IE 360 • Statistical Forecasting and Time Series

Class Project

Report

Tevfik Buğra Türker - 2019402120
Ömercan Mısırlıoğlu - 2020402261
Hüseyin Emre Bacak - 2021402279
Bora Polater - 2021402294

22.05.2023

TABLE OF CONTENTS

1. Describing the Observations

1.1. Time Series Observations

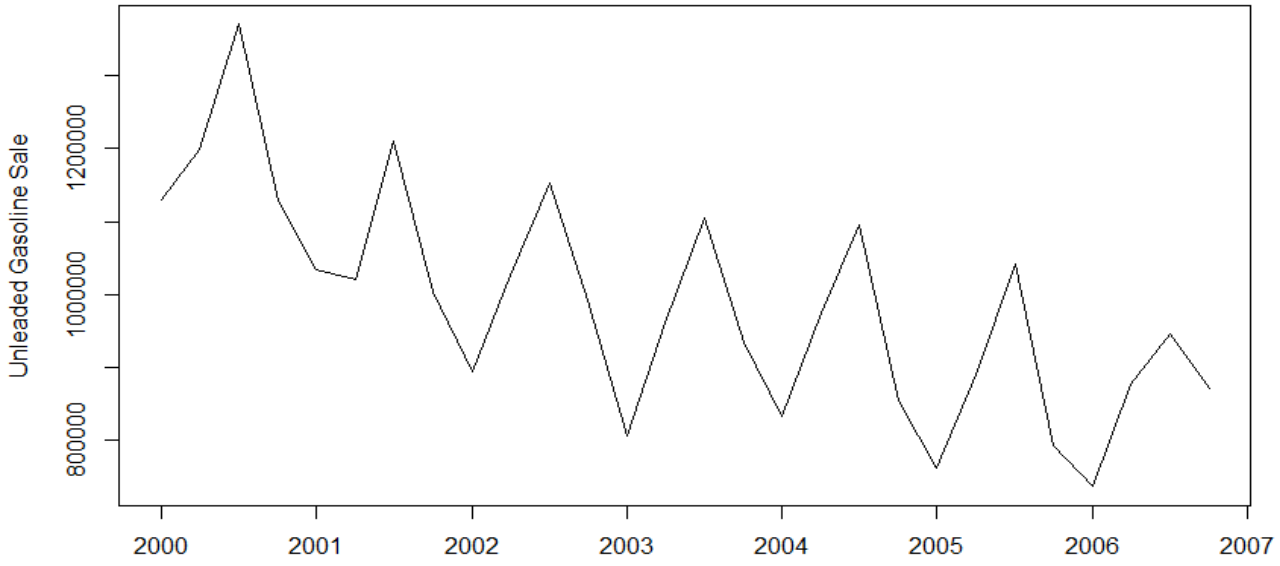


Figure 1.1.1 - Time series plot for Unleaded Gasoline Sale (UGS)

The gradual decline in the variable over the observed period indicates decreasing trend, suggesting a potential underlying shift. The presence of recurring patterns and fluctuations within regular intervals suggests there is seasonality in UGS.

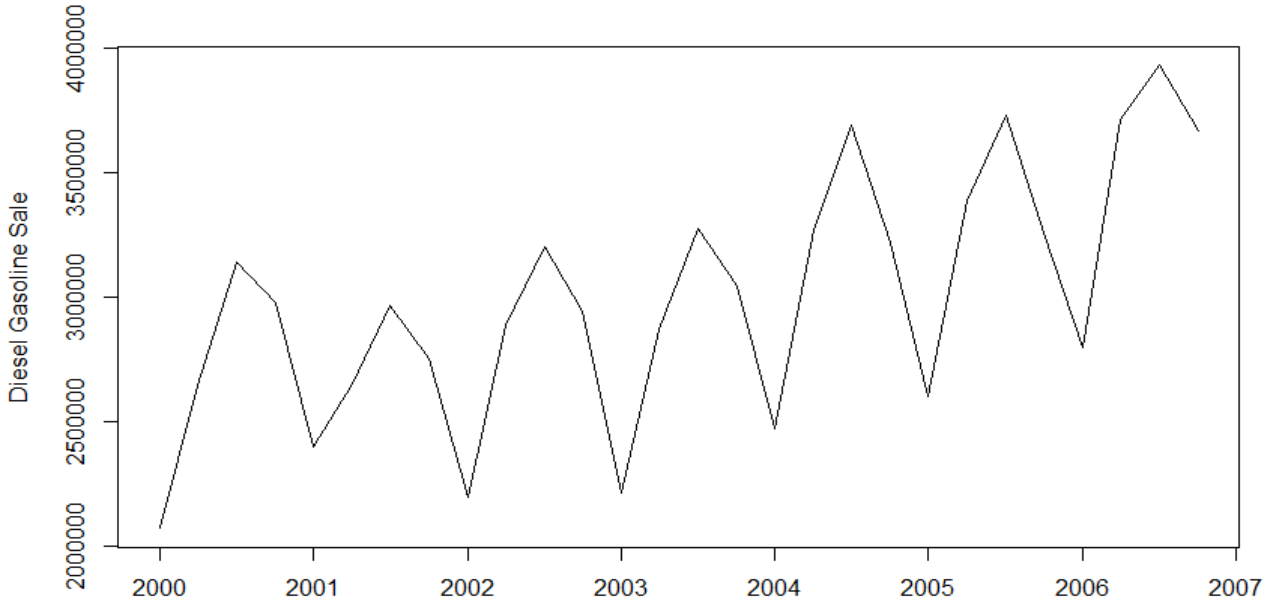


Figure 1.1.2 - Time series plot for Diesel Gasoline Sale (DGS)

The gradual rise in the variable over the observed period indicates increasing trend, suggesting a potential underlying growth or upward shift. Additionally, the presence of recurring patterns and fluctuations reveals seasonality in DGS.

1.2 Autocorrelation Functions Observations

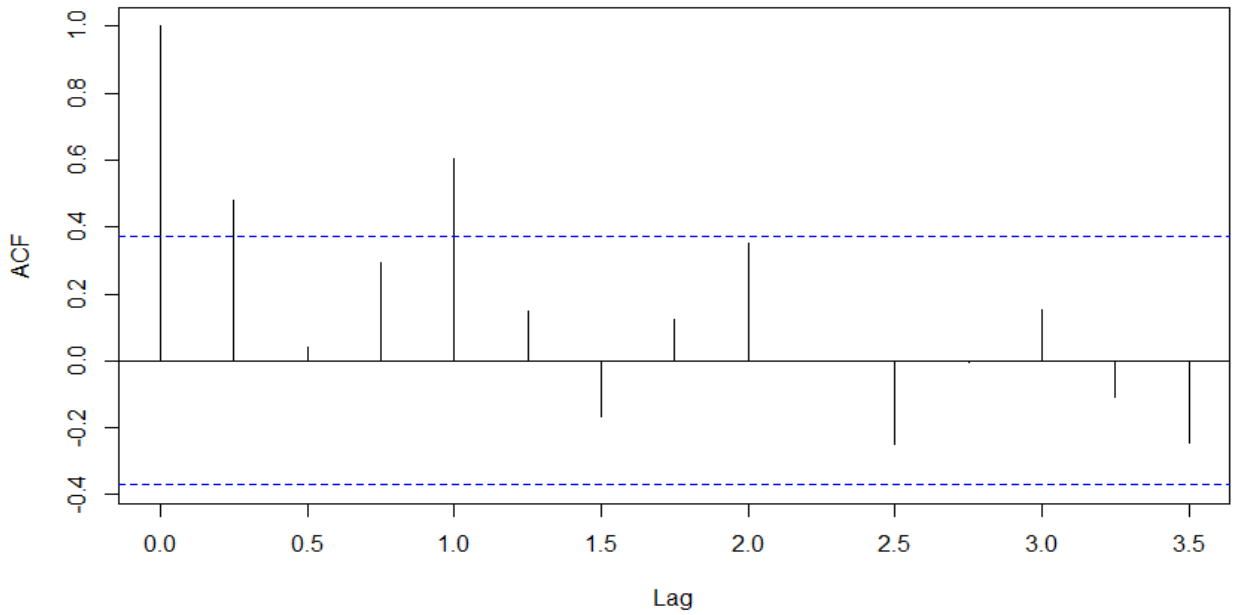


Figure 1.2.1 - ACF plot for Unleaded Gasoline Sale (UGS)

The ACF plot reveals spikes at lag 4, 8, 12, and so on. These spikes indicate significant correlations between the variable and its lagged values at these specific intervals and presence of seasonality in the UGS data.

