Modelling Suncor Energy Share Price by a Multiple Linear Regression

December 22, 2017

Section 1: Purpose

This report inspired by Capital Asset Pricing Model (CAPM), attempting to model stock returns based on regression. In CAPM, we are determining how our stock is related to the overall market, using a regression structure. In this report, we extend this model by adding commodity trading prices for oil, gold, gas and currency exchange rate as explanatory variables.

This report consists of four sections as mentioned below,

1.Constructing dataset using Quantmod package, providing summary and simple graphic analysis on the dataset.

2.Obtaing a model for estimating stock performance on US market, Canada to US currency exchange rate, prices for oil, gas and gold. Together with some diagnostic checks including lattice plots, variance inflation factor bar chart, visualization of correlation matrix, Durbin-Watson.

3.Modyfing the model by introducing the logarithmic returns for both response and the predictors. Also checking the assumption of linear regression for proposed model.

4.Checking for Model validation; deciding whether the numerical results obtained from regression analysis are acceptable as description of the data. This section includes test and train validation according to RMSE, and time Series diagnostic checks for residuals.

5.Conclusion and recommendation.

Introduction to Suncor Dataset

For this project we load raw data from yahoo finance and Fred source via Quantmod package in R. Working with raw data could be very difficult. The most challenging part is to convert a dataset with missing data into a clean dataset. In Suncor Dataset, there are some days that the price is not available. Also, choosing acceptable risk-free security or other explanatory variables need researching, and spending time to find the best choices. The dataset consists of five years (2012 to 2017) of records of the daily close price of Suncor energy INC., S&P500 from yahoo finance website, and daily future prices for crude oil, Natural gas, gold, Canada to US exchange rate From the Federal Reserve Bank of St. Louis Economic Data (FRED) website. In second model we consider 3-month Treasury constant maturity rate as the risk-free rate. which can be collected from FRED website.

Variable description

r.SU: Logarithmic daily return of stock price for Suncor Energy Inc. which is a Canadian integrated energy company based in Calgary, Alberta. It specializes in production of synthetic crude from oil sands.

r.SP500:Logarithmic daily return of index price for Standard & Poor's 500, often abbreviated as the S&P 500, which is an American stock market index based on the market capitalizations of 500 large companies having common stock listed on the NYSE.

r.GOLD: Logarithmic daily return of gold fixing price 10:30 A.M. (London time) in London Bullion Market.

r.OIL: Logarithmic daily return of crude Oil Price: West Texas Intermediate (WTI) - Cushing, Oklahoma, based in US dollars per barrel.

r.GAS: Logarithmic daily return of Henry Hub Natural Gas Spot Price, based in US dollars per million BTU.

r.CAUS: Logarithmic daily return of Canadian Dollars to One U.S. Dollar exchange rate.

Summary of data output and graphs

Some libraries are needed for the rest of the project:

```
library(zoo)
library(graphics)
library(ggfortify)
library(knitr)
library(car)
library(RColorBrewer)
library(quantmod)
library(Metrics)
library(psych)
```

