                            OLS   Adj. R-squared:                  0.873
Method:                 Least Squares   F-statistic:                     980.7
Date:                Wed, 22 Apr 2020   Prob (F-statistic):               0.00
Time:                        16:26:35   Log-Likelihood:                -6080.4
No. Observations:                 996   AIC:                         1.218e+04
Df Residuals:                     988   BIC:                         1.222e+04
Df Model:                           7
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       1718.7891    129.387     13.284      0.000    1464.884    1972.694
coffee        -7.3646      0.286    -25.755      0.000      -7.926      -6.803
corn          -1.5057      0.234     -6.426      0.000      -1.966      -1.046
cotton        13.4272      0.840     15.982      0.000      11.778      15.076
gold           0.6347      0.066      9.673      0.000       0.506       0.763
lumber         0.3235      0.074      4.343      0.000       0.177       0.470
oil            9.4012      0.589     15.960      0.000       8.245      10.557
wheat         -0.4842      0.142     -3.416      0.001      -0.762      -0.206
==============================================================================
```

The plot for actual values versus predicted values is given below, where the blue line for the plot indicates the actual 20% test data and the orange line along with it is the prediction values from the OLS Model:



8. ACCURACY OF MULTIPLE LINEAR REGRESSION

The Root Mean Square of Forecast Error using OLS is 485.79238819659855
The Mean of Forecast Error using OLS is 452.08047618280233
The standard error using OLS is 487.74729504151753

The p-value for the constant and the predictor variables are all 0.000. The null hypothesis and the alternate hypothesis suggests that:

H 0 : The coefficient is zero and has no effect on the model.

H a : The coefficient is not zero and has an effect on the model.

Since the p values are less than 0.05 we can reject the null hypothesis and say that all the predictor variables have meaningful addition to the model because changes in the predictor value are related to the changes in the response variable. Therefore there is no need for eliminating any feature.

9. ARMA

a. GPAC

As the GPAC result table from the above implementation will not be visible if attached in this section, please refer to the **gpac.pdf** file submitted along with the report. The two estimated orders for the ARMA parameter estimation are:

i. ARMA (15,16) and
ii. ARMA (16,23)

b. Parameter Estimation

The estimated parameters for the following ARMA Process are:

1. **ARMA (15,16):**

The AR coeff a0 is: -0.5966705635134957
The AR coeff a1 is: -0.2163960923095659
The AR coeff a2 is: -0.6805130515182428
The AR coeff a3 is: -0.6901177746513014
The AR coeff a4 is: 0.08403302096484255
The AR coeff a5 is: -0.1604230405642471
The AR coeff a6 is: -0.4958666182040948
The AR coeff a7 is: -0.235843340834847
The AR coeff a8 is: 0.06824785887979806
The AR coeff a9 is: -0.4388543315460225
The AR coeff a10 is: -0.8678596020133884
The AR coeff a11 is: -0.26768642575053786
The AR coeff a12 is: -0.4919379816335416
The AR coeff a13 is: -0.7811728903697143
The AR coeff a14 is: -0.38614067327570767
The MA coeff b0 is: 0.5909690429075963
The MA coeff b1 is: 0.1561093304558162
The MA coeff b2 is: 0.6947755938392512
The MA coeff b3 is: 0.68275560351528
The MA coeff b4 is: -0.14240728761303798
The MA coeff b5 is: 0.14873108542190883
The MA coeff b6 is: 0.5549062449033725
The MA coeff b7 is: 0.13864945223098016
The MA coeff b8 is: -0.1285348862660888
The MA coeff b9 is: 0.4523949570774292
The MA coeff b10 is: 0.8937197525066068
The MA coeff b11 is: 0.23359727777416667
The MA coeff b12 is: 0.5313188303164804
The MA coeff b13 is: 0.7708717293218298

The MA coeff b14 is: 0.2871556695060188
The MA coeff b15 is: -0.0272701551020267