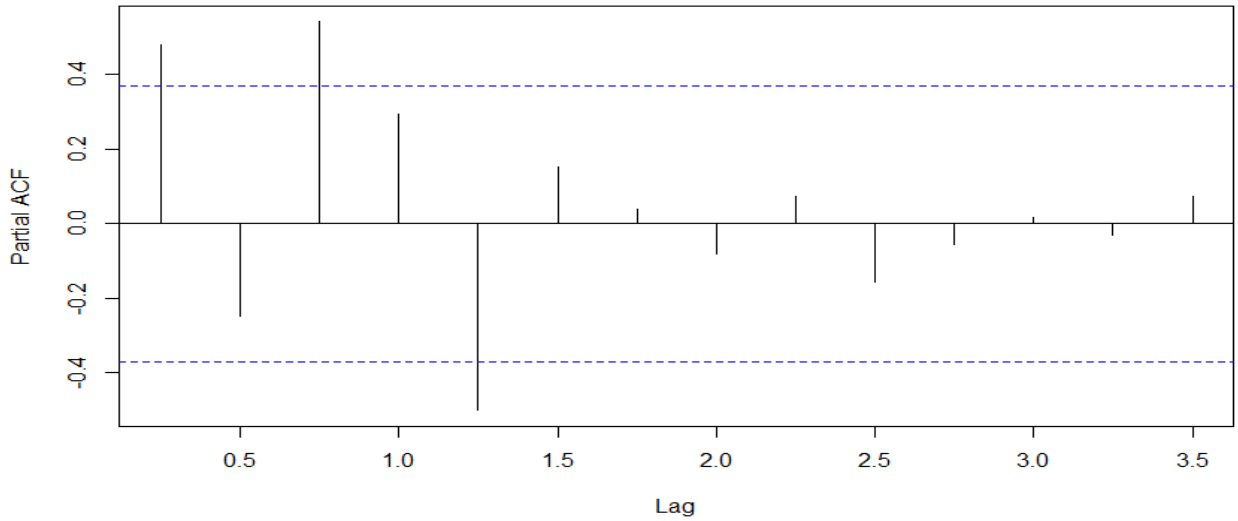


Figure 1.2.2 - PACF plot for Unleaded Gasoline Sale (UGS)

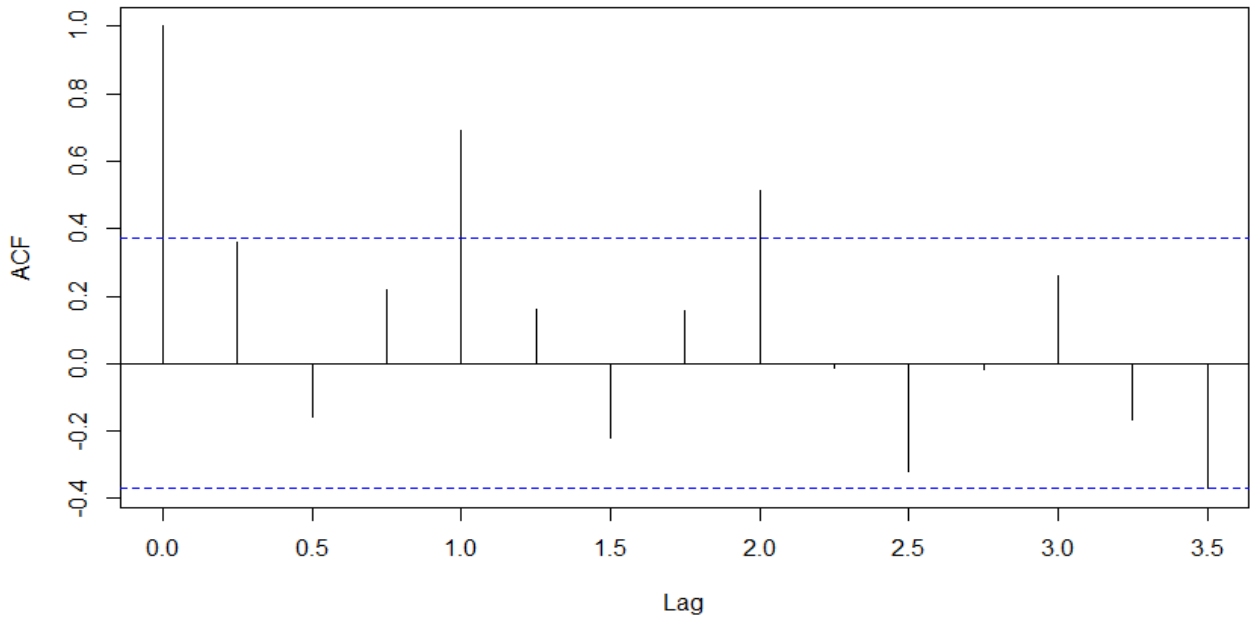


Figure 1.2.3 - ACF plot for Diesel Gasoline Sale (DGS)

The ACF plot reveals spikes at lag 4, 8, 12, and so on. These spikes indicate significant correlations between the variable and its lagged values at these specific intervals and presence of seasonality for the DGS just like the UGS data.

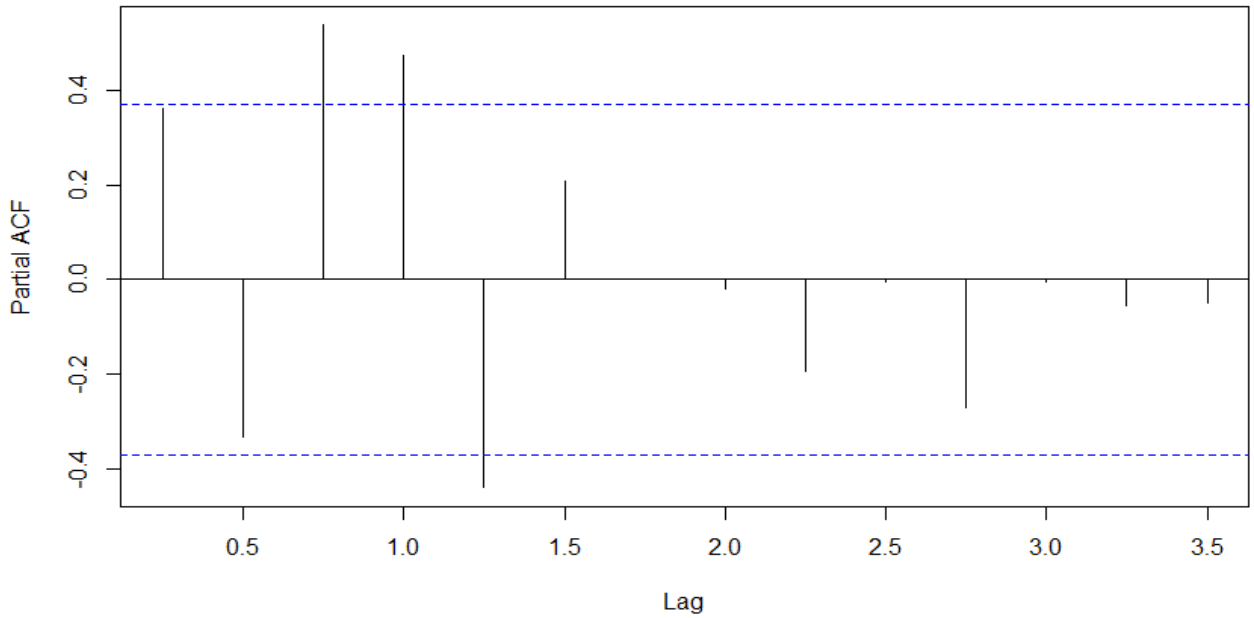


Figure 1.2.4 - PACF plot for Diesel Gasoline Sale (DGS)

2. Method A: Forecasting with Time Series Analysis

2.1. Checking for Preliminary Transformation to Induce Stationarity

As discussed in Section 1.1, presence of both seasonality and trend in both UGS and DGS datasets can be seen from Figure 1.1.1 and 1.1.2. Also, the presence of seasonality can be seen in ACF plots of DGS and UGS as discussed in Section 1.2 (Figure 1.2.1 and 1.2.3)

2.2. Utilizing the Time Series

Initially, in order to eliminate the trend component, non-seasonal difference was applied to both datasets. This involves taking the difference between consecutive observations, effectively removing any underlying trend present in the data. This differencing technique helped in achieving a stationary series, allowing for a more accurate analysis of the remaining seasonal patterns.