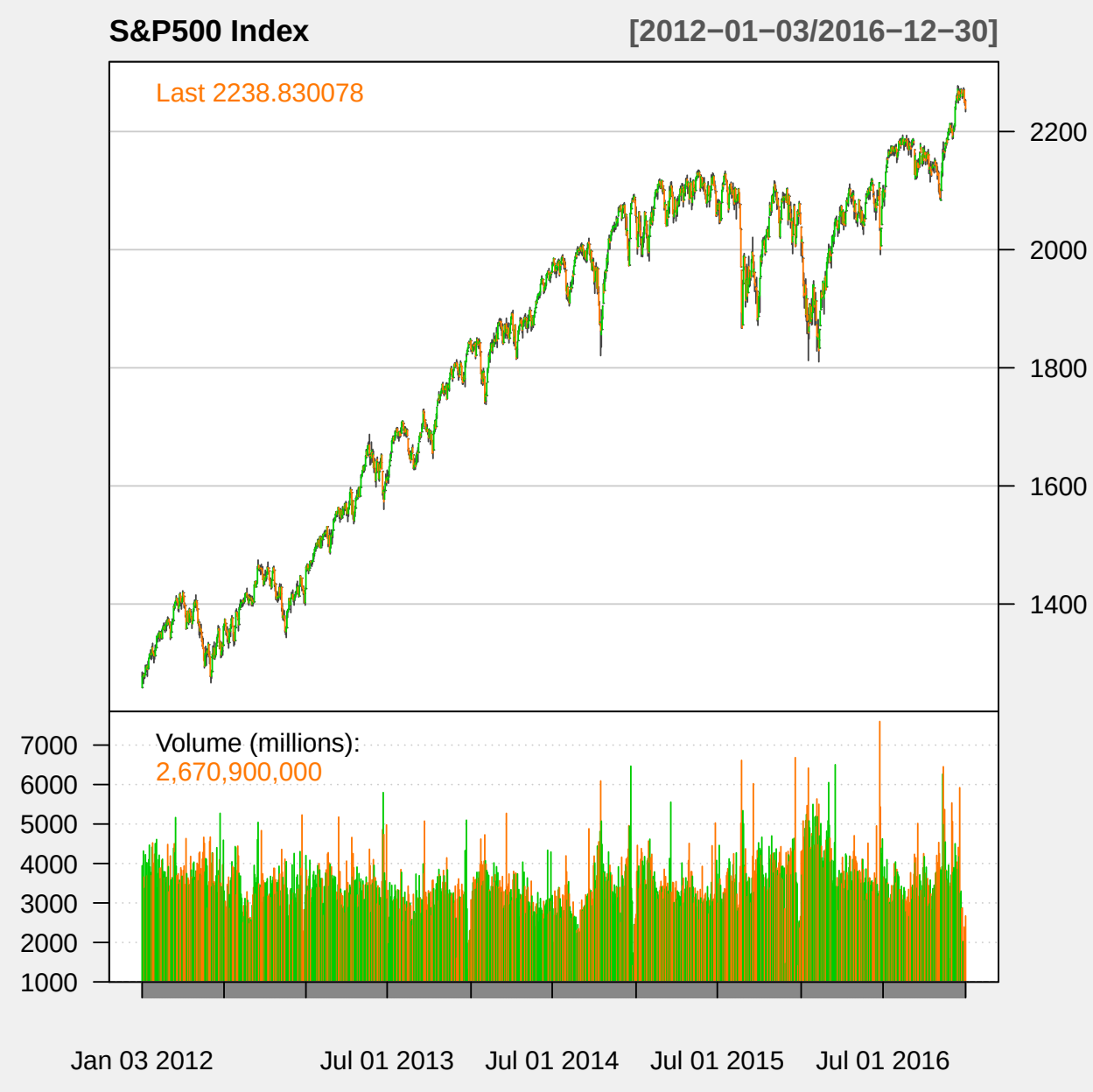

Figure 1: Daily close price of S&P500 from yahoo finance website

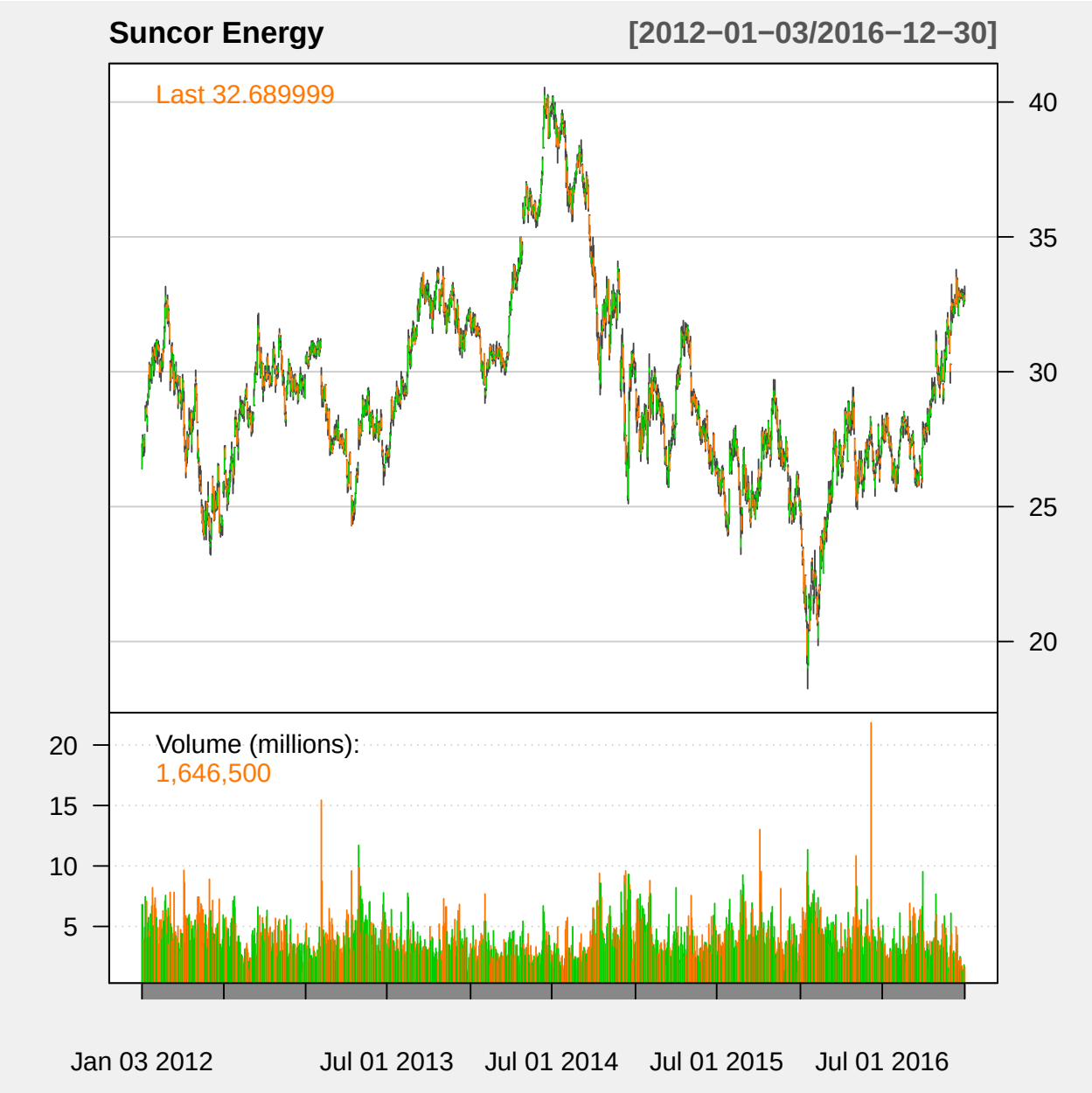

Figure 2: Daily close price of Suncor energy INC. from yahoo finance website

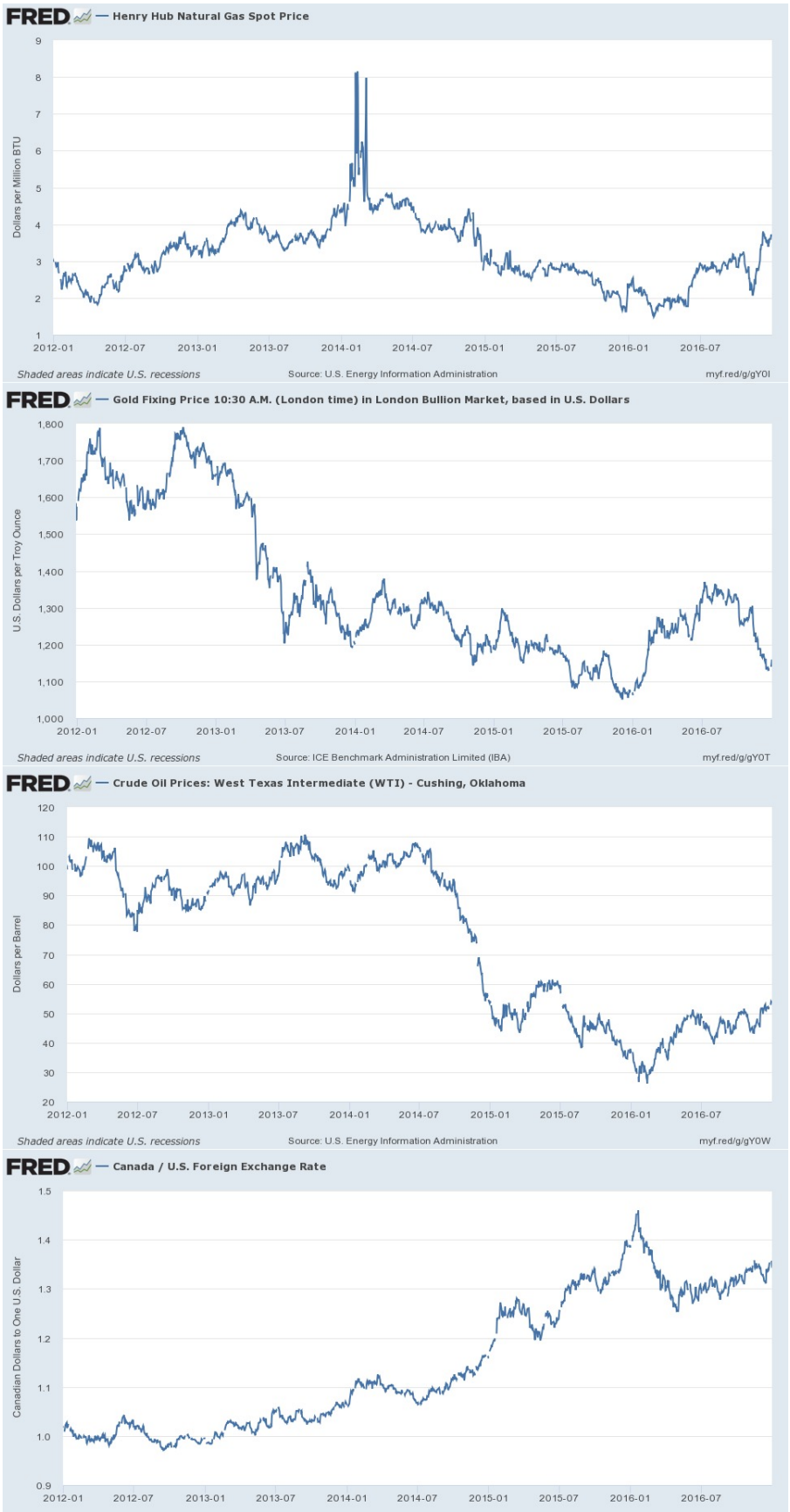

Figure 3: Daily prices of crude oil, Natural gas, gold and Canada to US exchange rate from 2012 to 2017 extracted from FRED website

All variables are numerical and the dataset doesn't have any categorical variable.There is no missing value as well.

Table 1: Dataframe Summary

Statistic	N	Mean	St. Dev.	Min	Max
SU	1,258	29.359	3.550	19.128	40.243
SP500	1,258	1,822.753	283.582	1,277.060	2,271.720
GOLD	1,258	1,349.769	197.101	1,050.600	1,790.000
CAUS	1,258	1.148	0.135	0.971	1.459
OIL	1,258	75.416	25.653	26.190	110.620
GAS	1,258	3.201	0.897	1.490	8.150

Section 2: Why use the logarithm of returns, rather than price or raw returns?

The purpose of this section is to demonstrate the idea behind logarithmic transformation of returns on our dataset. First, I look at scatter plot matrix to roughly determine if there is a linear correlation between my variables. Second, I run a regression of daily Suncor stock price on other variables, and look at the diagnostic checks to see how well is the plots. finally, I use Durbin-Watson test to detect the presence of autocorrelation.

Correlation matrix

1. The correlation matrix suggests that CAUS, OIL and GAS have significant correlations with the response variable SU.
2. It seems that, there is a collinearity between gas and oil variable.
3. The relationship between Suncor stock price and other variables are not linear.

