Time Series Forecast for Gas Price

Zihe Yang (zy151)
Xueyan Liu (xl491)
Jiamin Zhong(jz644)

1. INTRODUCTION

Recently gas prices have attracted media attention as the national average for a gallon of gasoline has swung from nearly $4 a gallon nationwide, to just under $1 in the wake of the COVID-19 pandemic. Before the pandemic, geo-political tensions, hurricane seasons, flooding or even the increased travel demand during the summer, all of these will affect gas price. As we observed, fluctuations in oil prices always affect people's lives. If discretionary spending is hampered by higher gasoline costs, it can even have knock-on effects throughout the broader economy.

At the individual level, higher gas prices mean that each of us pay more at the pump, leaving less to spend on other goods and services. Higher prices also mean that shoppers will tend to drive less - including to places like to malls or shopping centers. Indeed, academic and industry studies provide support for this, showing that driving miles are directly tied to gas prices. While shoppers may not drive, they do switch to online shopping more when gas prices rise. According to Marin Software, searches for online shopping do increase dramatically along with an increase in gas prices. However, all retailers are further squeezed as they are forced to pass on the higher expenses that they themselves experience, which are associated with increased shipping costs to consumers. Anything that has to be shipped or transported - from apples to electronics - could cost more as gas prices rise. This is especially true for products, or components for products, that are manufactured overseas. Likewise, many products that contain plastics or synthetic materials are based in part on petroleum and refining. Higher oil prices means higher prices for these materials too.

Higher commuting costs also affect the workplace, some businesses in colleges have implemented 4-day weeks to limit the financial burdens of commuting. Behind the price of gas can be an issue for job seekers causing them to turn down jobs that would require a significant drive. Rising gas prices can cause some companies to backoff for hiring plans. Out of concern about the economy's health, less discretionary spending also creates a lower sales and a need for fewer workers.

All in all, gas price can be a critical factor in the economy - affecting everything from consumer spending to the price of airline tickets to hiring practices. To understand its trend and to predict its behavior can help households with budget planning, even can become an investment reference for people who are interested in energy commodities.

2. DATA

The dataset used in this research is released by the National Agency of Petroleum, Natural Gas and Biofuels (ANP in Portuguese). It contains weekly reports of Regular Gasoline, Diesel, Hydrous Ethanol, Natural Gas and LPG prices across the country from year 2004 to year 2019. These datasets bring the mean value per liter, min value per liter, max value per liter, number of gas stations analyzed and other information grouped by regions and states across the country.

Each gas has different functions in our daily life. Natural gas can power the heating systems, stove tops, water heaters and dryers in our home. LPG is a flammable mixture of hydrocarbon gases used as fuel in heating appliances, cooking equipment, and vehicles. Gasoline and diesel are two widely used fuels in transportation. Hydrous Ethanol is also used for vehicle refueling. With this dataset, we are able to explore each gas's behavior along the time.

There are totally 21 attributes in this dataset. The detailed description of each attribute is attached in Appendix.

3. METHODS & RESULTS

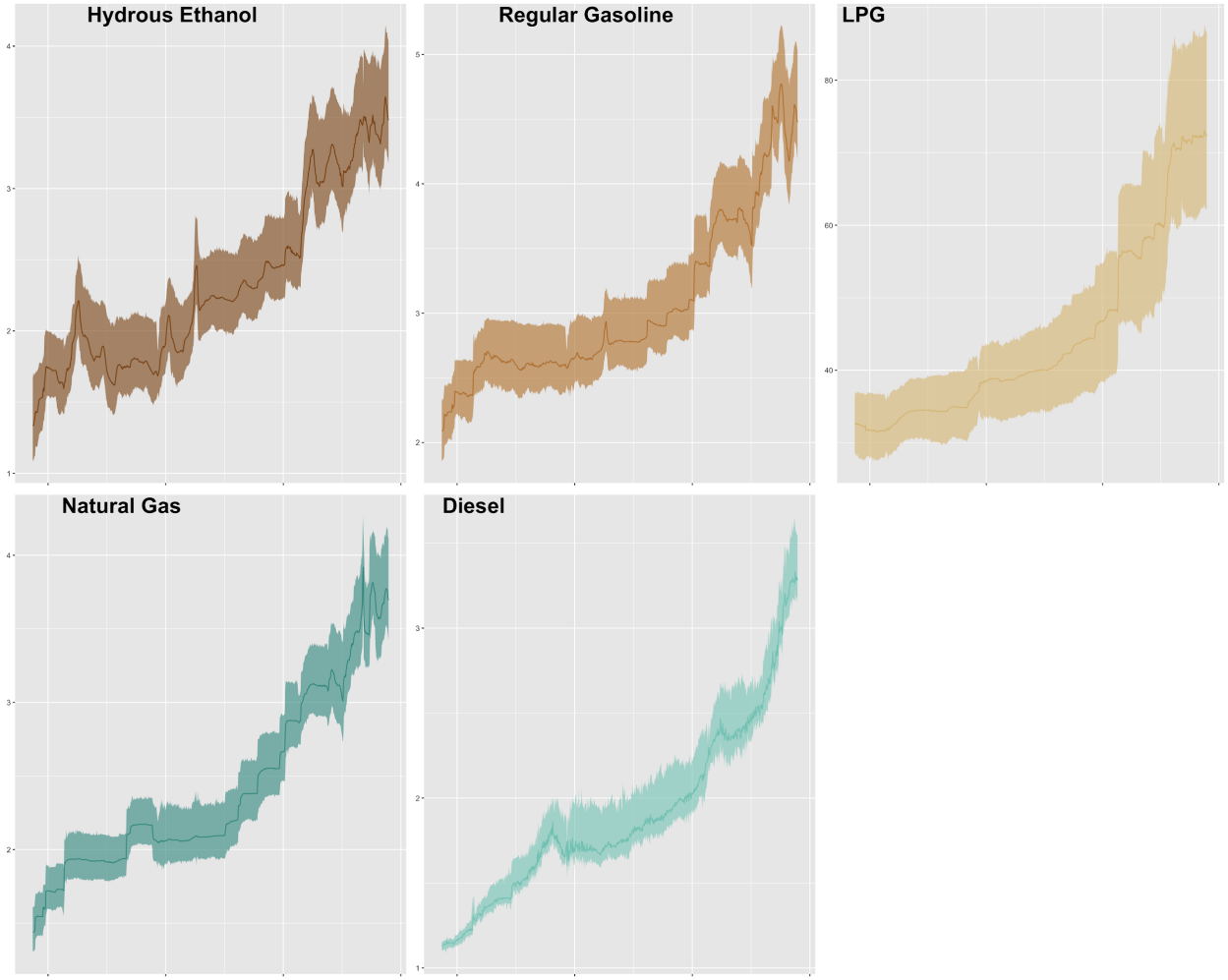
3.1 Exploratory Data Analysis

In order to better understand data, we applied Exploratory data analysis, which refers to the critical process of performing initial investigations on data to discover the pattern of data.[1]Our project's goal is to predict the gas prices of 5 different products. First of all, we grouped data by the gas products, computing the weekly average price in Brazil. Then visualized it to see how each product's price fluctuates over time.

3.2.1 Time series plot

The first figure shows each gas product weekly time series plot with the boundaries of price fluctuation in that week from 2004 to 2019. We can see that Natural Gas and Diesel have relatively narrow boundaries, indicating that these two products have smaller weekly price changes. And LPG's price has more dramatic weekly fluctuations since it has widest boundaries.

[1] Patil, P. (2018, May 23). What is Exploratory Data Analysis? Retrieved from https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15



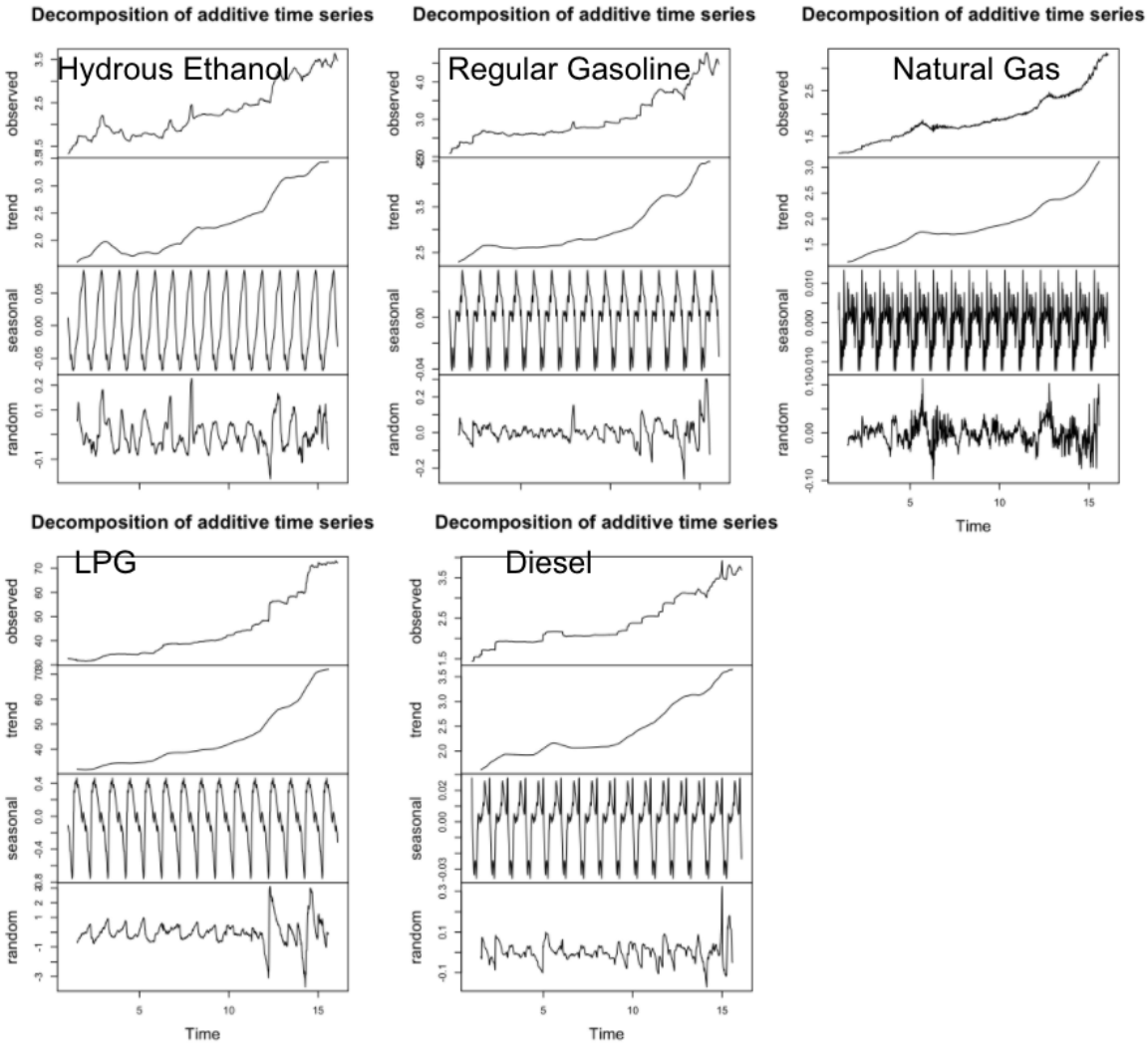
The second plot(the figure below) displayed the price change for each product per year. Hydrous Ethanol's plot is the most chaotic, which means its price fluctuations are greater than the other four products'. The price of Natural Gas's price fluctuations is the most stable one. It is clear that the price is increasing year by year, except for 2007-2010. Instead of rising slowly, the price shows relatively large fluctuations, especially for Hydrous Ethanol. At the same time, this plot also shows that in each year, the price is relatively stable (most lines tend to be straight). And in the same year, different products have similar patterns. For example in 2015, all products have the same shape curve (spikes appear at both ends, flat middle part.)

3.2.2 Time series decomposition

Time series decomposition offers an abstract of time series, it considers each time series curve as a combination of level, trend, seasonality, and noise components.