Taking non-seasonal difference for the UGS:

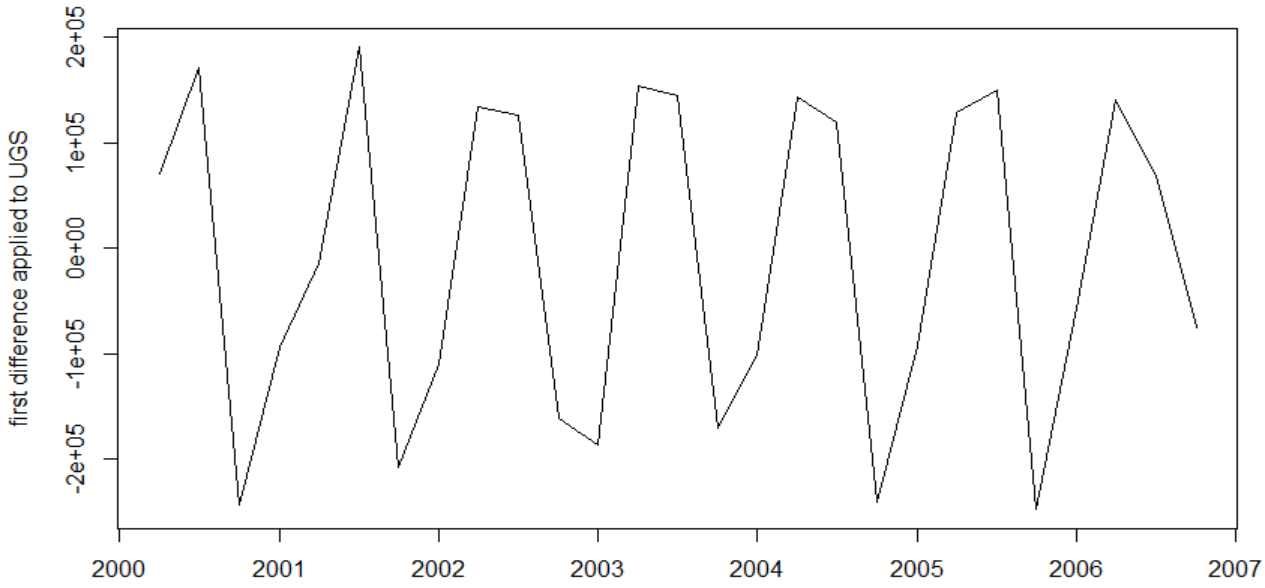


Figure 2.2.1 – Time Series plot for Unleaded Gasoline Sale (UGS) after taking non-seasonal difference

After taking the first non-seasonal difference, by looking at the Figure 2.2.1 stationary series achieved, yet seasonality still remains in the UGS data.

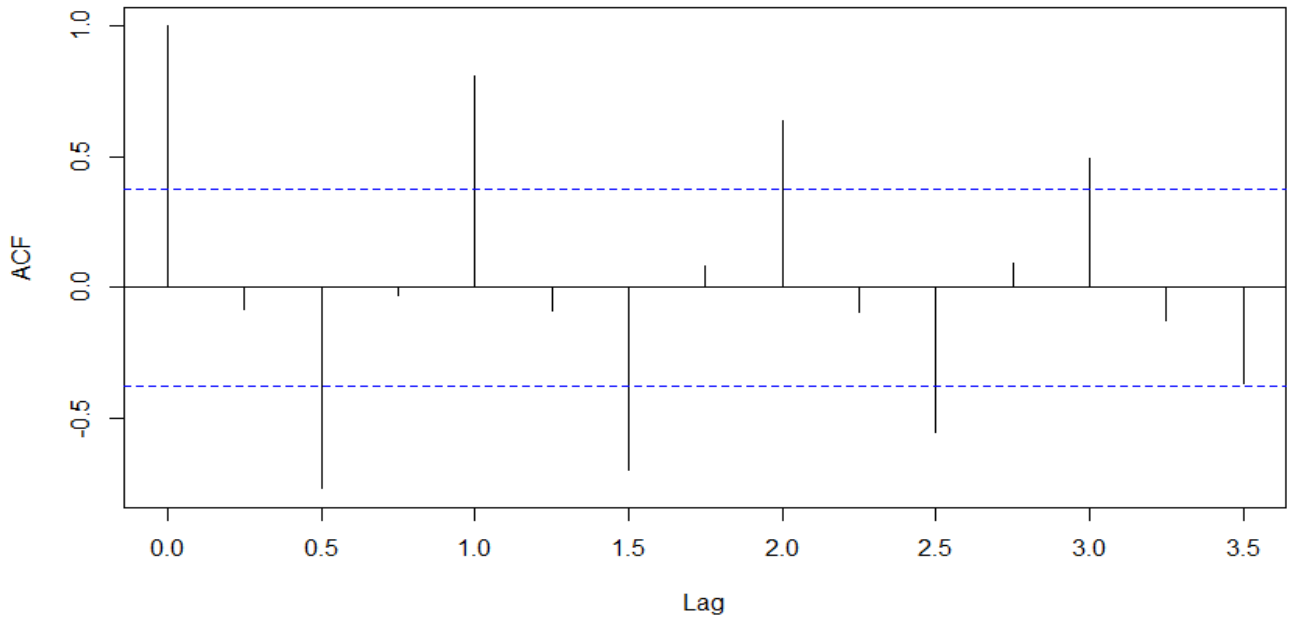


Figure 2.2.2 – ACF plot for Unleaded Gasoline Sale (UGS) after taking non-seasonal difference

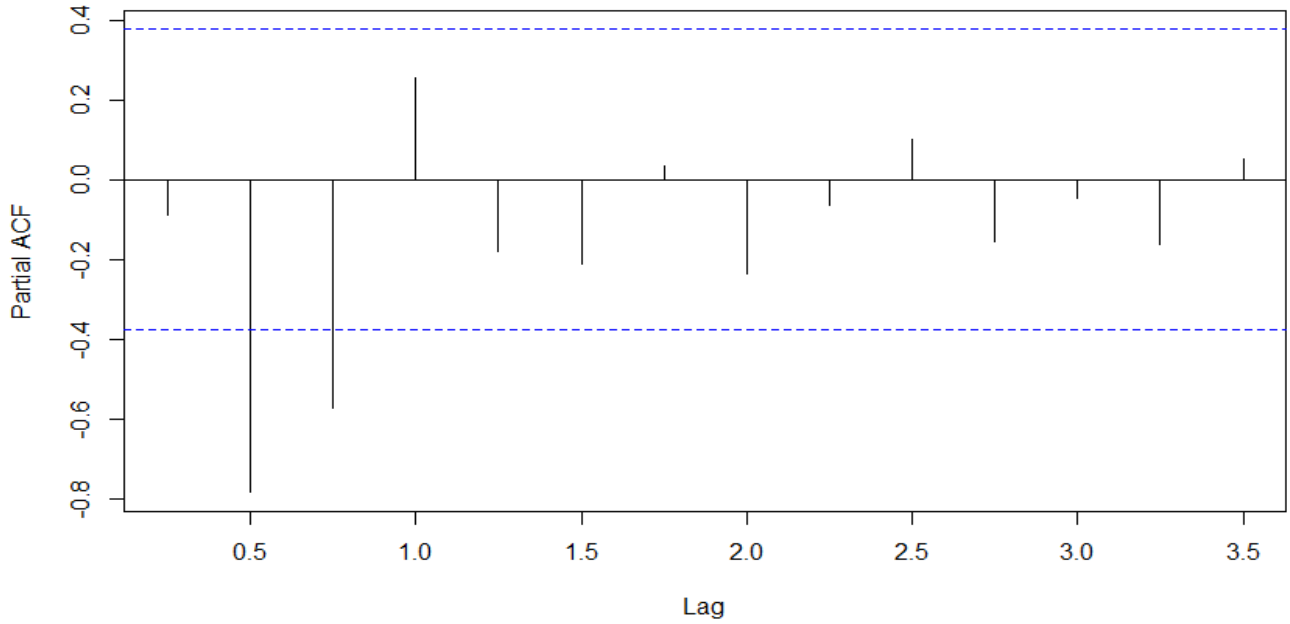


Figure 2.2.3 – PACF plot for Unleaded Gasoline Sale (UGS) after taking non-seasonal difference

By looking at the ACF plot of UGS after taking the non-seasonal difference (Figure 2.2.3), seasonality effect still can be seen. Thus, a seasonal difference should be applied to data with order=4.