ARMA Model Results

==

Dep. Variable:	y	No. Observations:	995
Model:	ARMA(15, 16)	Log Likelihood	-4380.637
Method:	css-mle	S.D. of innovations	19.618
Date:	Wed, 22 Apr 2020	AIC	8825.273
Time:	03:35:58	BIC	8982.161
Sample:	0	HQIC	8884.916

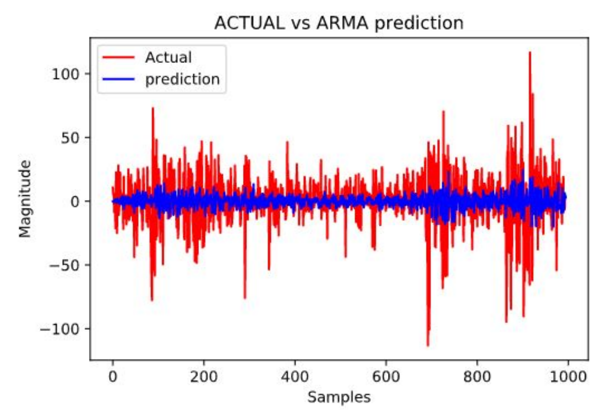
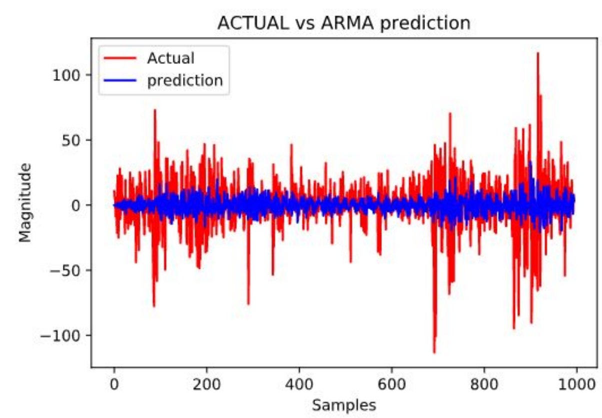
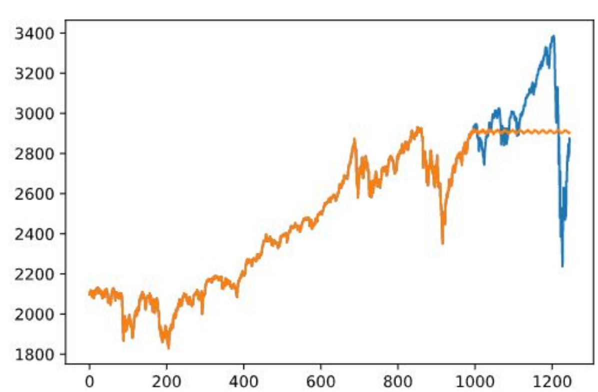
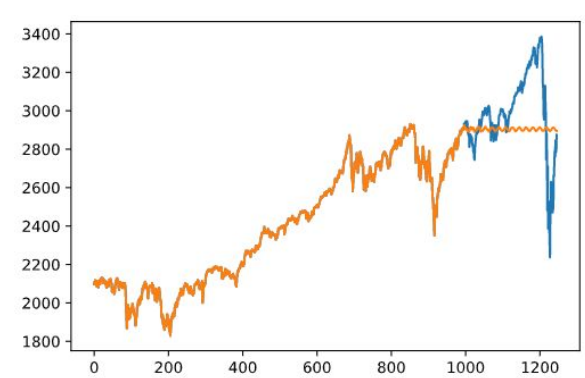
2. **ARMA (16,23):**

The AR coeff a0 is: -0.9000143894355711
The AR coeff a1 is: -0.674022908307425
The AR coeff a2 is: -0.8327394781772389
The AR coeff a3 is: -1.1551563360730701
The AR coeff a4 is: -0.9637304337894667
The AR coeff a5 is: -0.2308027882683602
The AR coeff a6 is: -0.4073584590846961
The AR coeff a7 is: -1.0446590540933827
The AR coeff a8 is: -0.46695174888518415
The AR coeff a9 is: -0.3221501571823652
The AR coeff a10 is: -1.0211771241027883
The AR coeff a11 is: -1.1349788052506382
The AR coeff a12 is: -0.8714824486279609
The AR coeff a13 is: -0.6881426645133217
The AR coeff a14 is: -0.7945877677317983
The AR coeff a15 is: -0.7708628210390632
The MA coeff b0 is: 0.898237124924133
The MA coeff b1 is: 0.6184391424370909
The MA coeff b2 is: 0.8224703941509651
The MA coeff b3 is: 1.1239185311687212
The MA coeff b4 is: 0.90454549464499
The MA coeff b5 is: 0.1631398986801479
The MA coeff b6 is: 0.4061185163291549
The MA coeff b7 is: 1.0070493117502126
The MA coeff b8 is: 0.37568459340266
The MA coeff b9 is: 0.24736892085732912
The MA coeff b10 is: 1.1090515226338518
The MA coeff b11 is: 1.141305757303169
The MA coeff b12 is: 0.8189329772561263
The MA coeff b13 is: 0.6508320528551754
The MA coeff b14 is: 0.7576518503236143
The MA coeff b15 is: 0.7148662643024575
The MA coeff b16 is: -0.0759523717621692
The MA coeff b17 is: -0.03216224094500768
The MA coeff b18 is: -0.011735793608416355
The MA coeff b19 is: -0.03703132247902313
The MA coeff b20 is: 0.07148460561549057
The MA coeff b21 is: 0.023088143540584492
The MA coeff b22 is: 0.011232630001094101

ARMA Model Results

==

Dep. Variable:	y	No. Observations:	995
Model:	ARMA(16, 23)	Log Likelihood	-4366.636
Method:	css-mle	S.D. of innovations	19.282
Date:	Wed, 22 Apr 2020	AIC	8813.272
Time:	06:27:26	BIC	9009.381
Sample:	0	HQIC	8887.825

ARMA(15,16)	ARMA(16,23)
	
Root Mean Square of Forecast Error is 223.31308551239488 Mean of Forecast Error is 78.20938526714954 The standard error is 3545.098777609819	Root Mean Square of Forecast Error is 224.98044440749692 Mean of Forecast Error is 82.71875920064676 The standard error is 3571.5681265389258
	

10. CONCLUSION

From Holt Winters:

The Holt Root Mean Square of Forecast Error is 211.02252647826106
The Holts Mean of Forecast Error is -23.206037709241826
The standard error using holt is 211.8683115374463

From OLS:

The Root Mean Square of Forecast Error using OLS is 485.79238819659855
The Mean of Forecast Error using OLS is 452.08047618280233
The standard error using OLS is 487.74729504151753

From ARMA(15,16):

Root Mean Square of Forecast Error is 223.31308551239488
Mean of Forecast Error is 78.20938526714954
The standard error is 3545.098777609819

From ARMA(16,23):

Root Mean Square of Forecast Error is 224.98044440749692
Mean of Forecast Error is 82.71875920064676
The standard error is 3571.5681265389258

Looking at the RMSE, mean forecast error and the standard errors we can conclude that Holt Winters Method is the best model for the data collected for S&P500 market index and the commodities. However if we are to consider only ARMA then ARMA(15,16) performed slightly better than ARMA(16,23). Multiple linear regression performed the worst amongst the three. The data did not show any seasonality. Moreover the nature of the stock market is highly volatile and several factors other than just commodities affect the stock market. The huge drop in the data at the 20% testing set indicates the stock market crash due to the COVID-19 pandemic resulting in high forecast errors.