The plot below shows the decomposition of each product's gas price from 2004 to 2019. All the products have a clear and stable upward trend. For LPG, regular gasoline and Diesel, the variances of residuals increase from left to right which may be a sign of heteroskedasticity. And the residual variances of Natural Gas and Hydrous Ethanol are volatile, we may need to conduct the volatility clustering in the model for further analysis.

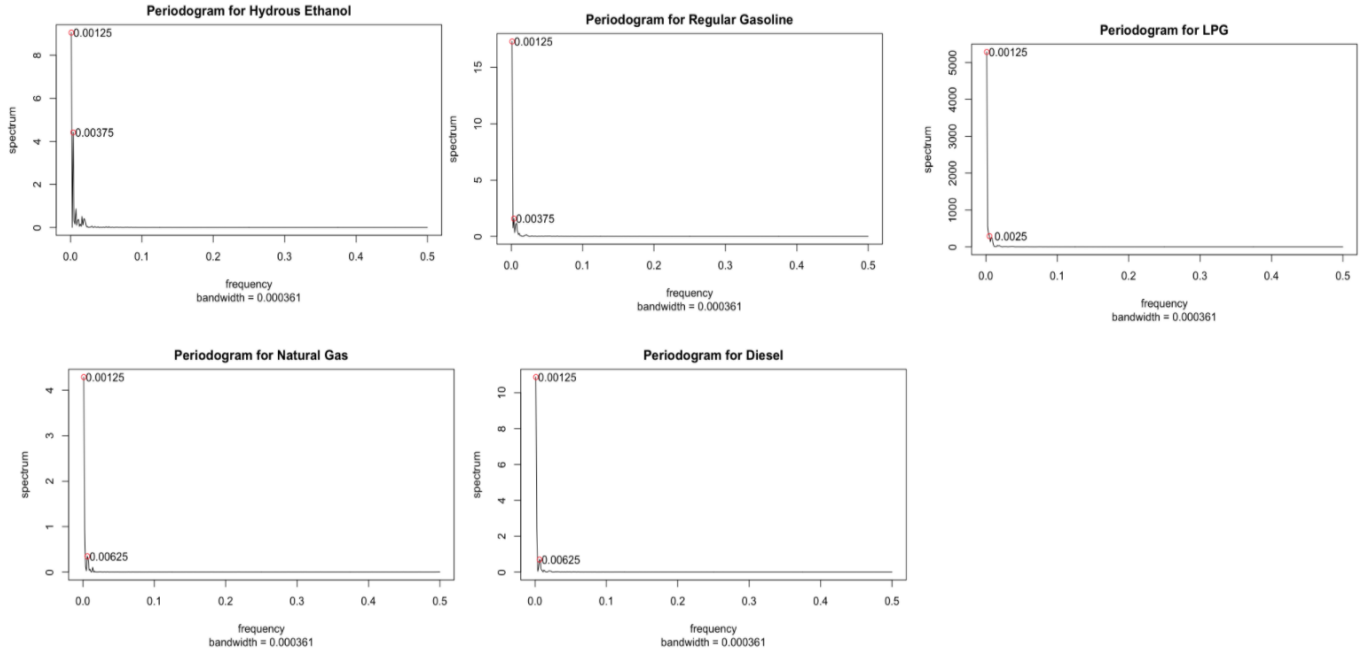
Because our data is weekly data, we set frequency equals 52 here for applying the decompose function in R. Although the decomposition shows there is clear seasonality of each product, in fact the seasonality is weak. We will talk about it in the next section (Spectral Analysis).



3.2.3 Spectral Analysis

For time series data, periodic behavior is an important indicator to analyze the data pattern. Spectral analysis is a technique that allows us to discover underlying periodicities. To perform spectral analysis, data is transformed from time domain to frequency domain by fourier transform, and then calculate power spectrum to determine what frequencies dominate the variance.

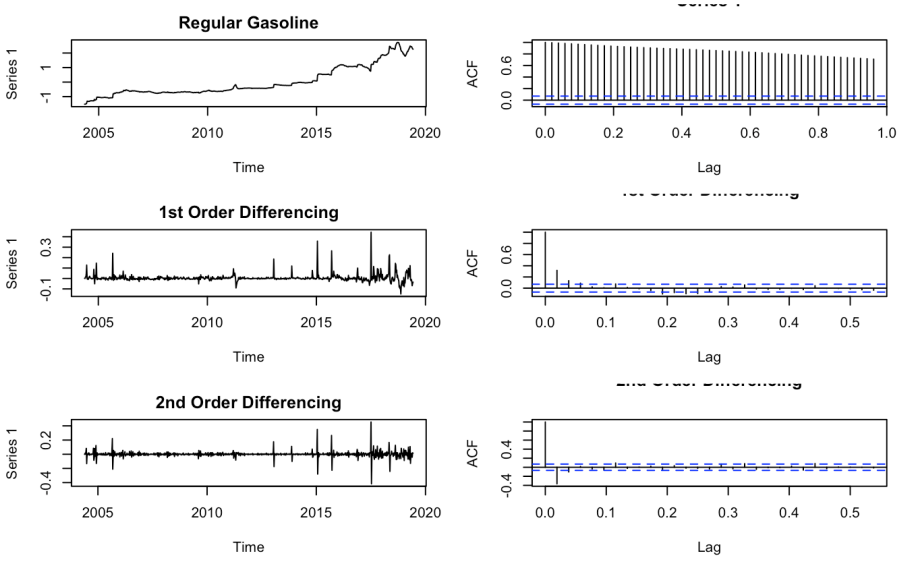
The figure below presents the periodogram plot for each product. All the products have the same result: the highest spike at frequency = 0.00125, which means 800 time points(week) pre cycle. And the largest the second spike among these products is at frequency = 0.00375 (266 week pre cycle). Such small frequencies indicate that all the products have a weak seasonality.



3.2 Forecasting - Arima

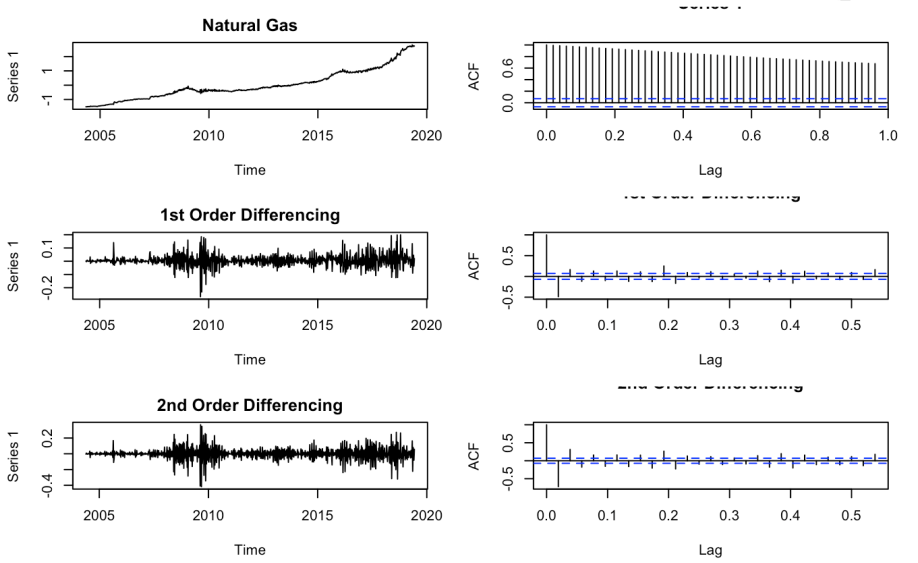
3.2.1 Stationarize the Series

To forecast its trend, we will start with ARIMA Modeling. First, we want to stationarize the series. As we can see, at the beginning, both of them are increasing and not stationary. After first order differencing, gasoline is stationary. For natural gas, it is more stationary after the second difference. To double check, we conduct the Dickey – Fuller test and it shows after the first differencing, both of them are stationary. So, we choose d = 1 for gasoline and d = 1or 2 for natural gas.



Augmented Dickey-Fuller Test

data: ng.diff.ts
Dickey-Fuller = -48.003, Lag order = 0, p-value = 0.01
alternative hypothesis: stationary

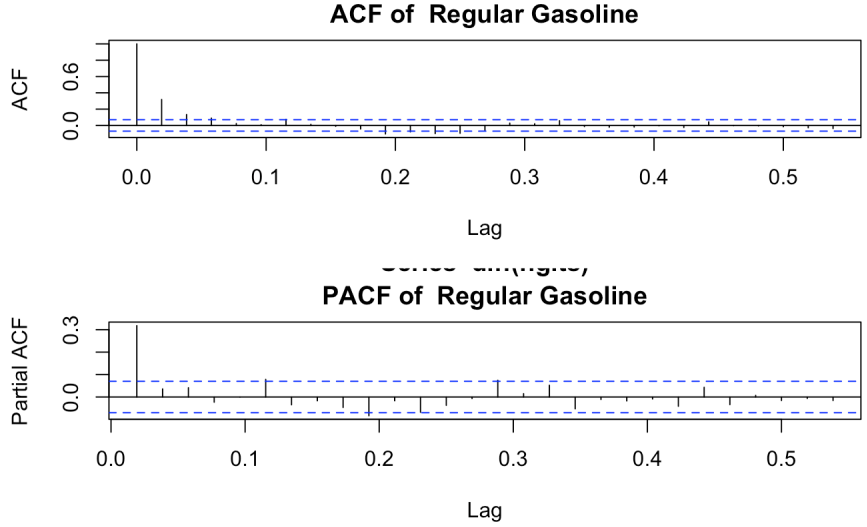


Augmented Dickey-Fuller Test

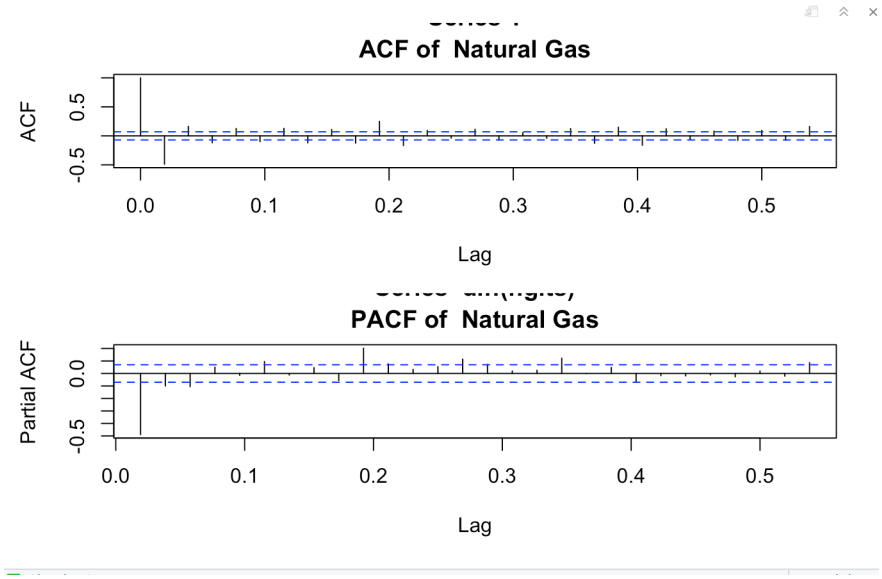
data: ng.diff.ts
Dickey-Fuller = -20.096, Lag order = 0, p-value = 0.01
alternative hypothesis: stationary

3.2.2 Find Optimal Parameters

Next, we plot ACF and PACF to find optimal parameters p and q. For regular gas, the ACF shows us the value of q should be 0 or 1. The PACF shows the value of p should be 1.



For natural gas, the ACF shows us the value of q should be 0, 1, 2 or 3. The PACF shows the value of p should be 1.



3.2.3 Build ARIMA Model

With the parameters in hand, we can now try to build an ARIMA model. The value found here might be an approximate estimate and we need to explore more (p,d,q) combinations. The one

with the lowest BIC and AIC should be our choice. Below are the optimal results for Regular Gas and Natural Gas.