Taking non-seasonal difference for the DGS:

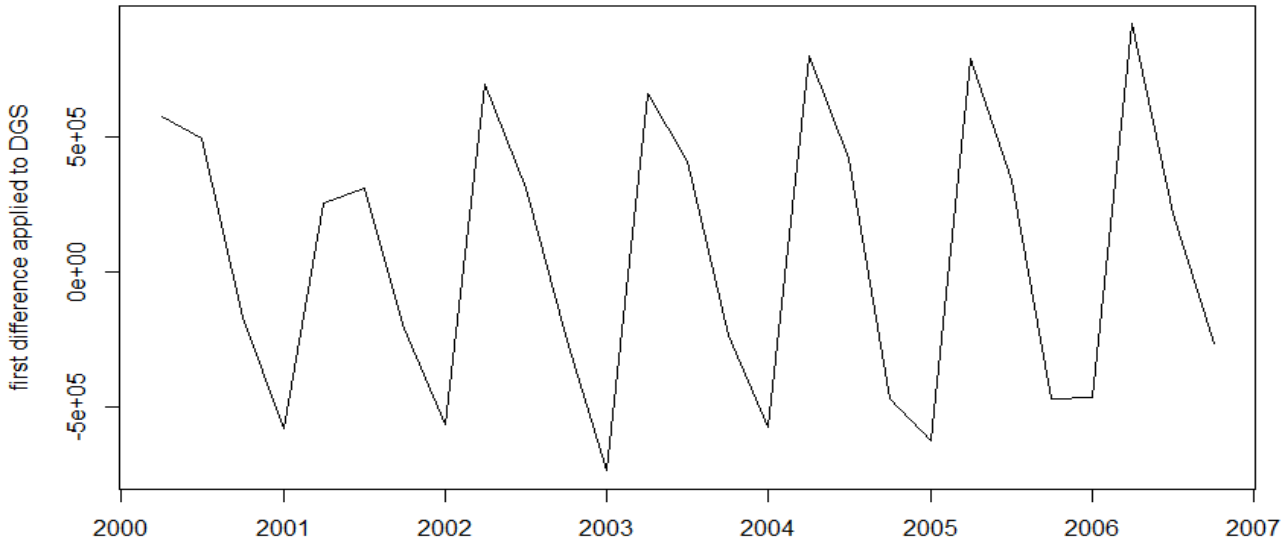


Figure 2.2.4 – Time Series plot for Diesel Gasoline Sale (DGS) after taking non-seasonal difference

Similarly, after taking the first non-seasonal difference, by looking at the Figure 2.2.4 stationary series achieved, yet seasonality still remains in the DGS data.

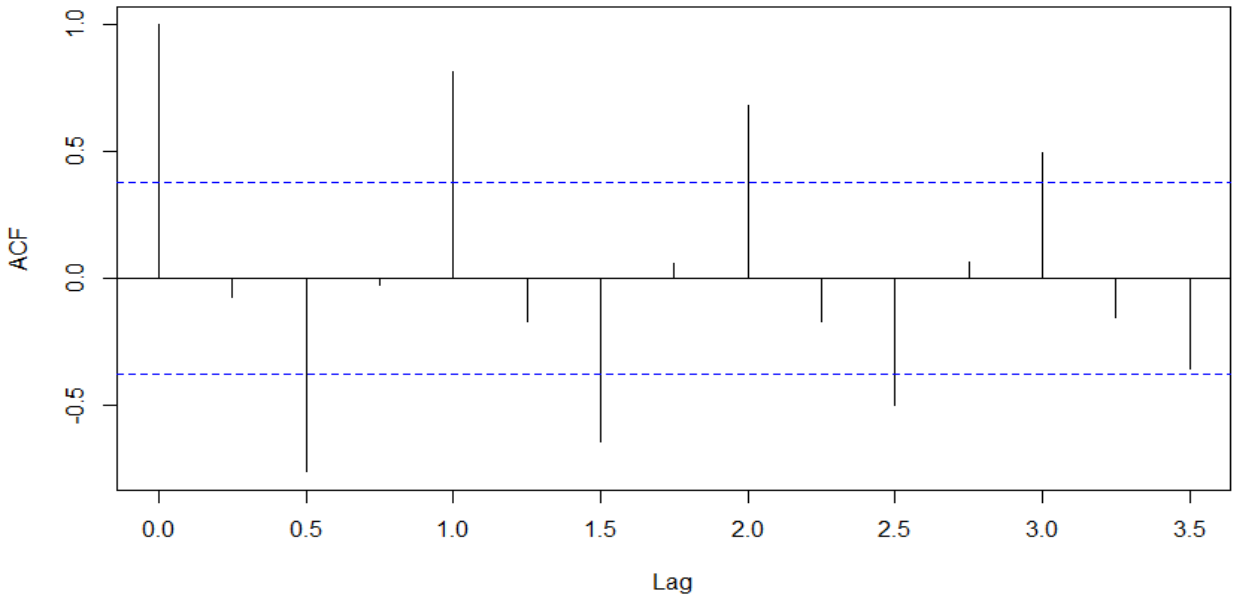


Figure 2.2.5 – ACF plot for Diesel Gasoline Sale (DGS) after taking non-seasonal difference

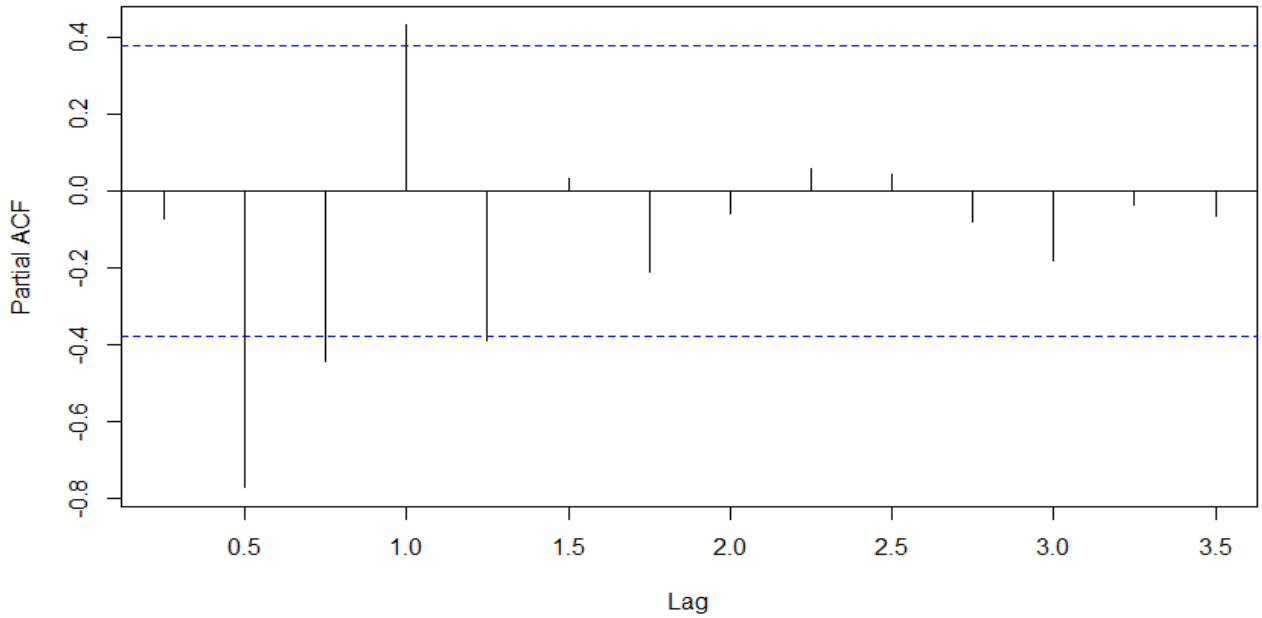


Figure 2.2.6 – PACF plot for Diesel Gasoline Sale (DGS) after taking non-seasonal difference

Similarly, by looking at the ACF plot of DGS after taking the non-seasonal difference (Figure 2.2.6), seasonality effect still can be seen. Thus, a seasonal difference should be applied to data with order=4.

Taking seasonal difference for the UGS:

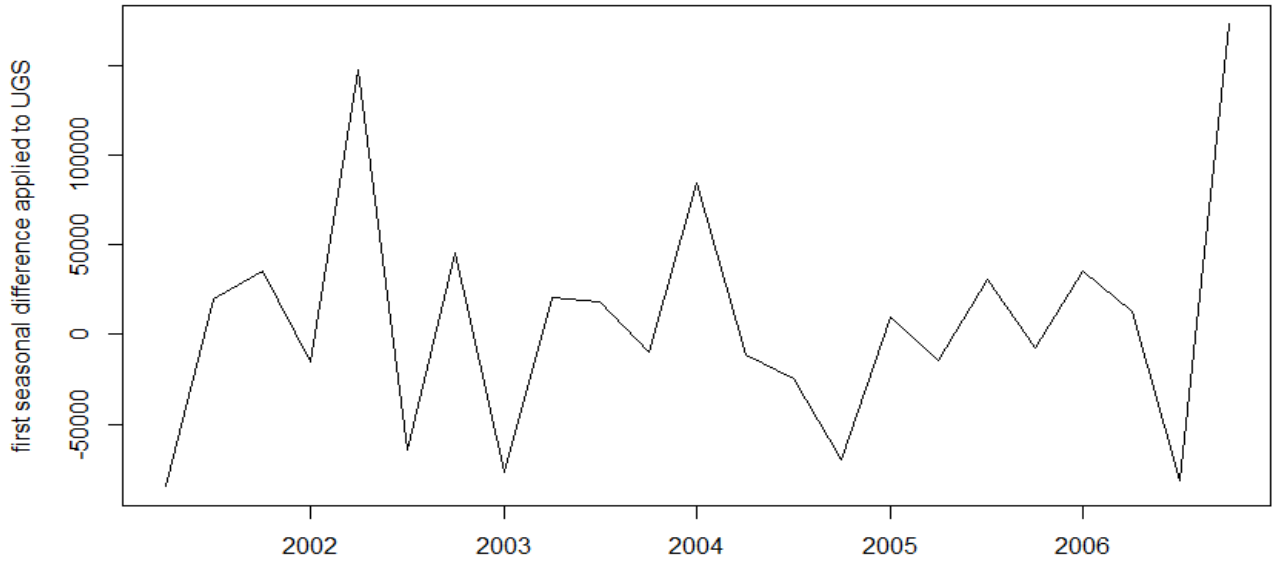


Figure 2.2.7 – Time Series plot for Unleaded Gasoline Sale (UGS) after taking seasonal difference