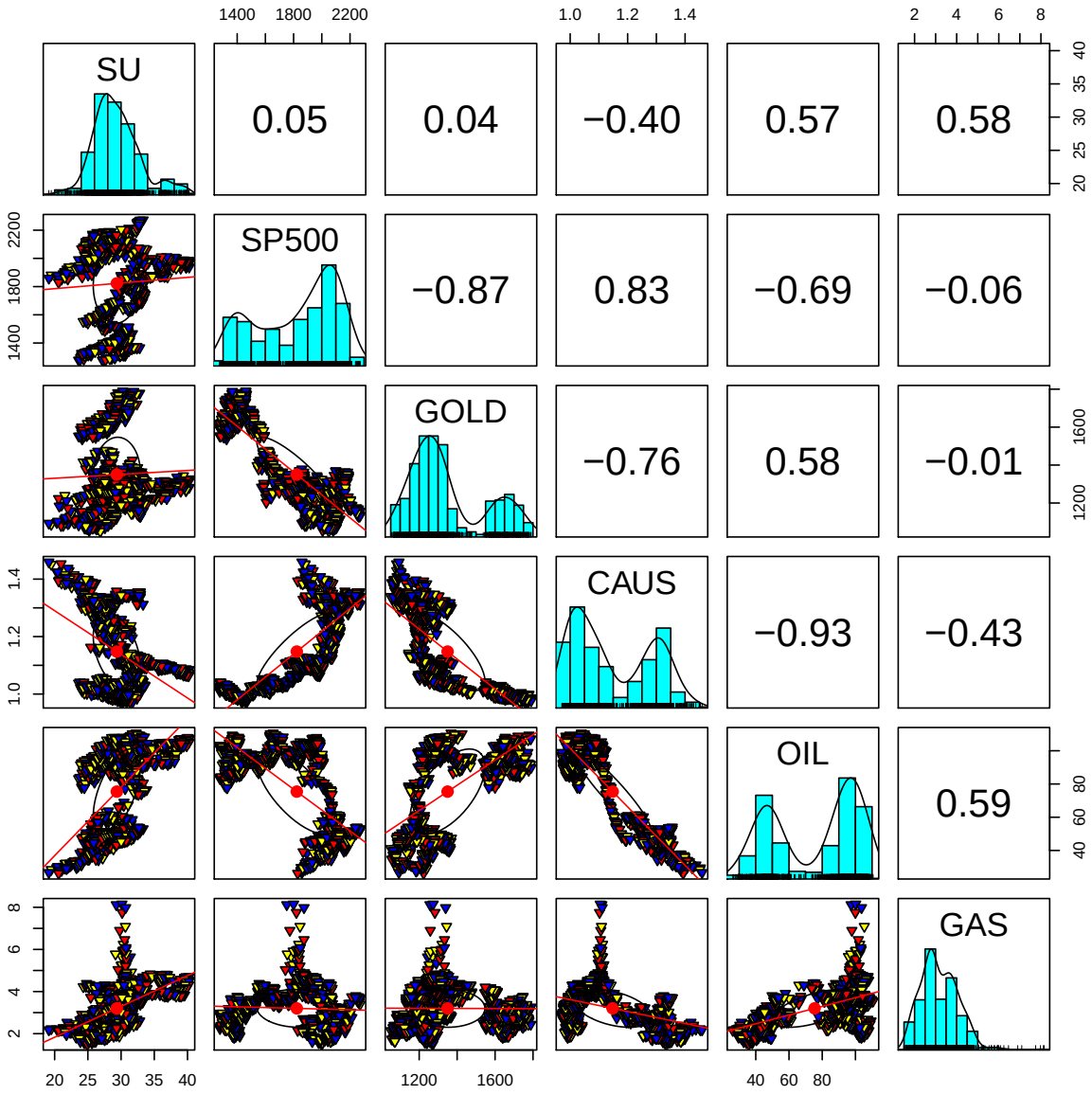

Figure 4: Correlation Matrix

MODEL 1:

$$SU = \beta_0 + \beta_1(SP500) + \beta_2(OIL) + \beta_3(GOLD) + \beta_4(CA/US) + \beta_5(GAS) + error.$$

Where:

SU: Daily stock price for Suncor Energy Inc. which is a Canadian integrated energy company based in Calgary, Alberta. It specializes in production of synthetic crude from oil sands, based US dollars.

SP500: daily index price for Standard & Poor's 500, often abbreviated as the S&P 500, which is an American stock market index based on the market capitalizations of 500 large companies having common stock listed on the NYSE, based US dollars.

GOLD: daily gold fixing price 10:30 A.M. (London time) in London Bullion Market, based US dollars.

OIL: daily crude Oil Price: West Texas Intermediate (WTI) - Cushing, Oklahoma, based in US dollars per barrel, base US dollars.

GAS: daily of Henry Hub Natural Gas Spot Price, based in US dollars per million BTU, based US dollars.

CAUS: daily Canadian Dollars to One U.S. Dollar exchange rate.

Summary of model 1 and diagnostic plots:

Table 2: Model Summary

	Dependent variable:
	SU
SP500	0.017^{***}
	(0.0004)
GOLD	0.006^{***}
	(0.001)
CAUS	-5.357^{***}
	(1.508)
OIL	0.162^{***}
	(0.007)
GAS	-0.484^{***}
	(0.087)
Constant	-14.166^{***}
	(2.578)
Observations	1,258
R^2	0.758
Adjusted R^2	0.757
Residual Std. Error	1.750 (df = 1252)
F Statistic	784.039^{***} (df = 5; 1252)
Note:	*p<0.1; **p<0.05; ***p<0.01

- All explanatory variables are significant at 0.01% level.