The best Arima model for Regular gas is ARIMA(1, 1, 0).

```
ARIMA(1,1,0) with drift
Box Cox transformation: lambda= 0.8965418

Coefficients:
         ar1   drift
      0.2956  0.0050
s.e.  0.0341  0.0018

sigma^2 estimated as 0.001275:  log likelihood=1501.17
AIC=-2996.34   AICc=-2996.31   BIC=-2982.35
```

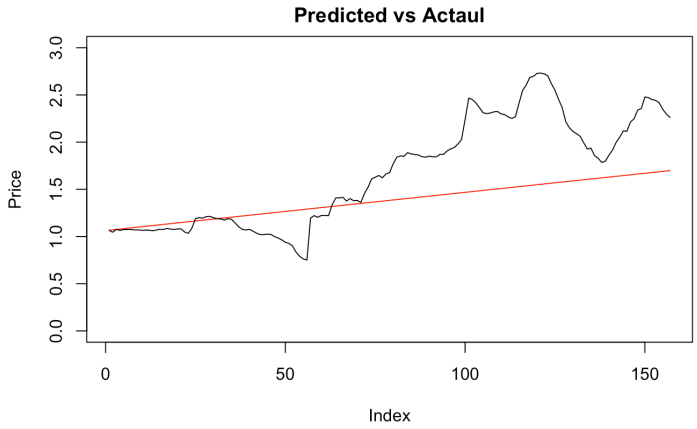
The best Arima model for Regular gas is ARIMA(1, 2, 3).

```
ARIMA(1,2,3)

Coefficients:
          ar1      ma1      ma2     ma3
      -0.9863  -0.5503  -0.9023  0.4986
s.e.   0.0094   0.0344   0.0276  0.0354

sigma^2 estimated as 0.00155:  log likelihood=1421.56
AIC=-2833.11   AICc=-2833.03   BIC=-2809.8
```

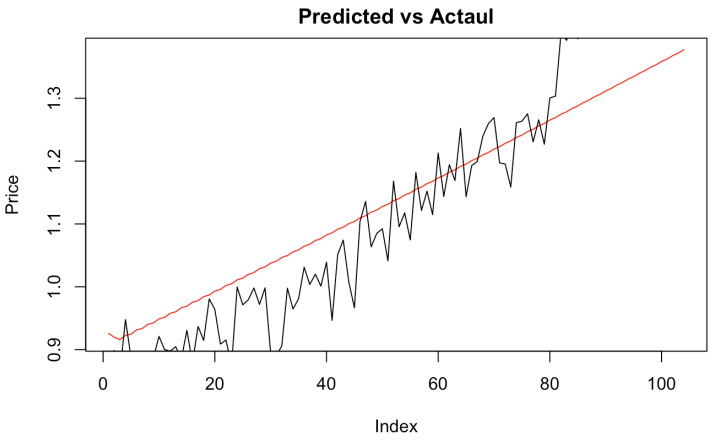
3.2.4 Make Prediction

Next, we will conduct predictions with our selected model to check the difference between our predictions and actual results.

From both graphs below, our models do follow the trend of actual data. But we noticed that the RMSE of natural gas is higher than regular gas. So we want to make further analysis for residuals.



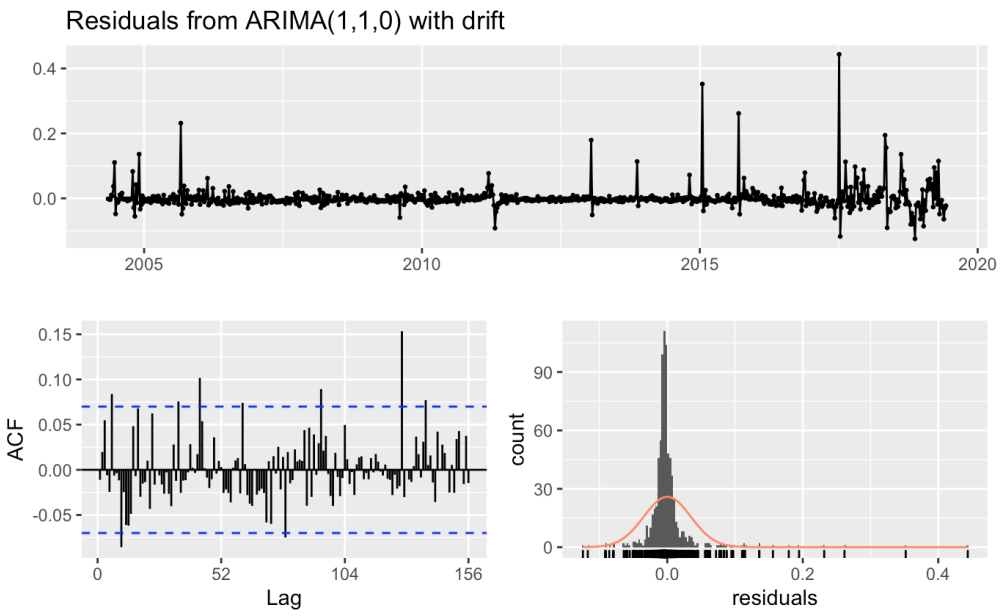

```
Regular Gasoline
                     ME      RMSE       MAE       MPE     MAPE      ACF1 Theil's U
Test set -0.2198984 0.3676017 0.2833857 -14.26648 19.20893 0.9705804  110.7162
```



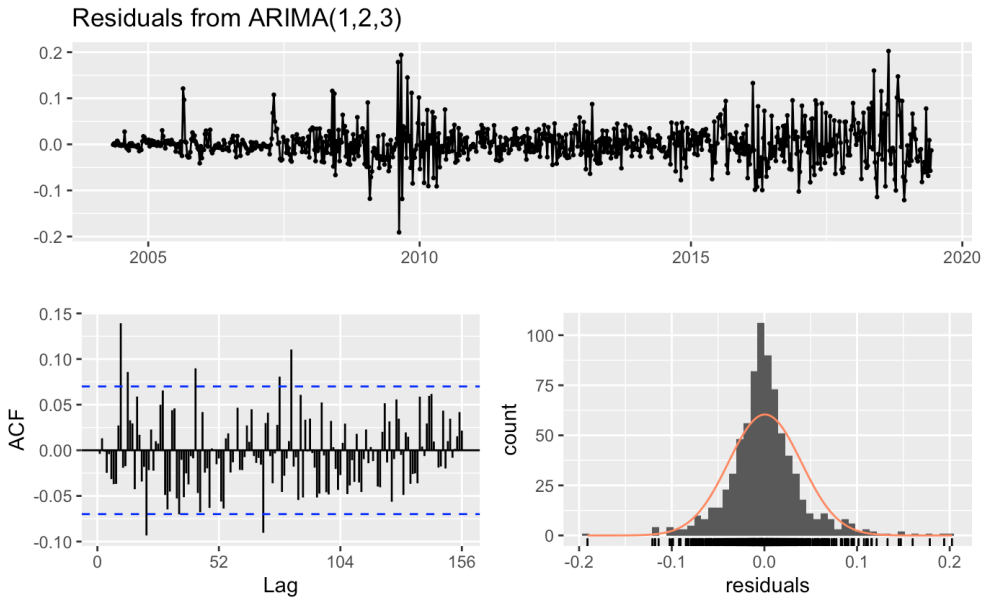

```
Natural Gas
                    ME      RMSE       MAE       MPE     MAPE      ACF1 Theil's U
Test set -0.310312 0.5415289 0.3562559 -20.07125 24.46913 0.9778159  90.60961
```

3.2.5 Model Diagnostics

Below are residuals analysis for both regular gas and natural gas.
For regular gas, we could notice residuals are welled-normalised distributed and it almost concentrated at mean = 0 with small variance.



For natural gas, we could notice obviously residuals have wide variance. It shows us we need to conduct further analysis for natural gas.

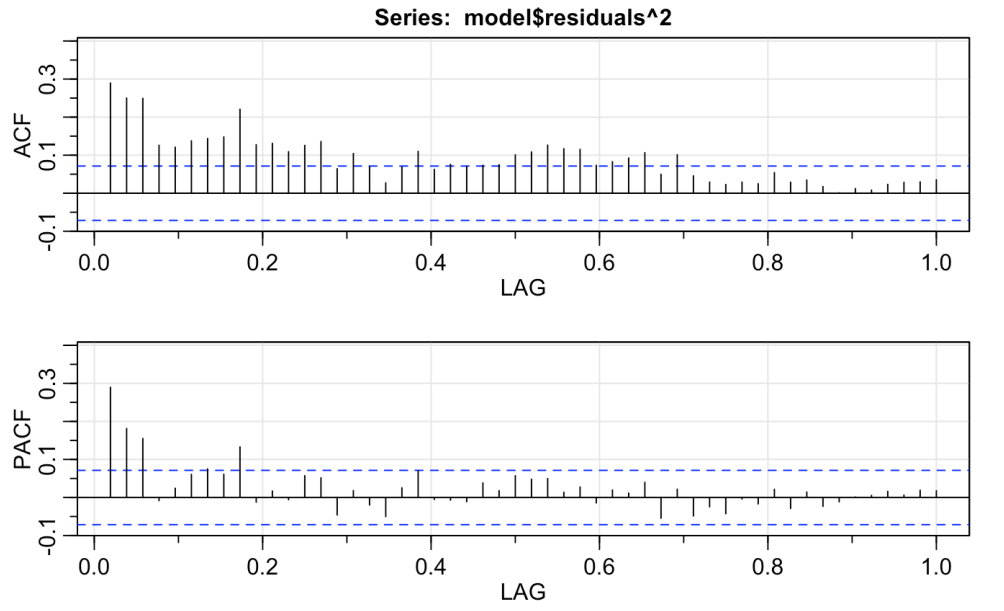


3.3 Forecasting - Volatility clustering

3.3.1 GARCH Model

We will use the GARCH model to simulate the residuals of Natural Gas. GARCH model is a very good approach to minimize the volatility effect.

First, we plot the ACF and PACF of squared residuals to find GARCH model parameters. In ACF, we noticed that the value of q should be selected among a wide range. In PACF, we noticed the value of p should be 1 or 2.



In the end, we select the Garch(1, 1) as our model with the lowest AIC. Below is the output of Garch(1,1).

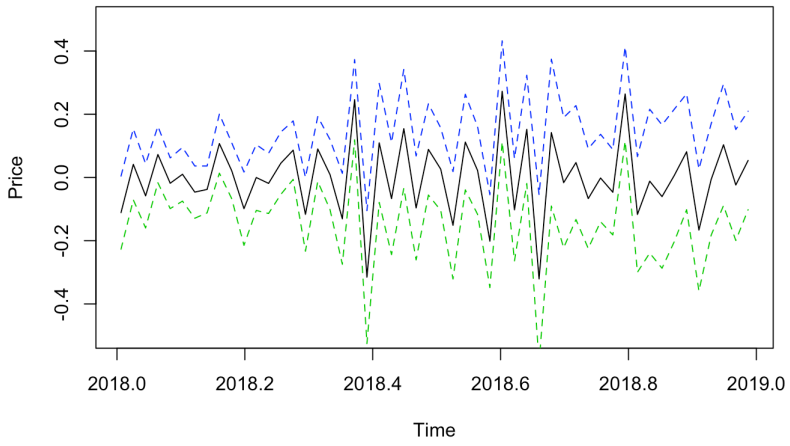
```
garch(x = model$residuals^2, order = c(1, 1), trace = F)

Model:
GARCH(1,1)

Residuals:
      Min        1Q    Median        3Q       Max
1.543e-07 1.985e-02 1.162e-01 3.968e-01 9.031e+00

Coefficient(s):
    Estimate  Std. Error  t value Pr(>|t|)
a0 2.327e-07   1.224e-08    19.02   <2e-16 ***
a1 5.322e-02   2.856e-03    18.64   <2e-16 ***
b1 9.361e-01   2.385e-03   392.55   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Diagnostic Tests:
        Jarque Bera Test

data:  Residuals
X-squared = 35764, df = 2, p-value < 2.2e-16


        Box-Ljung test

data:  Squared.Residuals
X-squared = 4.6781, df = 1, p-value = 0.03055
```