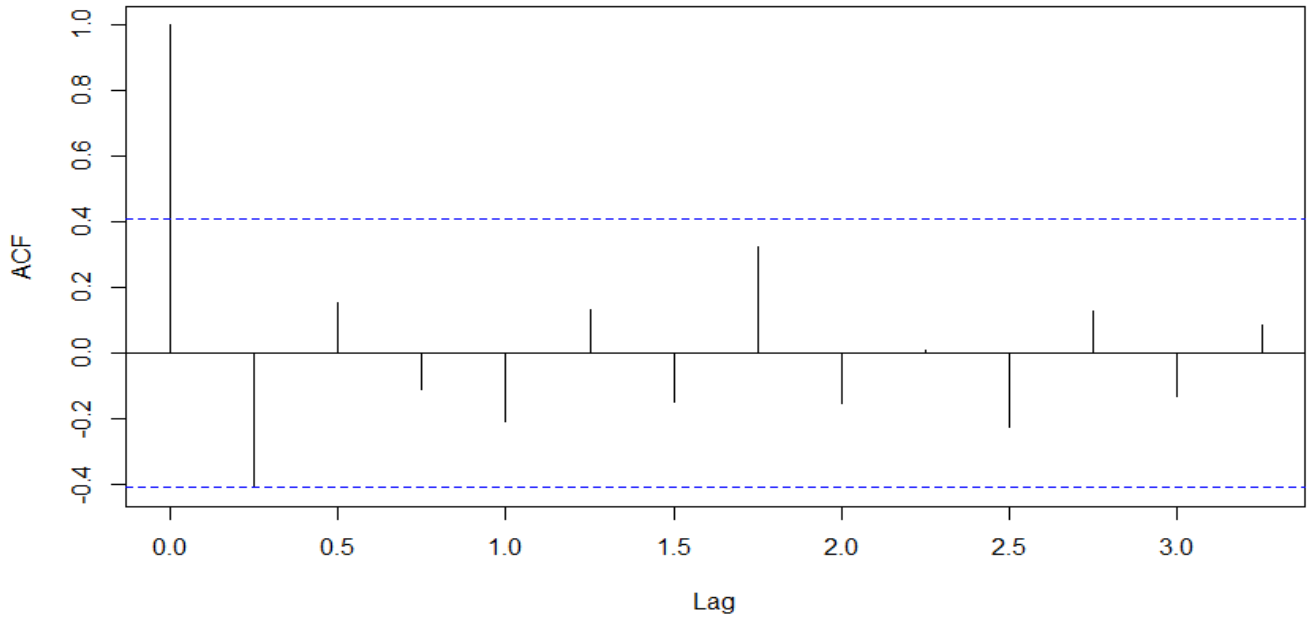


Figure 2.2.8 – ACF plot for Unleaded Gasoline Sale (UGS) after taking seasonal difference

For non-seasonal periods, ACF plot shows a significant cut-off after lag 1 for the non-seasonal differences. This suggests a non-seasonal Moving Average (MA) component in the ARIMA model, specifically a non-seasonal MA(1) can be used.

For seasonal periods, ACF plot for the seasonal differences exhibits a pattern that dies out to zero. This pattern suggests a seasonal Autoregressive (AR) component in the ARIMA model. Specifically, a seasonal AR(1) can be used.

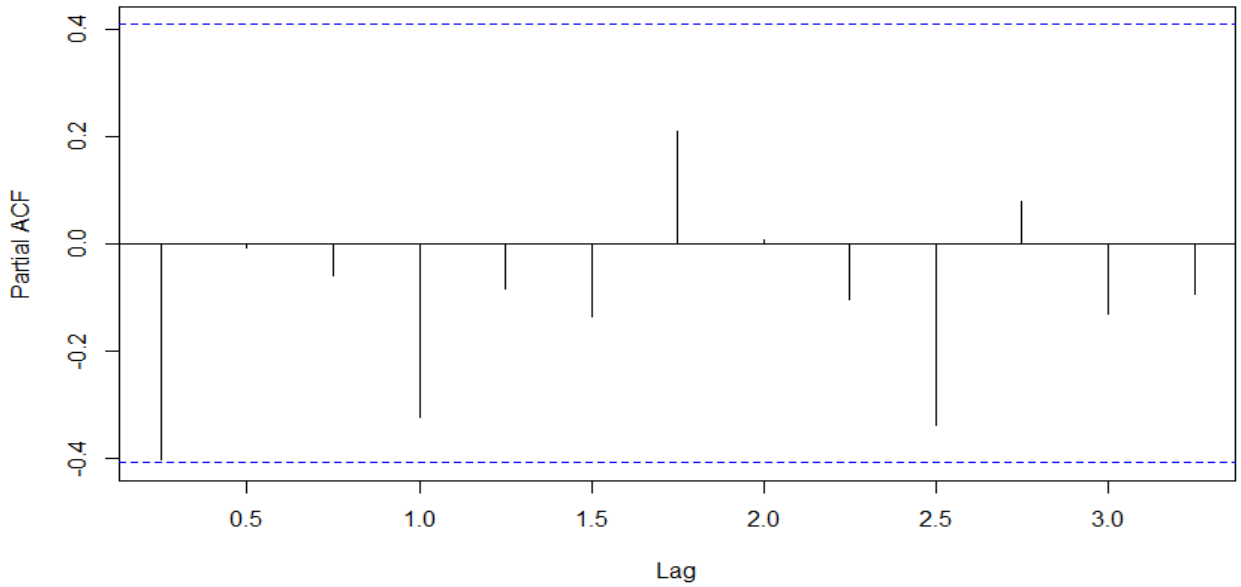


Figure 2.2.9 – PACF plot for Unleaded Gasoline Sale (UGS) after taking seasonal difference

Taking seasonal difference for the DGS:

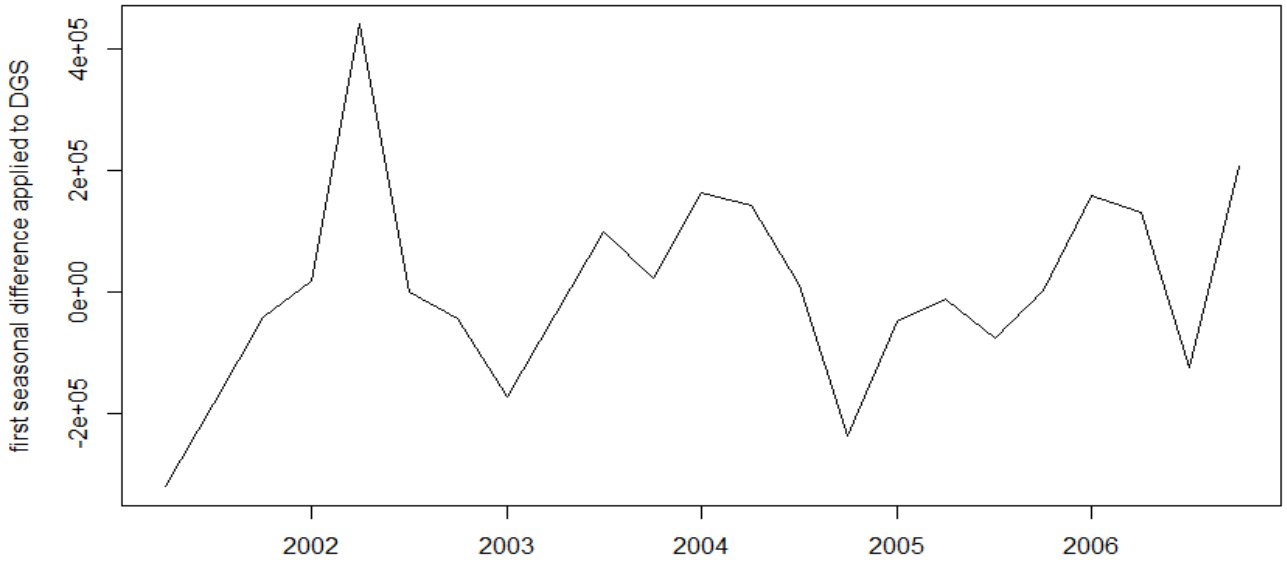


Figure 2.2.10 – Time Series plot for Diesel Gasoline Sale (DGS) after taking seasonal difference