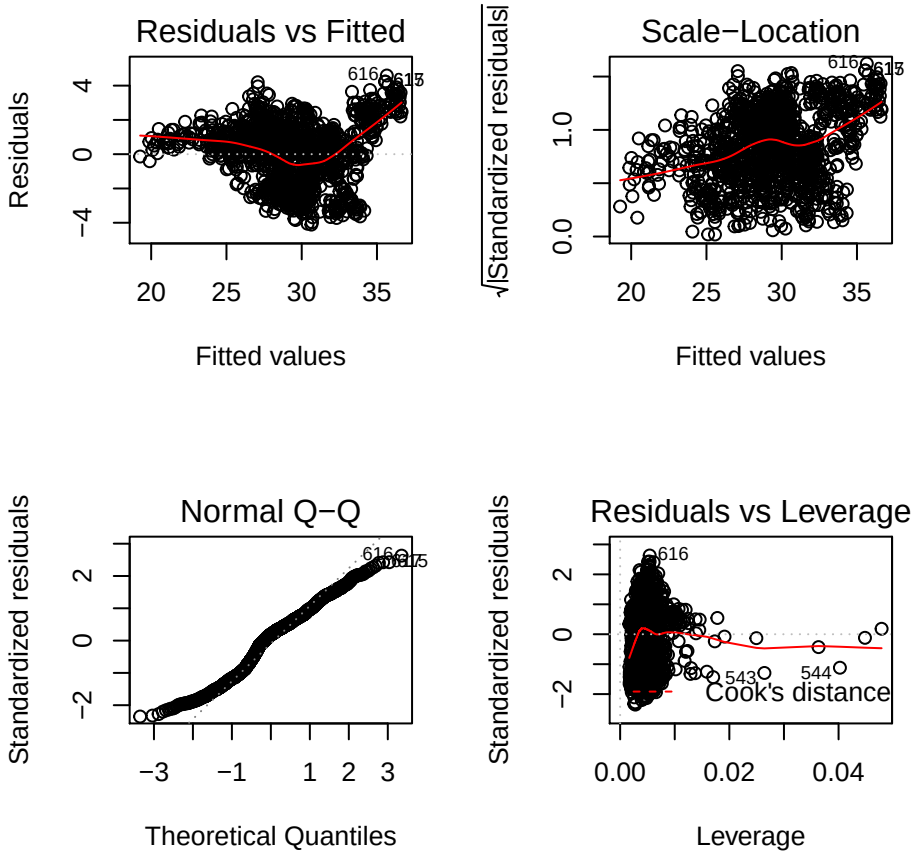

Figure 5: Basic Regression Diagnostic Checks

- Although the R^2 value is quite high (75.79%), the fitted line plots of SU vs explanatory variables suggests that the relationship between Suncor Energy stock price and other variables is not linear.
- The residual versus fitted plot also suggests that the relationship is not linear. Because the lack of linearity dominates the scale location plot, we can not use the plot to evaluate whether the error variances are equal.
- We must fix the non-linearity problem before we assess the assumption of equal variances.
- The Q-Q plot suggests that the error terms are not normal; There is sufficient evidence to conclude that the error terms are not normally distributed.

VIF test

The variance inflation factors (VIF)for CAUS and OIL are high. The higher the value of VIF, the higher the collinearity.

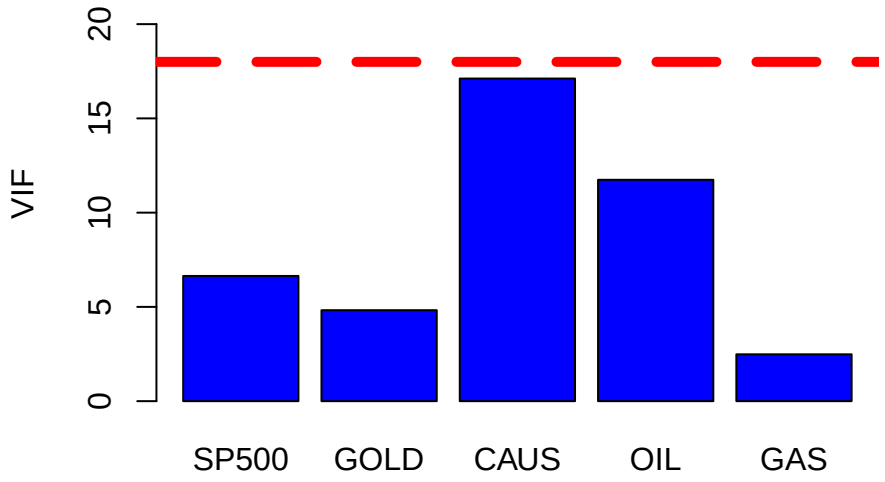

Figure 6: VIF for input variables.