3.3.2 ARIMA + GARCH

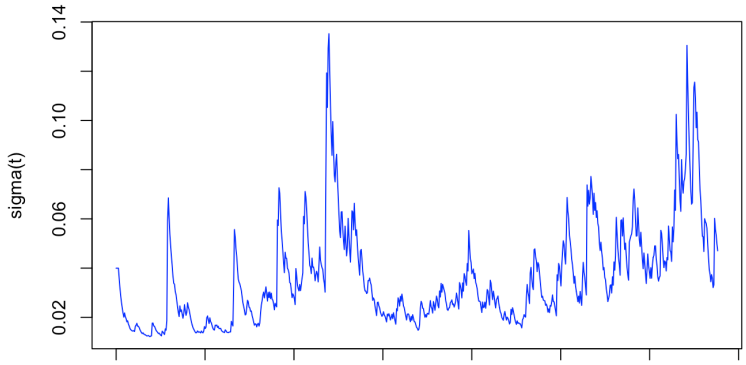
Now we could combine our ARIMA model and GARCH model together to predict the natural gas price. Below is the output of the combined model.

```
 GARCH Modelling

Call:
 garchFit(formula = ~arma(1, 3) + garch(1, 1), data = diff(ng.diff.ts))

Mean and Variance Equation:
 data ~ arma(1, 3) + garch(1, 1)
<environment: 0x7ff4bddbc9e0>
 [data = diff(ng.diff.ts)]

Conditional Distribution:
 norm

Coefficient(s):
         mu           ar1           ma1           ma2           ma3         omega        alpha1
 1.9267e-05  -9.4037e-01  -5.7990e-01  -8.1367e-01   4.1014e-01   2.2300e-05   2.6976e-01
      beta1
 7.6244e-01

Std. Errors:
 based on Hessian

Error Analysis:
         Estimate  Std. Error  t value Pr(>|t|)
mu       1.927e-05   1.792e-05    1.075  0.28235
ar1     -9.404e-01   3.081e-02  -30.524  < 2e-16 ***
ma1     -5.799e-01   4.936e-02  -11.748  < 2e-16 ***
ma2     -8.137e-01   6.186e-02  -13.153  < 2e-16 ***
ma3      4.101e-01   4.477e-02    9.162  < 2e-16 ***
omega    2.230e-05   7.650e-06    2.915  0.00356 **
alpha1   2.698e-01   4.076e-02    6.619 3.62e-11 ***
beta1    7.624e-01   2.527e-02   30.173  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Below is the prediction result with the ARIMA + GARCH model. We could see the variance is smaller at the beginning of Year 2018 and becomes larger while reading the end of Year 2018.



The variance change is matched with the conditional variance plot below.



3.4 Forecasting - Periodic Analysis

3.4.1 Prediction

ARIMA Model with Fourier Terms is typically used when data has a long seasonal period. We coerce the 5 product prices into weekly data, which means the frequency is 52. It is a challenge to fit an ARIMA model with data having a long seasonal period, such as daily or weekly data, since seasonal versions of ARIMA models are designed for shorter periods such as monthly or quarterly data. The problem is that there are m−1 parameters to be estimated for the initial seasonal states where m is the seasonal period. So for large m, the estimation becomes almost impossible. For such data we can use a Fourier series approach where the seasonal pattern is modelled using Fourier terms with short−term time series dynamics allowed in the error, although in this case the seasonality is assumed to be fixed, i.e. the pattern is not allowed to change over time.

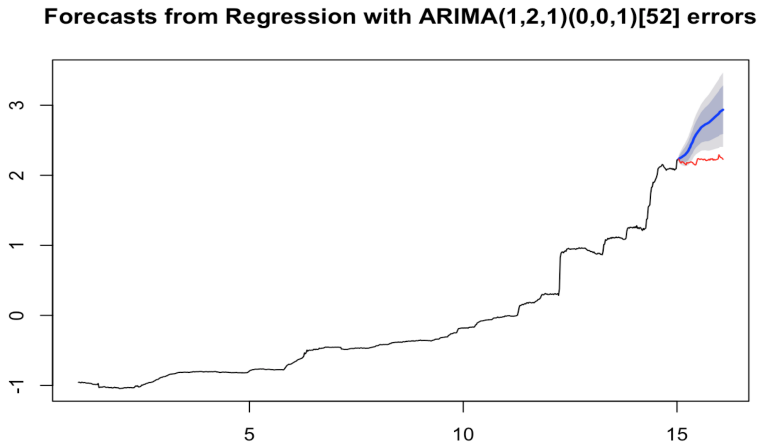
3.4.2 Model Diagnostics

The model is considered adequate only if the p-value associated with the Ljung-Box Q Statistic is higher than a given significance, 0.05.

3.4.3 Models for Products Respectively

3.4.3.1 LPG

Prediction:

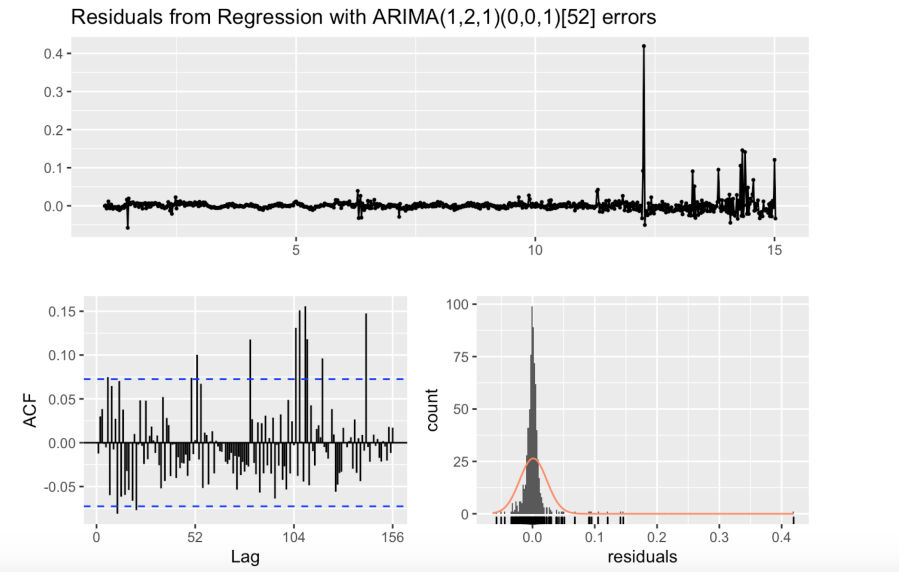



```
## $aicc
## [1] -3479.988
##
## $i
## [1] 2
##
## $fit
## Series: lpg.trn
## Regression with ARIMA(1,2,1)(0,0,1)[52] errors
##
## Coefficients:
##          ar1      ma1    sma1   S1-52   C1-52    S2-52   C2-52
##       0.2013  -0.9884  0.1026  -0.017  -0.029  -0.0077  0.0154
## s.e.  0.0370   0.0058  0.0419   0.013   0.013   0.0064  0.0064
##
## sigma^2 estimated as 0.0004828:  log likelihood=1748.09
## AIC=-3480.19   AICc=-3479.99   BIC=-3443.47
```

Model Diagnostics:

```
##                       ME       RMSE         MAE          MPE      MAPE
## Training set  0.001128669 0.02183694 0.008386885   0.05765925  2.929964
## Test set     -0.400693602 0.44917555 0.400715022 -18.05379783 18.054757
##                    MASE        ACF1 Theil's U
## Training set 0.03815339 -0.01230901        NA
## Test set     1.82292200  0.93789436  27.56134
```

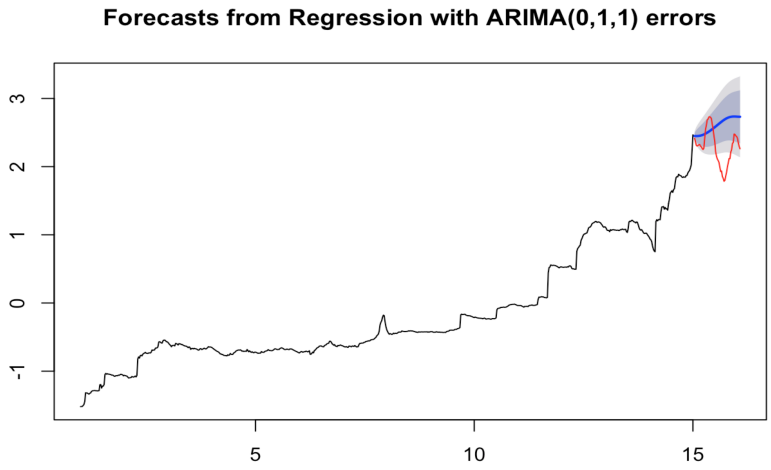
```
checkresiduals(bestfit$fit)
```



The plot shows that the prediction does not follow the true trend of prices. The distribution of residuals are left-skewed. p-value = 0.05519, which suggests the model is somewhat adequate. However the plot tells us the result is not close to the actual data. Therefore, the model is not performing well on predicting the price of LPG.

3.4.3.2 Regular Gasoline

Prediction:

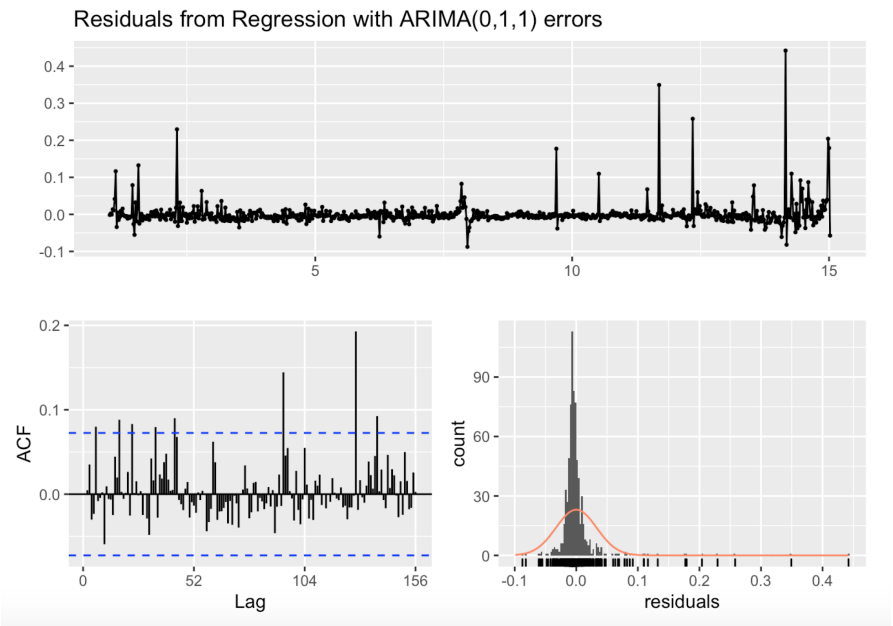


```
## $aicc
## [1] -2891.996
##
## $i
## [1] 2
##
## $fit
## Series: rg.trn
## Regression with ARIMA(0,1,1) errors
##
## Coefficients:
##          ma1   drift    S1-52   C1-52    S2-52    C2-52
##       0.2371  0.0054  -0.0444  0.0013  -0.0034  -0.0016
## s.e.  0.0367  0.0015   0.0177  0.0177   0.0088   0.0088
##
## sigma^2 estimated as 0.001096:  log likelihood=1453.08
## AIC=-2892.15   AICc=-2892   BIC=-2860.01
```

Model Diagnostics:

```
##                       ME       RMSE        MAE        MPE      MAPE
## Training set -5.337257e-06 0.03294528 0.01386406   0.777765  3.314007
## Test set     -3.163084e-01 0.44756366 0.36734711 -15.551011 17.447740
##                    MASE          ACF1 Theil's U
## Training set 0.05668099 -0.0008170911        NA
## Test set     1.50183963  0.9716730033  8.405515
```

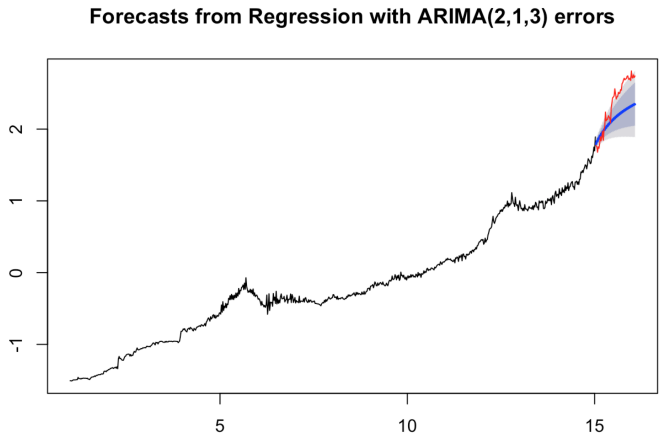
```
checkresiduals(bestfit$fit)
```



The prediction line looks close to the true trend. However, the distribution of the residuals is still left-skewed. From the p-value of Ljung-box, the residuals are autocorrelated.

3.4.3.3 Natural Gas

Prediction:

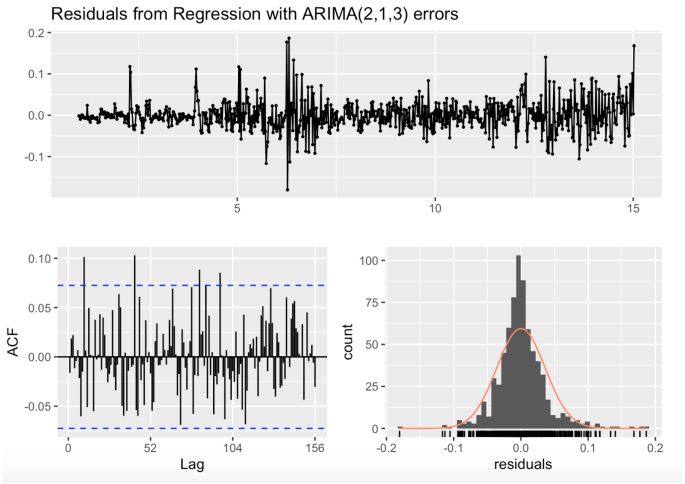



```
## $aicc
## [1] -2759.368
##
## $i
## [1] 1
##
## $fit
## Series: ng.trn
## Regression with ARIMA(2,1,3) errors
##
## Coefficients:
##           ar1     ar2      ma1      ma2     ma3   drift    S1-52    C1-52
##       -0.0361  0.9360  -0.5758  -0.8623  0.5465  0.0049  -0.0025  -0.0008
## s.e.   0.0300  0.0289   0.0417   0.0343  0.0338  0.0014   0.0081   0.0081
##
## sigma^2 estimated as 0.001309:  log likelihood=1388.81
## AIC=-2759.62   AICc=-2759.37   BIC=-2718.29
```

Model Diagnostics:

```
##                       ME       RMSE        MAE      MPE      MAPE
## Training set 4.543788e-05 0.03595606 0.02476605 3.610143 13.28872
## Test set     2.218213e-01 0.29509045 0.25368435 8.296008 10.09520
##                    MASE        ACF1 Theil's U
## Training set 0.1060918 -0.01626525        NA
## Test set     1.0867224  0.92434679  3.503991
```

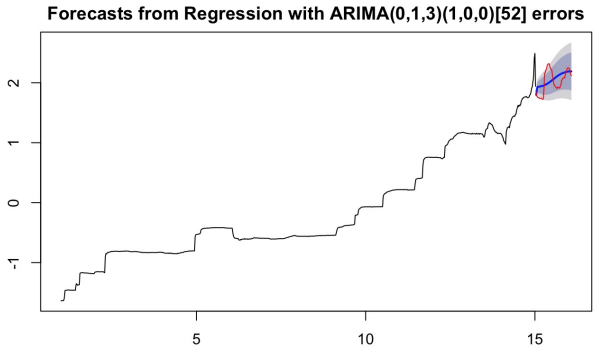
```
checkresiduals(bestfit$fit)
```



The prediction line looks close to the true trend. However, the distribution of the residuals is close to normal. From the p-value of Ljung-box, P-value = 0.08053, which suggests the residuals are not autocorrelated. The model is adequate for predicting the prices of Natural Gas. However, the plot suggests this model is not perfect.

3.4.3.4 Diesel

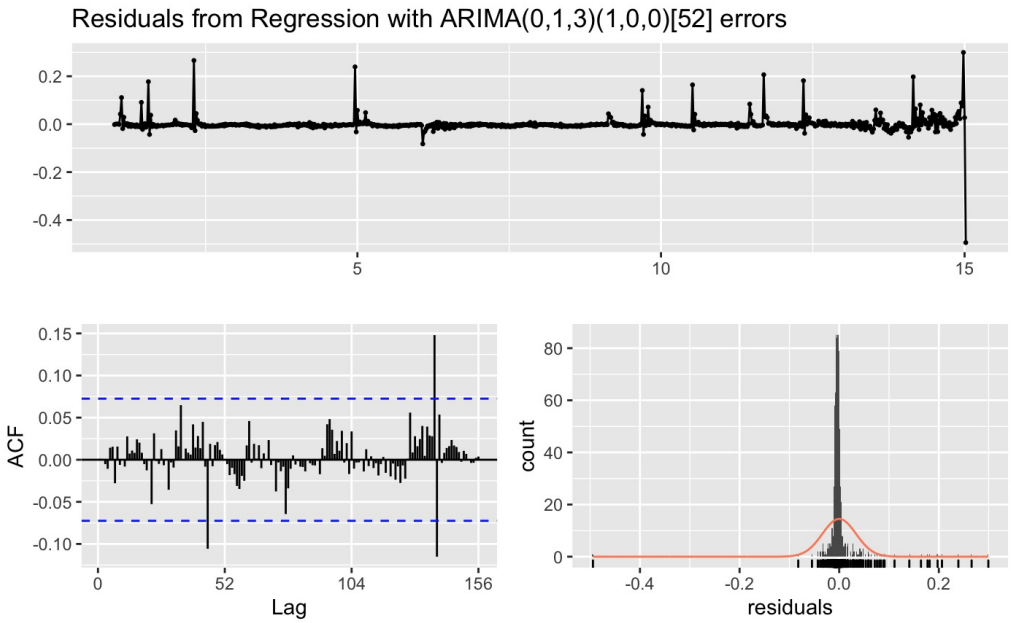
Prediction:




```
$aicc
[1] -2859.29

$i
[1] 1

$fit
Series: d.trn
Regression with ARIMA(0,1,3)(1,0,0)[52] errors

Coefficients:
         ma1      ma2      ma3    sar1   drift    S1-52   C1-52
      0.2457  -0.1889  -0.0840  0.0073  0.0049  -0.0310  0.0057
s.e.  0.0445   0.0424   0.0455  0.0536  0.0012   0.0145  0.0145

sigma^2 estimated as 0.001144:  log likelihood=1437.74
AIC=-2859.49   AICc=-2859.29   BIC=-2822.76
```

Model Diagnostics:

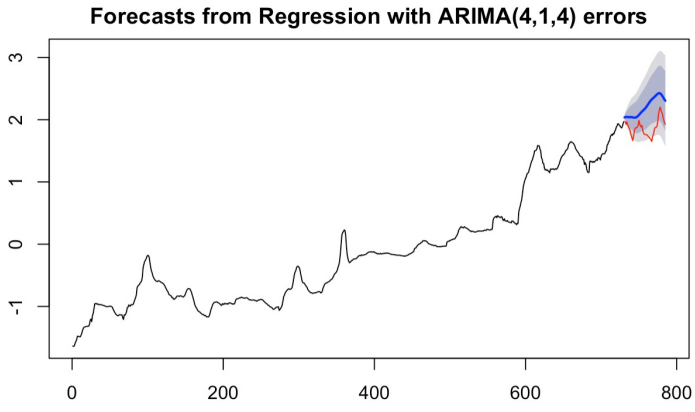
	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
Training set	-2.733988e-06	0.03364303	0.0119585	0.5527423	2.212726	0.04666222	0.001119424	NA
Test set	-2.834594e-02	0.17371994	0.1511915	-2.0785106	7.552033	0.58995091	0.942975907	3.138609



This forecast looks like a pretty good abstraction of the basic seasonal cycle projected into the future. P-value = 0.8787, which suggests the model is adequate for prediction. In general, the model works pretty well.

3.4.3.5 Hydrous Ethanol

Prediction:

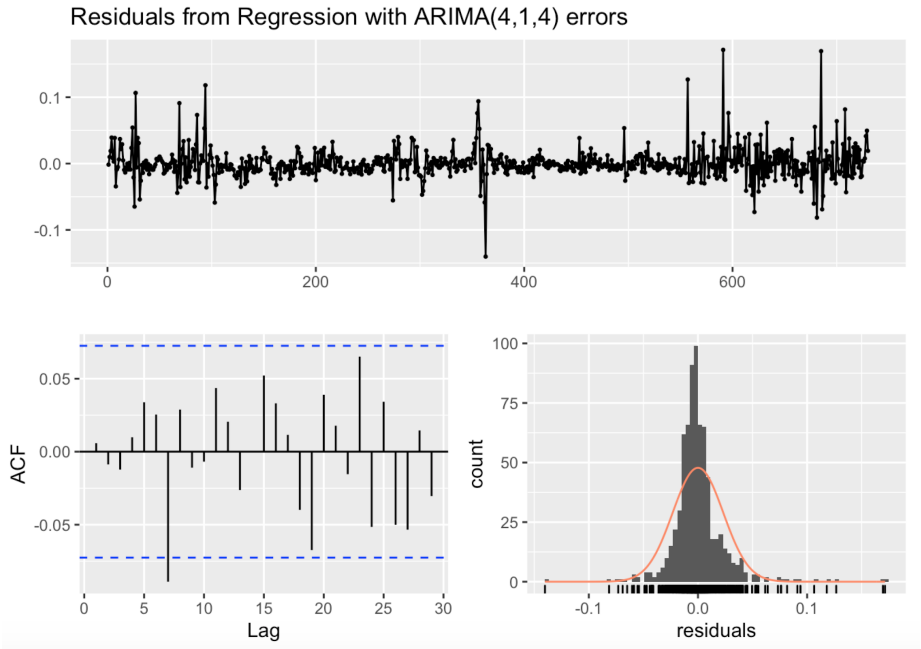


```
## $aicc
## [1] -3383.144
##
## $i
## [1] 4
##
## $fit
## Series: he.trn
## Regression with ARIMA(4,1,4) errors
##
## Coefficients:
##          ar1     ar2     ar3      ar4     ma1      ma2      ma3     ma4
##       -0.428  1.2199  0.3062  -0.5768  0.8487  -0.6804  -0.4336  0.3007
## s.e.   0.158  0.1865  0.0800   0.1098  0.1616   0.2558   0.1250  0.0749
##        drift    S1-52   C1-52    S2-52    C2-52    S3-52    C3-52    S4-52
##       0.0051  -0.1013  0.0407  -0.0200  -0.0004  -0.0109  -0.0028  -0.0105
## s.e.  0.0019   0.0222  0.0223   0.0114   0.0114   0.0070   0.0070   0.0044
##         C4-52
##       -0.0004
## s.e.   0.0044
##
## sigma^2 estimated as 0.0005489:  log likelihood=1710.05
## AIC=-3384.11   AICc=-3383.14   BIC=-3301.46
```

Model Diagnostics:

```
##                       ME       RMSE        MAE         MPE      MAPE
## Training set -6.535982e-06 0.0231389 0.01437184  -8.217705 18.38079
## Test set     -3.304496e-01 0.3692898 0.33044960 -18.086665 18.08667
##                    MASE        ACF1 Theil's U
## Training set  0.7847429 0.005819074        NA
## Test set     18.0434759 0.942132910  8.864018
```