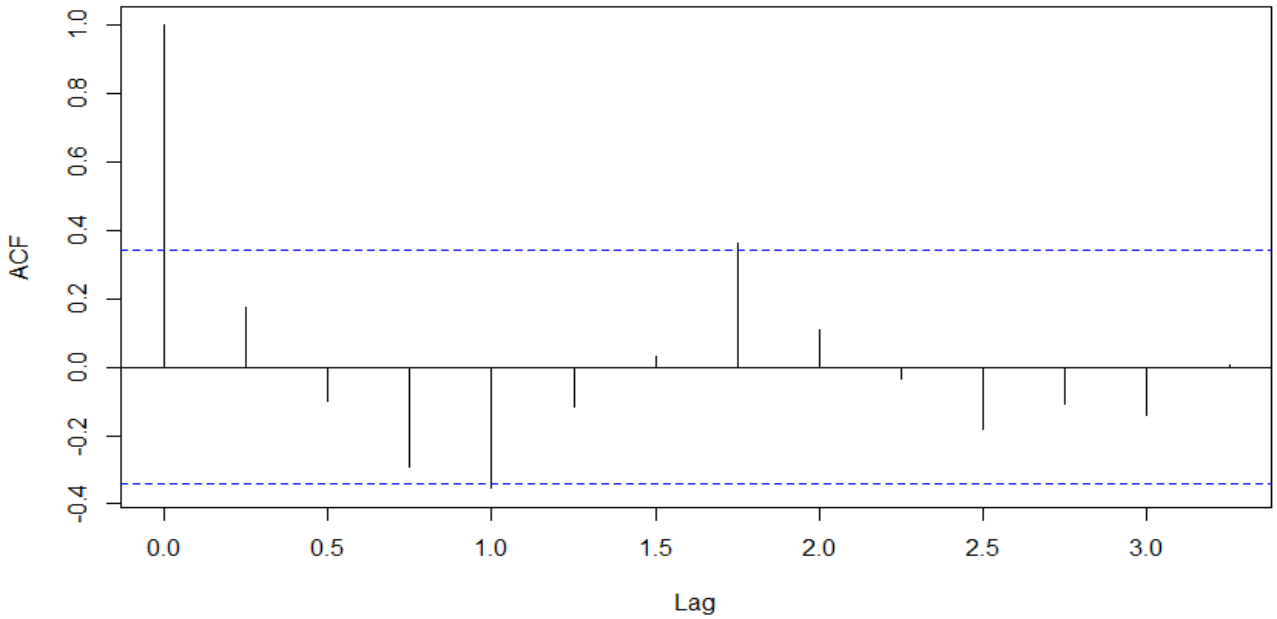


Figure 2.2.11 – ACF plot for Diesel Gasoline Sale (DGS) after taking seasonal difference

For non-seasonal periods, ACF plot (Figure 2.2.11) reveals that the autocorrelation gradually dies out to zero for the non-seasonal differences. This indicates Autoregressive (AR) component in the ARIMA model. Specifically, a non-seasonal AR(1) term can be considered to capture the residual autocorrelation beyond lag 1.

For seasonal periods, ACF plot (Figure 2.2.11) for the seasonal differences shows a significant cut-off after lag 4. This suggests need for a seasonal Moving Average (MA) component in the ARIMA model. Specifically, a seasonal MA(1) term can be considered to capture the residual autocorrelation at the seasonal lag.

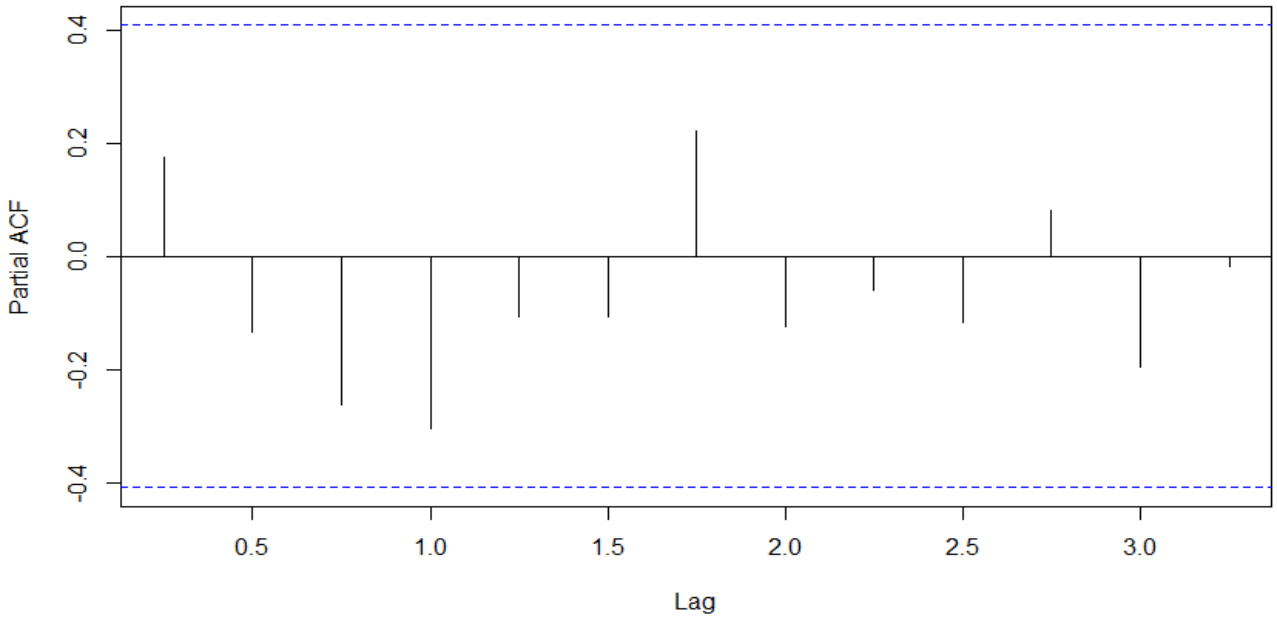


Figure 2.2.12 – PACF plot for Diesel Gasoline Sale (DGS) after taking seasonal difference

2.3. Initial ARIMA Model

For UGS:

For the initial ARIMA model based on inspections made in section 2.1 and 2.2, a seasonal autoregressive integrated moving average (SARIMA) model with parameters (0,1,1)(1,1,0) was performed. Output of initial model and corresponding ACF and PACF plots are:

```
AIC=572.8    AICc=574.07    BIC=576.21
```

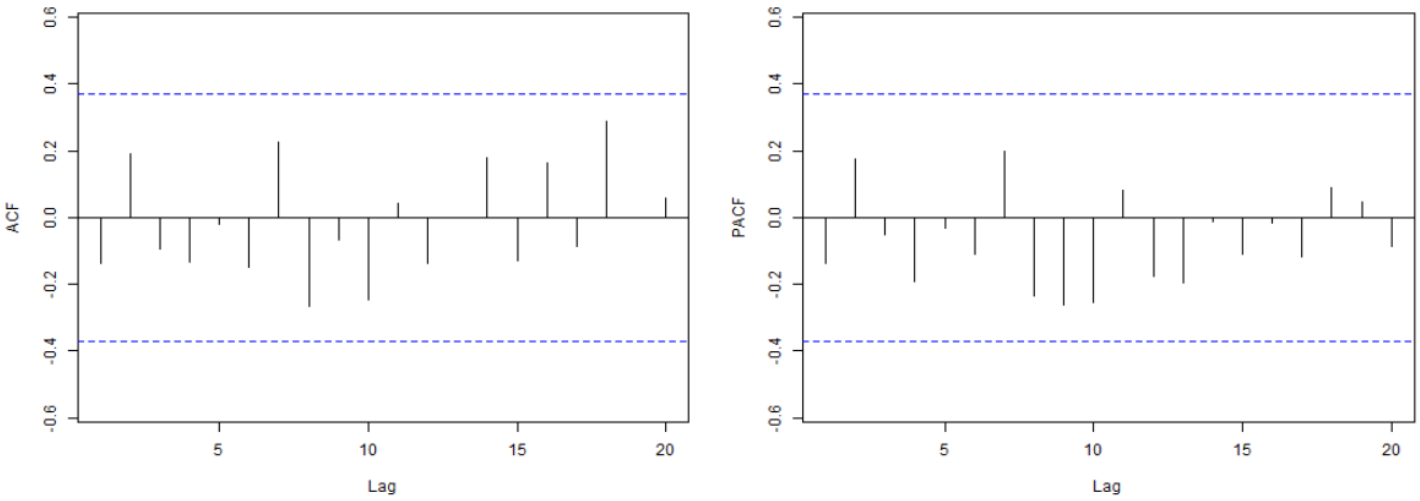


Figure 2.3.1 – ACF and PACF plots of initial ARIMA model for UGS

For DGS:

For the initial ARIMA model based on inspections made in section 2.1 and 2.2, a seasonal autoregressive integrated moving average (SARIMA) model with parameters (1,1,0)(0,1,1) was performed. Output of initial model and corresponding ACF and PACF plots are:

```
AIC=618.75   AICc=620.02   BIC=622.16
```

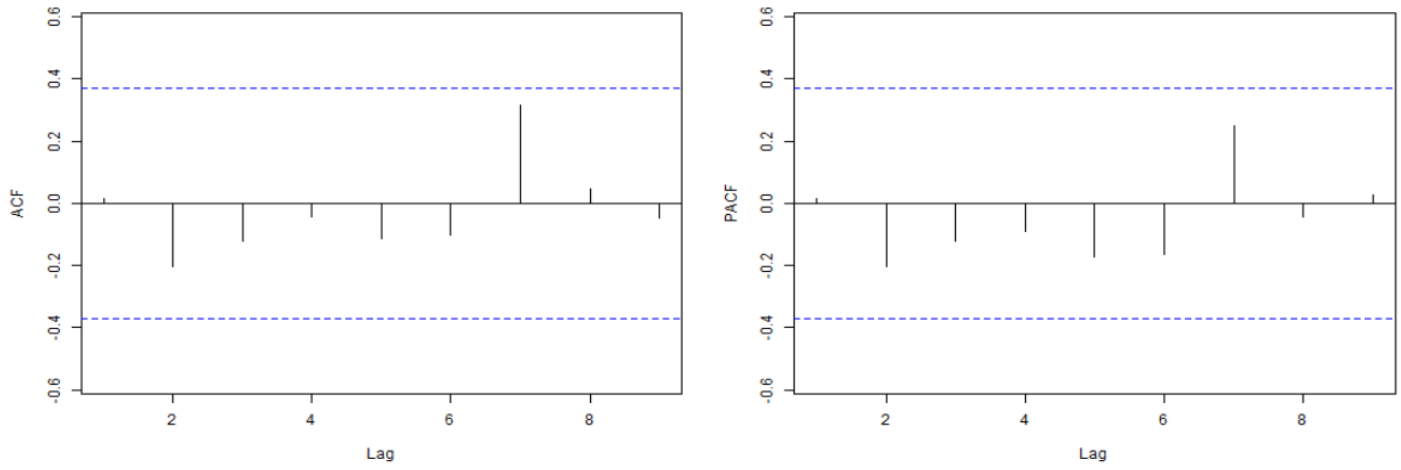


Figure 2.3.2 – ACF and PACF plots of initial ARIMA model for DGS

2.4. Neighborhood Search of Initial Model

For UGS:
For neighborhood search, six models are used for UGS including the initial model given in section 2.3.

For the second model, the number of regular AR coefficients are increased by 1. Resulting model is SARIMA (1,1,1)(1,1,0), which performed worse than the initial model. Output of the second model and corresponding ACF and PACF plots are:

```
AIC=572.22    AICc=574.44     BIC=576.76
```

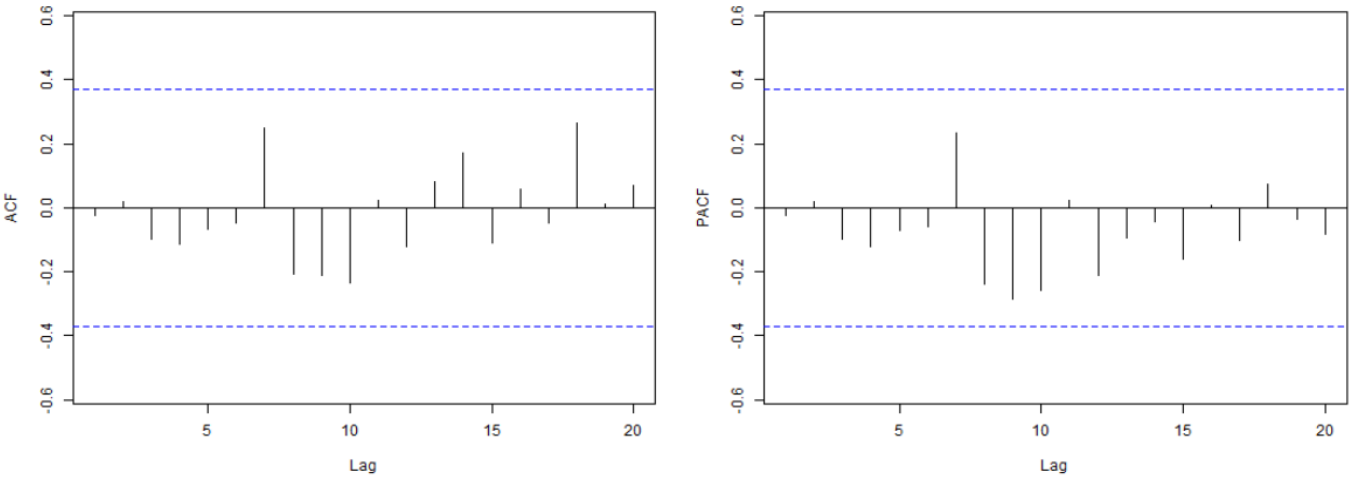
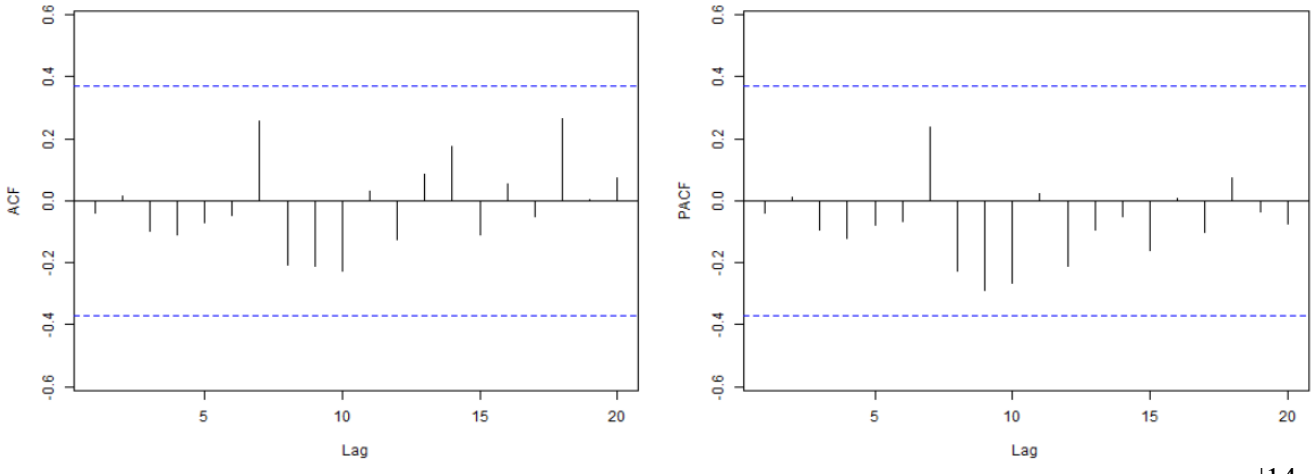


Figure 2.4.1 – ACF and PACF plots of second model for UGS

For the third model, the number of regular MA coefficients are increased by 1. Resulting model is SARIMA (1,1,0)(1,1,0), which performed better than the initial model. Output of the third model and corresponding ACF and PACF plots are:

```
AIC=570.24    AICc=571.5      BIC=573.64
```



For the fourth model, the number of regular AR coefficients are increased by 1 compared to the third model. Resulting model is SARIMA (1,1,0)(2,1,0), which performed worse than the third model but better than the first model. Output of the fourth model and corresponding ACF and PACF plots are:

```
AIC=571.95    AICc=574.17     BIC=576.49
```