Durbin-Watson Test

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.9739011    0.05068815       0
##  Alternative hypothesis: rho != 0
```

1.The null hypothesis is that, there is no autocorrelation. We can safely reject the null at 1%. There is evidence for autocorrelation.

2.Because the dataset is time series, and the most common issue with this type of data is autocorrelation, as we expected, independence of residuals has violated in this model. One of the common method to solve this problem, is transformation.

In finance, it is very common to work with the logarithmic return or continuously compounded return, also known as force of interest, and a standard approach is to apply a natural log transformation to the return of stock prices before fitting a regression model. One of the justification for this method is due to ease of interpretation for coefficients. For example, if a stock is priced at 3.570 USD per share at the close on one day, and at 3.575 USD per share at the close the next day, then the logarithmic return is: $\ln(3.575/3.570) = 0.0014$, or 0.14%.

Also, logarithmic returns are widely preferred over raw prices or returns in quantitative analysis of financial time series for various other reasons such as normalization (returns of different assets can be compared, their prices usually not), time-additivity, and other conveniences for classical statistics and mathematics.

Section 2

Separating data into training and testing sets is an important part of evaluating data mining models. Typically, when we separate a data set into a training set and testing set, most of the data is used for training, and a smaller portion of the data is used for testing. Analysis Services randomly samples the data to help ensure that the testing and training sets are similar. By using similar data for training and testing, we can minimize the effects of data discrepancies and better understand the characteristics of the model. For this purpose, I allocate randomly 75% of my data to training set, and 25% to test set. I do all my analysis on my training set.

MODEL 2:

Model 2 which is inspired by CAPM (Capital asset pricing model):

$$(R_{su} - r_{rf}) = \beta_0 + \beta_1(R_M - r_{rf}) + \beta_2(R_{oil}) + \beta_3(R_{gold}) + \beta_4(R_{ca/us}) + \beta_5(R_{gas}) + error.$$

where:

r.SU.0=r.SU-r_{rf}

r.SP500.0= r.sp500-r_{rf}

r_{rf} is the rate of return for a risk-free security. Other variables are the ones that we defined in Data section.

Correlation matrix:

- As we can see, the excess rate of return of market over risk-free rate(r.SP500.0), log return of oil price(r.OIL) and log return of Canada to US dollar Exchange rate(r.CAUS) have significant correlations with excess rate of return of Suncor energy stock price over risk-free rate(r.SU.0)
- All correlations are below 0.55, which shows that, there is no evidence of collinearity in new model.
- Correlation between r.SU.0 and r.CAUS is negative, and is positive with other variables.