```
checkresiduals(bestfit$fit)
```



The graph shows that our prediction for the price of Hydrous Ethanol is paralleling the trend of true data. There is no indication of autocorrelation between residuals from lag 1-20. The residuals are normally distributed. This is a good model for prediction for this product.

3.5 Forecasting - Multivariate Analysis

3.5.1 Methodology

Vector Autoregression model is an extension version of autoregressive model. An autoregressive model of order p, abbreviated AR(p), is the simple univariate model of the form of

$$x_t = \varphi_1 x_{t-1} + \ldots + \varphi_1 x_{t-p} + w_t$$

where w_t is assumed as white noise with mean zero and constant variance and $x_p$presented the lagged values[2](Shumway & Stoffer,2017) However, the model that only contains its own lagged value may not be able to explain all the variability for the more complex data. Thus VAR (stands for vector autoregression) is introduced. It extends the idea of AR models. In a VAR model we regard lagged values of multiple dependent variables as the regressors. For example, a VAR(p) model of two variables $X_t$and Y_t (k=2) is given by the equations[3]

$$x_t = \varphi_{1.0} + \varphi_{1.1} x_{t-1} + \ldots + \varphi_{1.p} x_{t-p} + \gamma_{1.1} y_{1.1} + \ldots + \gamma_{1.p} y_{1.p} + w_{1t}$$
$$y_t = \varphi_{2.0} + \varphi_{2.1} y_{t-1} + \ldots + \varphi_{2.p} y_{t-p} + \gamma_{2.1} y x_{2.1} + \ldots + \gamma_{2.p} x_{2.p} + w_{2t}$$

or in more compactly way:

$$X_t = \Phi + A_1 X_{t-1} + \ldots + A_p X_{t-p} + e_t$$
$$\text{where } X_t = \begin{pmatrix} x_t \\ y_t \end{pmatrix}, \Phi = \begin{pmatrix} \varphi_{1.0} \\ \varphi_{2.0} \end{pmatrix}, A_1 = \begin{pmatrix} \varphi_{1.1} & \gamma_{1.1} \\ \varphi_{2.1} & \gamma_{2.1} \end{pmatrix}, \cdots,$$
$$A_p = \begin{pmatrix} \varphi_{1.p} & \gamma_{1.p} \\ \varphi_{2.p} & \gamma_{2.p} \end{pmatrix} \text{ and } e_t = \begin{pmatrix} w_{1t} \\ w_{2t} \end{pmatrix}$$

Here ϕs and γs can be estimated using OLS on each equation. The matrix of ϕs and γs (A) is called variance-covariance matrix; it contains the variances of the variable on its diagonal elements and covariances of the errors on the off-diagonal elements. Thus the VAR model is more flexible than the AR model. As the one of the most successful and popular, VAR model is widely used for describing the dynamic behavior of data and predicting the future value, for example, in economics it can forecast the macroeconomic variable, such as GDP, Money supply and unemployment; or in financial, it can predict spot prices and future prices of securities orforeign exchange rates across markets. [4]

---

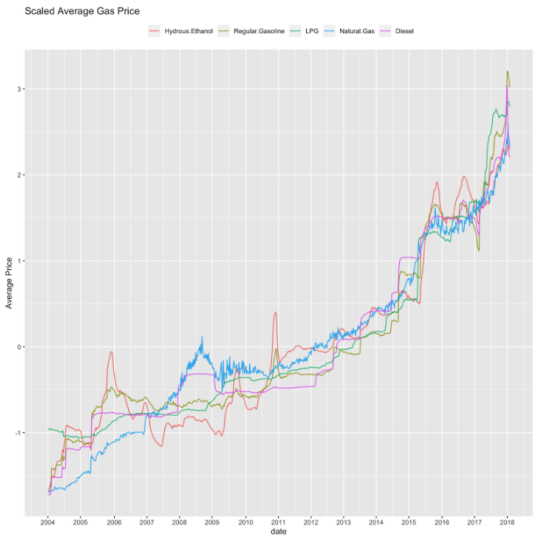
[2] Shumway, R. H., & Stoffer, D. S. (2017). Time series analysis and its applications: with R examples (3rd ed. New York: Springer. page 94

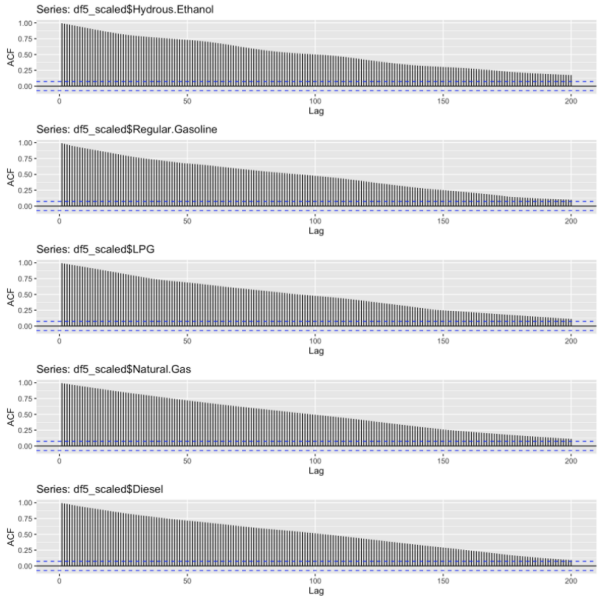
[3] Hanck, C., Arnold, M., Gerber, A., & Schmelzer, M. (2019, August 30). Introduction to Econometrics with R. Retrieved from https://www.econometrics-with-r.org/16-1-vector-autoregressions.html

[4] http://www.ams.sunysb.edu/~zhu/ams586/VAR_Lecture2.pdf

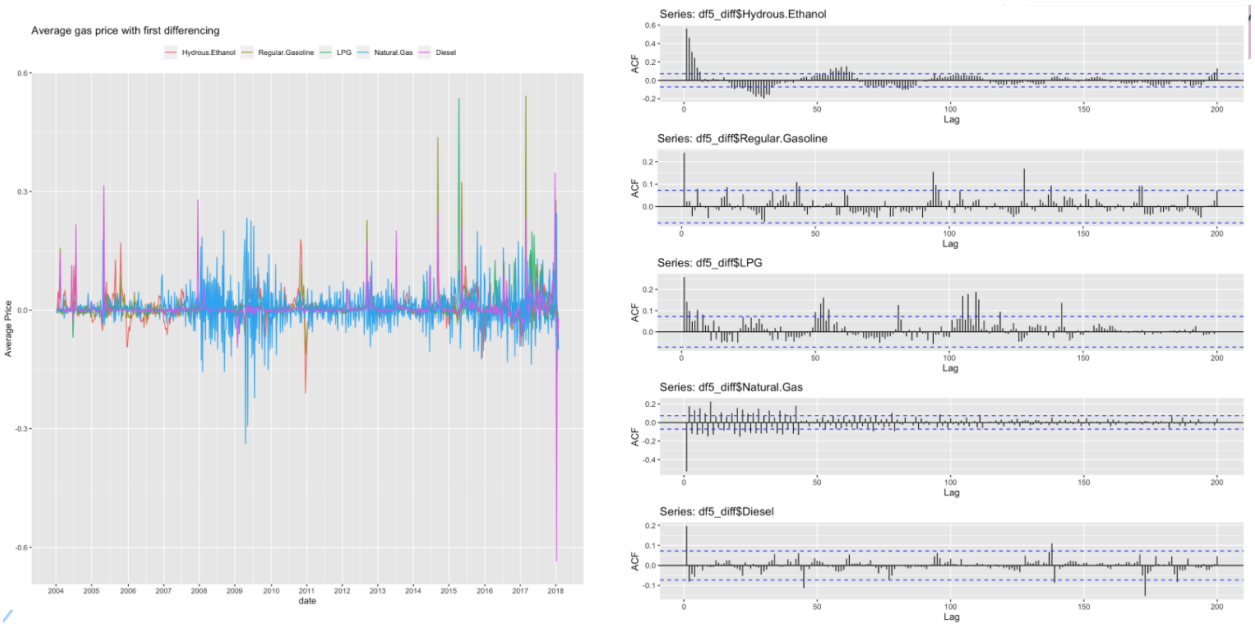
3.5.2 Stationary

Same as autoregression models, VAR models also assume that all the time series data are stationary. According to the figure below, the left plot shows all the products have clear trends and ACF plots hold a slow-decaying pattern. Clearly, all of our products' data are not stationary. Thus we take the differencing here.

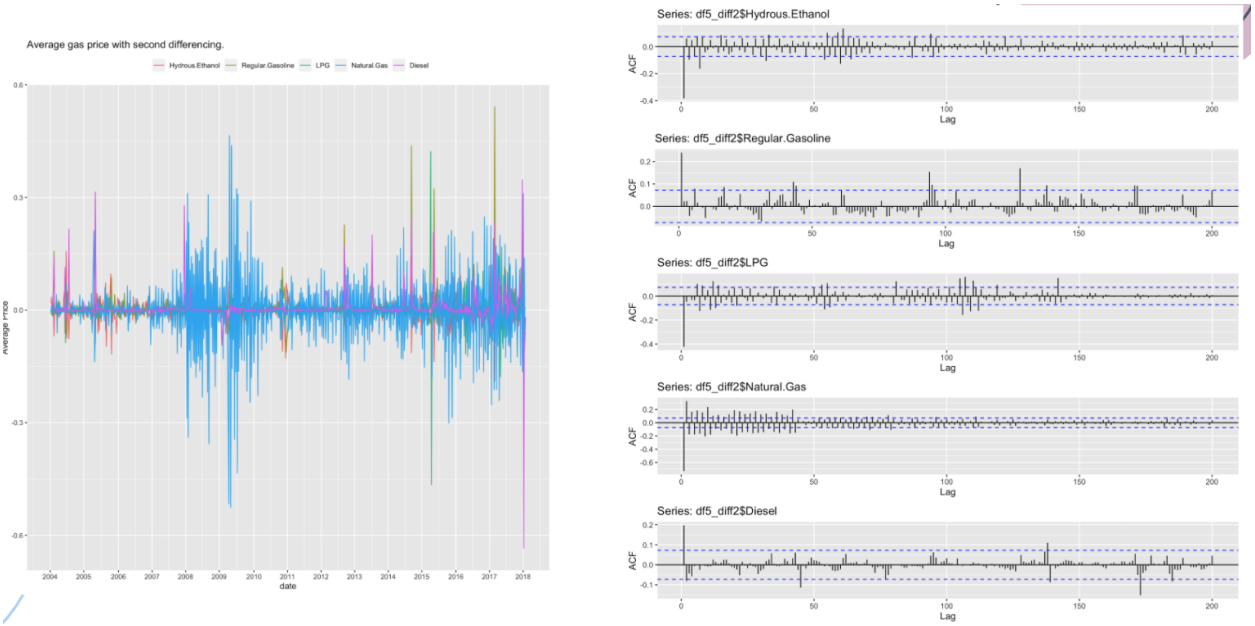




The figure below shows the average price with the first differencing, and their ACF. We noticed that Hydrous Ethanol and LPG still hold a little slow-decaying pattern on their ACF plot. Then we decided to take the second order differencing.



The figure below shows the average price with the 2 order differencing, and their ACF. However, some products' lags reach negative too soon. We considered it as the signal of over-differencing. Thus we decided to use the data with first order differencing to fit the VAR models.

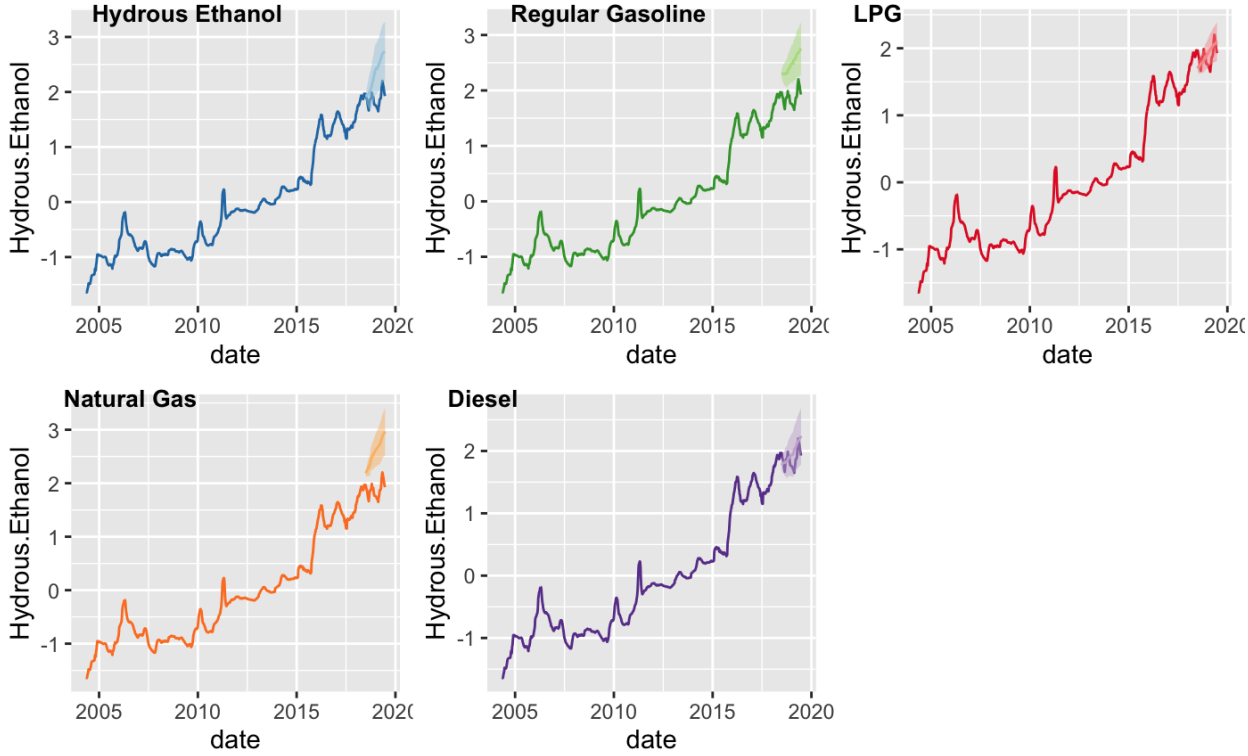


3.5.3 VAR models with product price as exogenous variables

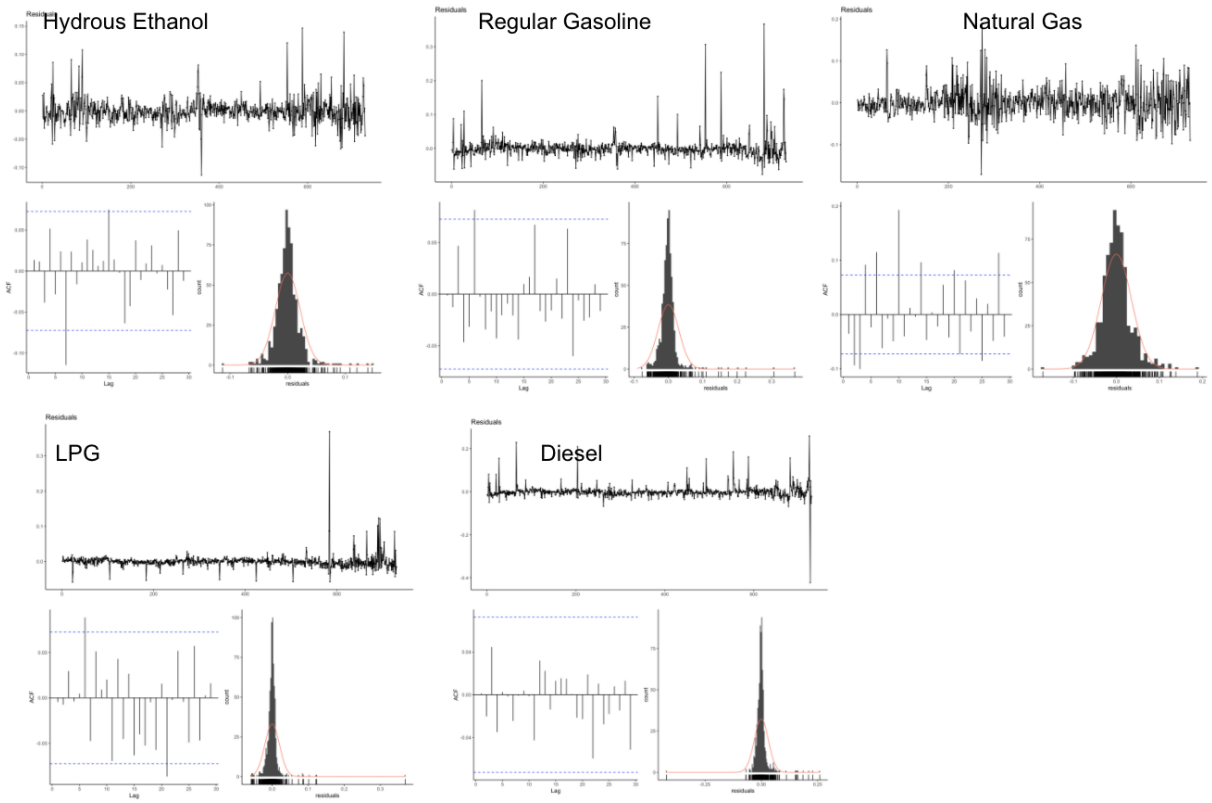
According to the EDA, we found that all the products reflect a similar pattern (long seasonality, increasing trend) over this 15 years. Thus, we guessed that these products may affect each other. So we first set other product prices as exogenous variables. In other words, the current price of

one product not only is impacted by his own lagged value but also receives the influence of other products' lagged values.

We fitted a VAR model with average prices of 5 products, and obtained a model with 0.9 R squared. The figure below shows the original data and their prediction interval area. Obviously, interval areas of Diesel, LPG and Hydrous Ethanol cover the real data, indicating that they have relatively good predictions. Then we also did some residuals analysis.



The figure below presents a time plot of the residuals, the corresponding ACF, and a histogram for each product. All the products' residuals almost perfectly met the conditions of stationary.



Furthermore, for all five products, the all products' p-value (as the screen shot displayed below) are larger than the significant level indicating rejection of the null hypothesis that the time series isn't autocorrelated. Therefore, we have enough evidence to conclude that: the residuals are white noise, and we obtain a good prediction.

```
        Box-Ljung test

data:  residuals(var.fit2)[, i]
X-squared = 0.1318, df = 1, p-value = 0.7166


        Box-Ljung test

data:  residuals(var.fit2)[, i]
X-squared = 0.0013848, df = 1, p-value = 0.9703


        Box-Ljung test

data:  residuals(var.fit2)[, i]
X-squared = 0.015084, df = 1, p-value = 0.9023


        Box-Ljung test

data:  residuals(var.fit2)[, i]
X-squared = 0.9275, df = 1, p-value = 0.3355


        Box-Ljung test

data:  residuals(var.fit2)[, i]
X-squared = 0.00098154, df = 1, p-value = 0.975
```

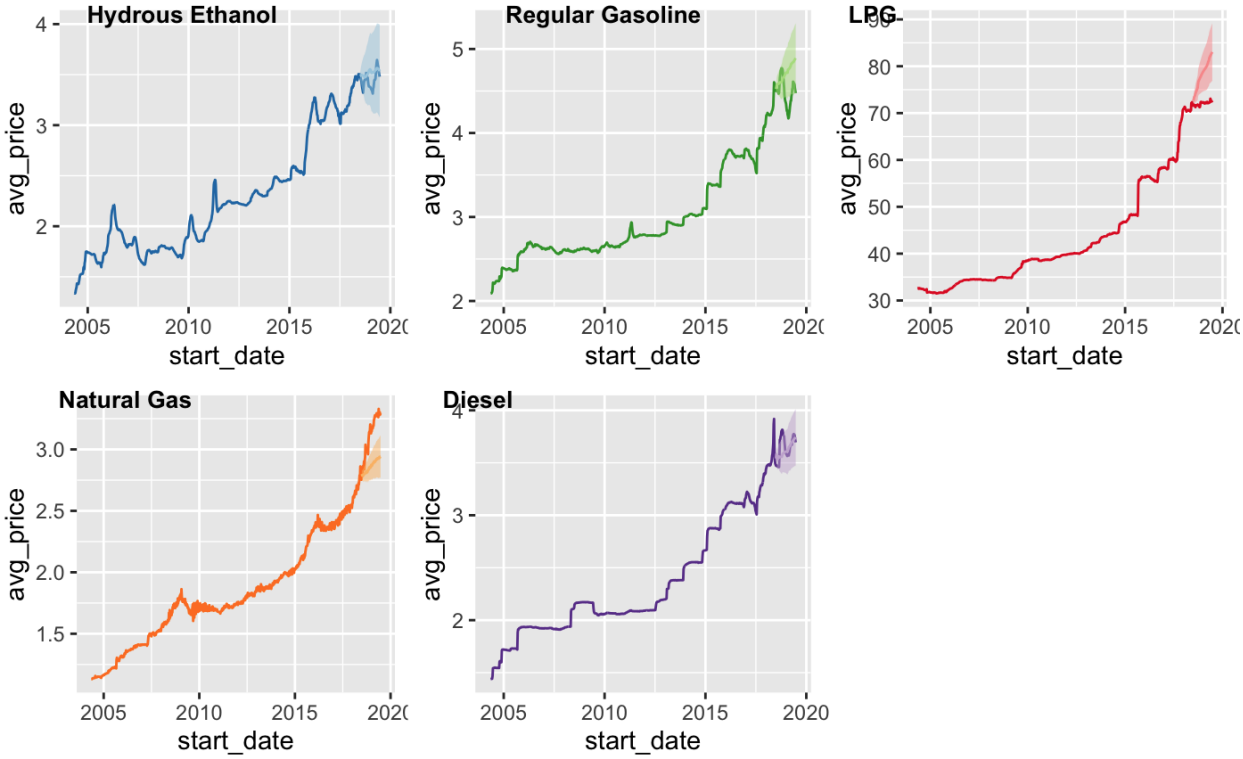
However, by the accuracy table (as the table displayed below), we noticed that this model led to a serious overfitting problem, since for all the products, the test errors are larger than the train error. And 0.9 R squared somehow also proved the overfitting problem. As the result, we decided to try other variables, to see whether we can fix the overfitting problem

	product	err_type	ME	RMSE	MAE	MPE	MAPE
1	Hydrous.Ethanol	train	-1.274e-04	0.02164	0.014323	8.8169	15.336
2	Regular.Gasoline	train	7.453e-05	0.03022	0.015172	1.0028	4.285
3	Natural.Gas	train	-8.667e-05	0.02038	0.009103	1.6804	4.020
4	LPG	train	-1.631e-04	0.03498	0.025236	3.0698	13.830
5	Diesel	train	-2.297e-05	0.03123	0.014716	0.5033	2.928
6	Hydrous.Ethanol	test	-4.647e-01	0.53156	0.466161	-25.1495	25.225
7	Regular.Gasoline	test	-1.991e-01	0.39598	0.332034	-10.4395	15.447
8	Natural.Gas	test	-3.780e-01	0.42877	0.377999	-17.0151	17.015
9	LPG	test	4.850e-01	0.52692	0.484975	19.3837	19.384
10	Diesel	test	8.067e-02	0.19910	0.133131	3.4072	6.190

3.5.4 VAR models with price information as exogenous variables

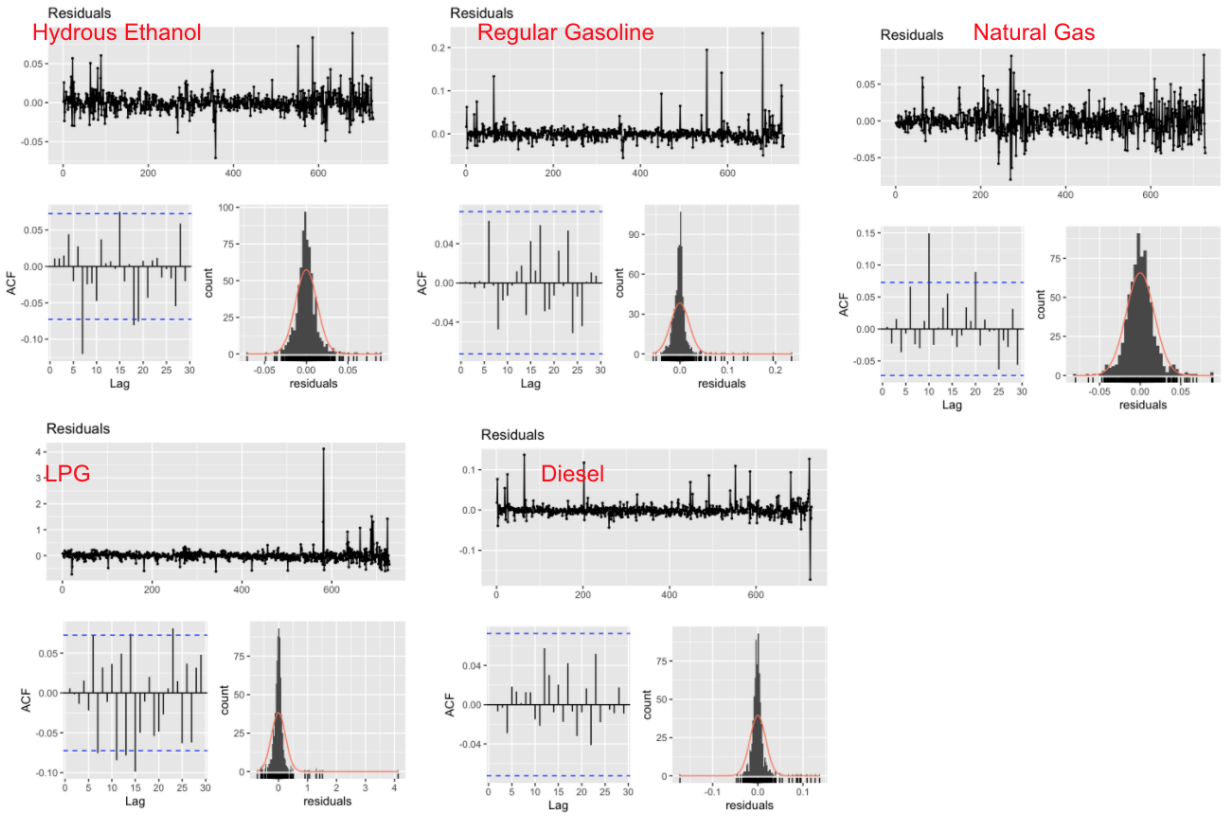
By EDA, we also realized that different products have different fluctuate intervals for each time collection. The first two models only assume that the current price is impacted by its average price. But from the perspective of econometrics[5](Mishra,2012), when measuring gas price, the other price information of the same time collection has significant influence too. Thus we also selected “average price of resale in each collection”, “minimum price of resale in each collection”, “maximum price of resale in each collection” , “standard deviation price of resale in each collection”and “coefficient of variation about distribution of each collection” as exogenous variables.

Implemented the model with new variables, we obtained the prediction (The figure below shows the original data and their prediction interval area.) This time, Hydrous Ethanol, Regular Gasoline and Diesel got relatively good predictions.



According to the residuals analysis. All the products’ residuals are met the features of white.

5 Mishra, P. (2012). Forecasting Natural Gas Price - Time Series and Nonparametric Approach . *Proceedings of the World Congress on Engineering, I*. Retrieved from https://pdfs.semanticscholar.org/b558/455f308c142f9be5990785022df8e91062d2.pdf



And all the p-values from Box-Ljung test are significant, indicating that the residuals are independent. The predictions by this models are also good.