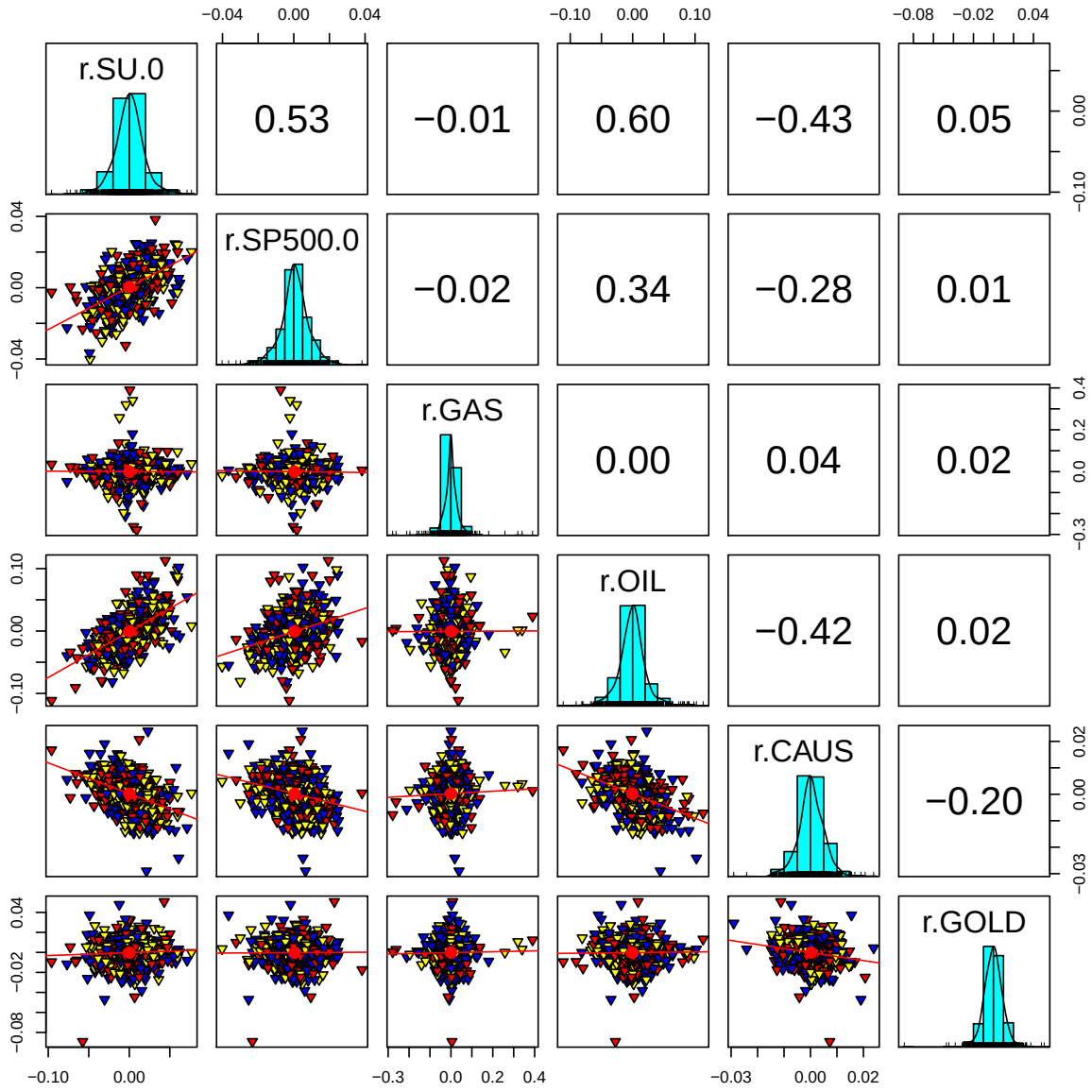

Figure 7: Correlation Matrix

Summary of model 2:

Table 3: Model Summary

	Dependent variable:
	r.SU.0
r.SP500.0	0.751^{***}
	(0.055)
r.OIL	0.354^{***}
	(0.021)
r.CAUS	-0.578^{***}
	(0.095)
r.GOLD	0.036
	(0.041)
r.GAS	0.004
	(0.010)
Constant	0.0002
	(0.0004)
Observations	943
R^2	0.517
Adjusted R^2	0.514
Residual Std. Error	0.013 (df = 937)
F Statistic	200.397^{***} (df = 5; 937)
Note:	*p<0.1; **p<0.05; ***p<0.01

The regression is definitely significant at less than 0.1%.

1.$R^2 = 0.5162$, The model explains approximately 52% of the variability in the logarithmic daily return of Suncor Energy Stock price.

2.Slope estimate for log daily return of S&P500, log Daily return of Crude Oil and log daily return of exchange rate (Canada to US dollar) is significant at less than 0.1%.

3.The intercept is almost 0, and the p-value for intercept is high. I don't have enough evidence to reject the null hypothesis that intercept is equal to 0 .IT makes sense in finance. I can interpret that as "the investment has earned a return adequate for the risk taken".

Stagewise

According to this procedure, the best model , based on BIC criteria ,is the one that includes the variables log daily return of S&P500, log daily return of oil and log daily return of exchange rate (Canada to US dollar).

Summary of final model (BIC model):

Table 4: Model Summary

	Dependent variable:
	r.SU.0
r.SP500.0	0.751^{***}
	(0.054)
r.OIL	0.352^{***}
	(0.021)
r.CAUS	-0.595^{***}
	(0.093)
Constant	0.0002
	(0.0004)
Observations	943
R^2	0.516
Adjusted R^2	0.515
Residual Std. Error	0.013 (df = 939)
F Statistic	334.087^{***} (df = 3; 939)
Note:	*p<0.1; **p<0.05; ***p<0.01

Diagnostic plots:

1.From the normal Q-Q plot we see that the assumption of residuals being normally distributed does not seem to be violated.

2.The Residuals vs fitted, and Scale-Location plots agree that the residuals are independently and identically distributed. There is no fan shape in the residual plot, however there are some sign of violation of homoscedasticity that we can check it by NCV test.

3.The Residuals vs Leverage plot shows that there is no concern.

This test has a p-value more that a significance level of 0.05, therefore we can not reject the null hypothesis that the variance of the residuals is constant.

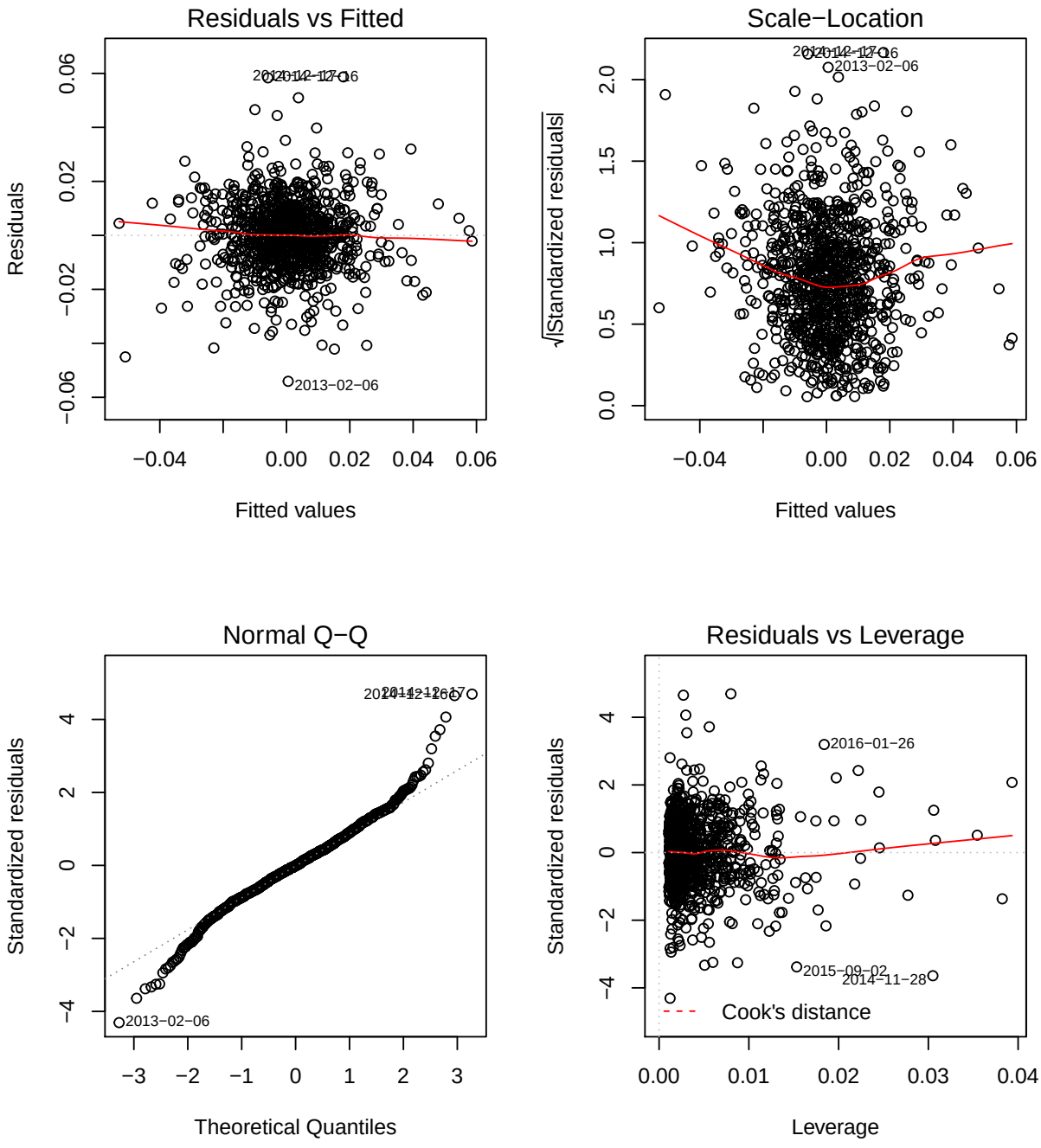

Figure 8: Basic Regression Diagnostic Checks

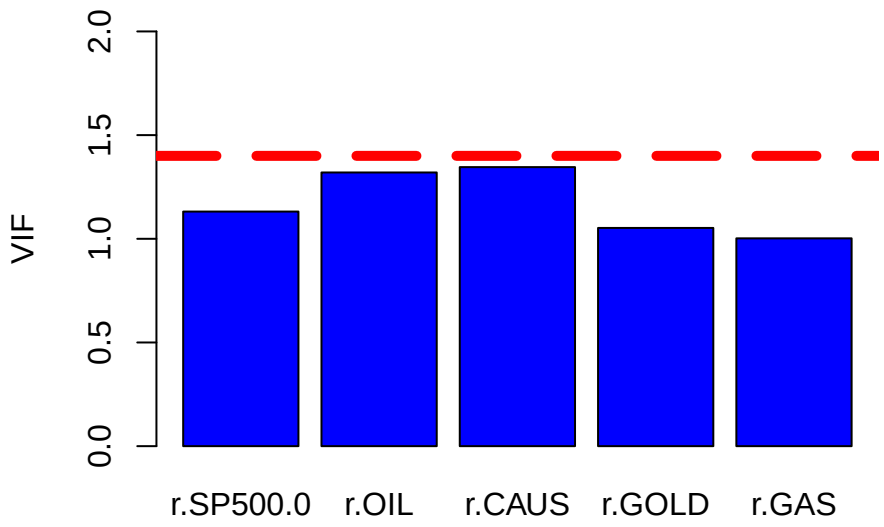

Figure 9: VIF for input variables

VIF test

Variance Inflation Factor(Figure 9) – In general, the VIFs of the linear regression indicate the degree that the variances in the regression estimates are increased due to multicollinearity. VIFs values are very small, indicate that multicollinearity is not a problem in this fit.

As we can see, the transformation helped me to overcome the problem with multicollinearity.

Durbin-Watson Test

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.9739011    0.05068815       0
##  Alternative hypothesis: rho != 0
```