```
        Box-Ljung test

data:  residuals(var.fit_HE)[, 3]
X-squared = 0.088645, df = 1, p-value = 0.7659


        Box-Ljung test

data:  residuals(var.fit_HE)[, 3]
X-squared = 0.000648, df = 1, p-value = 0.9797


        Box-Ljung test

data:  residuals(var.fit_HE)[, 3]
X-squared = 0.024801, df = 1, p-value = 0.8749


        Box-Ljung test

data:  residuals(var.fit_HE)[, 3]
X-squared = 0.0084985, df = 1, p-value = 0.9265


        Box-Ljung test

data:  residuals(var.fit_HE)[, 3]
X-squared = 0.0084985, df = 1, p-value = 0.9265
```

As for the overfitting problem, it still exists. (As the table below shows) The test error is still larger than the train errors. But since the gap between the two errors is shrinked. It somehow indicates that the VAR model with price information has a better performance on overfitting than the VAR model with other product prices.

Except for Natural Gas, the other four products hold smaller test errors by this model. It may indicate that natural gas's price is more sensitive to the other product price.

	product	err_type	ME	RMSE	MAE	MPE	MAPE
1	Hydrous.Ethanol	train	0.0000	0.0130	0.0086	-0.0004	0.3914
2	Regular.Gasoline	train	0.0000	0.0193	0.0097	-0.0015	0.3233
3	Natural.Gas	train	0.0002	0.2434	0.1173	-0.0017	0.2612
4	LPG	train	0.0000	0.0174	0.0125	-0.0075	0.6777
5	Diesel	train	0.0000	0.0175	0.0093	-0.0025	0.3953
6	Hydrous.Ethanol	test	-0.0758	0.1072	0.0885	-2.2501	2.6028
7	Regular.Gasoline	test	-0.2237	0.3139	0.2635	-5.1534	5.9929
8	Natural.Gas	test	-6.2491	6.7924	6.2491	-8.6603	8.6603
9	LPG	test	0.2526	0.2781	0.2532	7.9260	7.9479
10	Diesel	test	0.0281	0.1087	0.0815	0.6971	2.2119

4. CONCLUSION & DISCUSSION

We have fitted 3 distinct models- ARIMA+Garch model, ARIMA with Fourier Terms, and VAR models.

First, the ARIMA model shows a good fit with regular gas. But for natural gas, we have to conduct the GARCH model to fit its volatility as its residuals have unstable variance in the ARIMA model. It tells us the price of natural gas might be highly various to predict.

Second, the prediction results of the ARIMA model with Fourier Terms that we fitted for Diesel are the most accurate. The pattern seems like a pretty good abstraction of basic seasonal cycles into the future. The Ljung-Box also suggests that the model is adequate for predicting the prices of Diesel. However, surprisingly, the model does not perform as well on Diesel as on the other four products. Either the prediction line fails to follow the true data, or the residuals are non-stationary. The relatively higher RMSE means the prediction drifts away from the actual prices.

Last, we fitted VAR models using both average product price and other price information separately in order to compare how the model performs with different response variables. Eventually, we found that the VAR model fitted with average product price was the best for predicting price of natural gas, while the VAR model fitted with other price information was the best for all products except natural gas. As of the residuals of the models, either with average price or other price information as target variable, are not autocorrelated and seem to follow a normal distribution. Nevertheless, the problem is that it tends to overfit with both models.

In the end, by comparing results of all 3 models, Hydrous Ethanol prediction has the best performance on Arima Garch model; Natural Gas, LPG and Diesel perform better on Arima with Fourier term model; and Regular Gasoline can give the best prediction by VAR.

Based on the overall analysis from ARIMA+Garch model, ARIMA with Fourier Terms, and VAR models. We have the following conclusions:

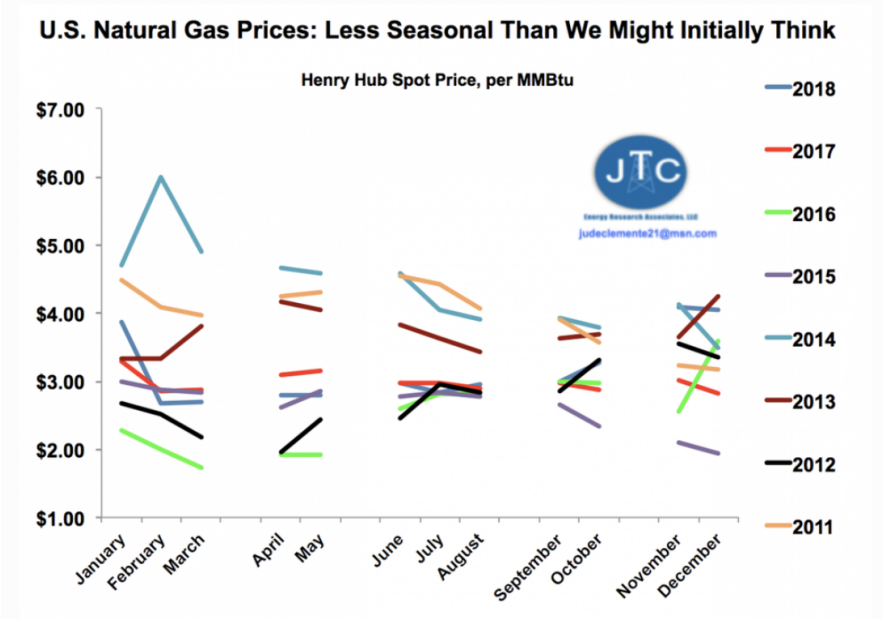
- Weak seasonality

  From our exploratory analysis on the five product prices, we conclude that each of the prices has weak seasonality, or long term seasonality. Therefore, the second model "ARIMA with Fourier Terms" returns relatively better results. As we all know, ARIMA models are often used for short-term-- monthly or quarterly data, so when the frequency is large, an ARIMA model with Fourier terms would give more accurate prediction of seasonal cycles.

  Does only gas prices in Brazil show weak seasonality? We made some further research on it. Let's take Natural Gas as an example here.

  For the U.S. natural gas market[6], technically, there are two seasons: Summer (April-Oct) and Winter (November-March). Gas is injected into the ground in Summer and gas is withdrawn in Winter to meet demand that rises well above production. According to the analysis article from Jude Clemente, when looking from 2011-2018, the seasonality of gas prices is not as apparent as one might think. As the graphic below shows, seasonal differences overall are not so clear.

[6] https://www.forbes.com/sites/judeclemente/2019/02/07/the-surprisingly-low-seasonality-of-u-s-natural-gas-prices/#44a3e7161f4a



For further study in the future, we may explore why these countries show weak seasonality even though gas prices are always considered as seasonal series.

- Stable upward trend
  In general, except for a few years, all products have a steady upward trend. In fact, because oil prices as a global market are deeply affected by inflation. According to the calculation of trading economics website[7], after we adjust the gas price by inflation, the increase in oil price is not large. And inflation is relatively stable data. We can see from our model that a stable trend is helpful for making a accurate time series prediction

Appendix

Variable	**Description**
start_date	Date of start the collection of data
end_date	Date of end the collection of data
region	Is a subdivision of Brazil. There are 5 Regions: Sul - South Sudeste - Southeast Centro-Oeste - Midwest Nordeste - Northeast

[7] https://tradingeconomics.com/brazil/gasoline-prices

	Norte - North.
state	Brazil has 26 states and one federal district
product	There are 6 types of fuel: Óleo diesel - Diesel Gasolina comum - Regular gasoline GLP - LPG Etanol hidratado - Hydrous Ethanol GNV - Natural gas Óleo Diesel S10 - Diesel S10
num_gas_stations	Number of gas stations consulted in the research
unit	Unit about each fuel in the research. I will explain better in the conversion topic.
avg_price	Average price of resale in each collection
sd_price	Standard deviation price of resale in each collection
min_price	Minimum price of resale in each collection
max_price	Maximum price of resale in each collection
avg_price_margin	Average of profit in release in reach collection
year	ANO or year: Year of each collection
month	MÊS or month: Month of each collection
coef_dist	Coefficient of variation about distribution of each collection
dist_max_price	Maximum price of distrition in each collection
dist_min_price	Minimum price of distrition in each collection
dist_sd_price	Standard deviation price of resale in each collection
dist_avg_price	Average price of distribution in each collection
coef_price	coefficient of variation about resale of each collection