The null hypothesis is that there is no autocorrelation. According to p-value, we don't have enough evidence to reject null hypothesis. there is no evidence of autocorrelation.

Chapter 4: Model validation

Final model:

$$(R_{su} - r_{rf}) = \beta_0 + \beta_1(R_M - r_{rf}) + \beta_2(R_{oil}) + \beta_3(R_{ca/us}) + error.$$

Comparing root mean square error for testing and training set:

```
 [,1]           [,2]
```

[1,] "RMSE.train" "0.0125607341169698" [2,] "RMSE.test" "0.0133473607179899"

1.I Predicted the test dataset and computed the RMSE on the test dataset and compared it with the training dataset.

2.According to the result from summary for final model and RMSE for test dataset, result seems good. RMSE for testing and training dataset is not very different.

Time Series Diagnostic Checks for Residuals in OLS Model

- In the time series plot, there is no obvious systematic departures from randomness such as trend, seasonality, clustering.
- The SACF plot is used to detect if there is strong autocorrelation present. According to plot, residuals are independent.
- There is no evidence of heteroscedasticity in this dataset.

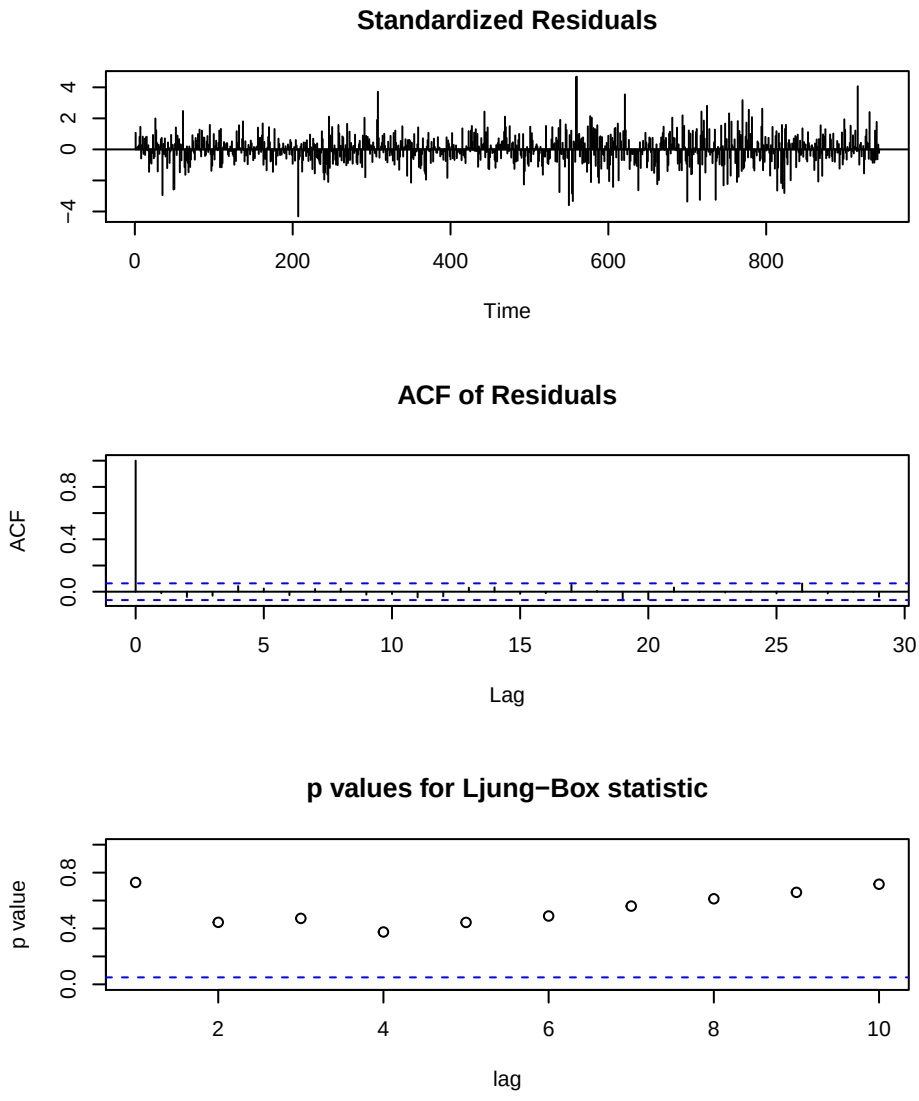

Figure 10: Time series plot

Section5: conclusion and recommendation

This report contributes to the determination of the character of excess of daily stock return of Suncor Energy over risk free rate (r.SU.0) based on excess of daily market return over risk free rate(r.SP500.0) and commodities including daily return of oil(r.OIL), gas(r.GAS), gold(r.GOLD), Canada to US currency exchange rate(r.CAUS). Our final analysis shows significant relationship between r.SU.0 values and r.SP500.0, r.OIL, r.CAUS. In last two chapters we checked for any violation of linear assumption, and could not find any obvious evidence, however In 2003 Granger and his collaborator Robert Engle were jointly awarded the Nobel Memorial Prize in Economic Sciences for demonstrating that, the stock market prices exhibit conditional variance changes depending on the past value. Therefore a Garch(1,1) could be a better fit for this type of data.