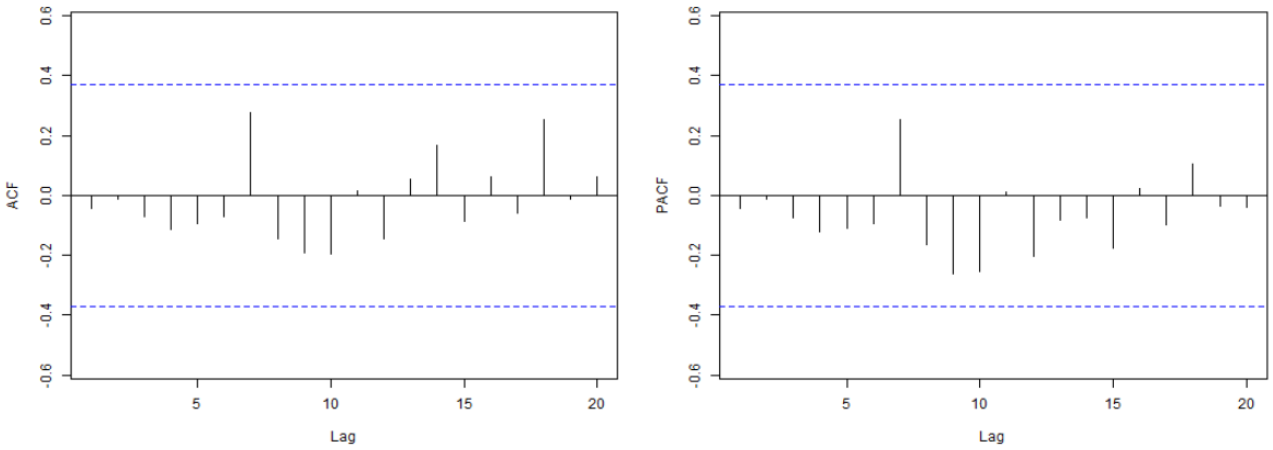


Figure 2.4.3 – ACF and PACF plots of fourth model for UGS

For the fifth model, the number of regular AR coefficients are decreased by 1 and seasonal MA coefficients increased by 1. Resulting model is SARIMA (1,1,0)(0,1,1), which performed better than the third model. Output of the fifth model and corresponding ACF and PACF plots are:

```
AIC=569.53    AICc=570.79     BIC=572.93
```

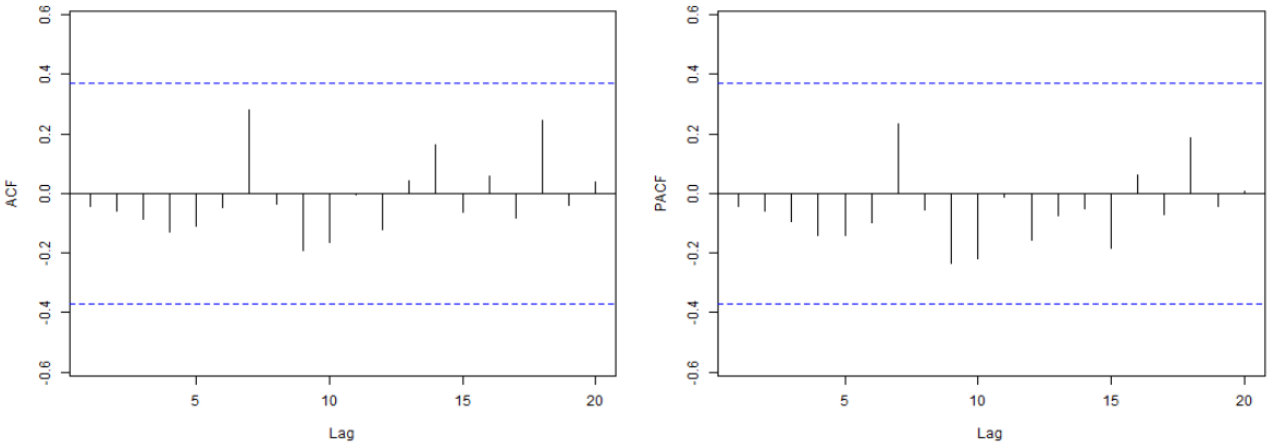


Figure 2.4.4 – ACF and PACF plots of fifth model for UGS

For the sixth model, the number of seasonal MA coefficients are increased by 1. Resulting model is SARIMA (1,1,0)(0,1,2), which performed worse than the fifth model. Output of the fifth model and corresponding ACF and PACF plots are:

```
AIC=571.49    AICc=573.71    BIC=576.03
```

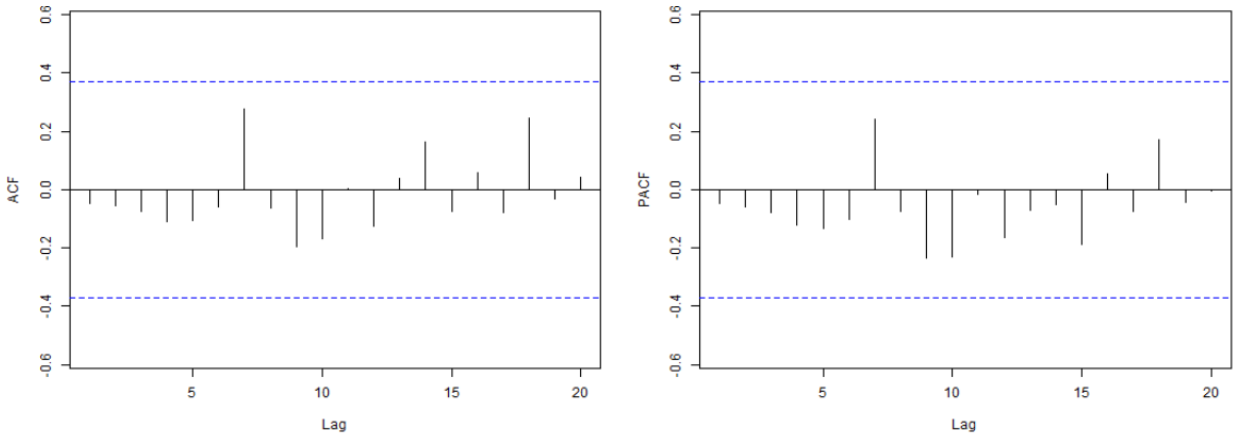


Figure 2.4.5 – ACF and PACF plots of sixth model for UGS

Finally, the auto.arima function is used for finding a suitable model for UGS. Resulting auto.arima model is SARIMA(1,1,0)(0,1,1) with:
`AIC=569.53    AICc=570.79    BIC=572.93.`

Comparing all six models and the auto.arima model according to their AIC, AICc and BIC values; the best model to forecast future prices for UGS is **Model 5** (which is the same model suggested by auto.arima) since it has the lowest values for these criteria.

For DGS:
For neighborhood search, six models are used for DGS including the initial model given in section 2.3.

For the second model, the number of regular MA coefficients are increased by 1 compared to the initial model. Resulting model is SARIMA (1,1,1)(1,1,0), which performed worse than the initial model. Output of the second model and corresponding ACF and PACF plots are:

```
AIC=618.96    AICc=621.19    BIC=623.51
```

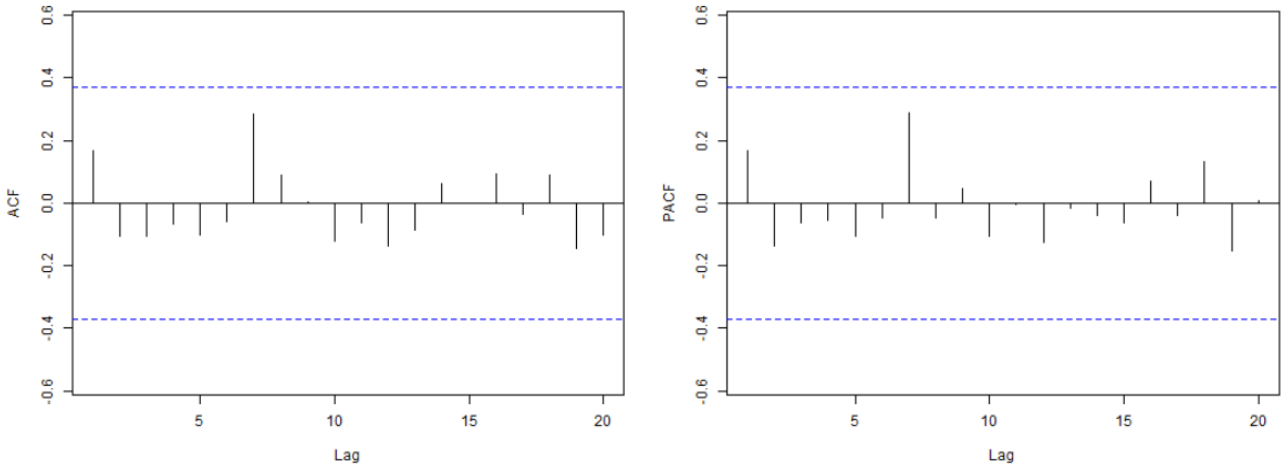


Figure 2.4.6 – ACF and PACF plots of the second model for DGS

For the third model, the number of seasonal AR coefficients are increased by 1 and seasonal MA coefficients are decreased by 1 compared to initial model. Resulting model is SARIMA (1,1,0)(1,1,0), which performed better than the initial model. Output of the third model and corresponding ACF and PACF plots are:

```
AIC=618.22   AICc=619.49   BIC=621.63
```

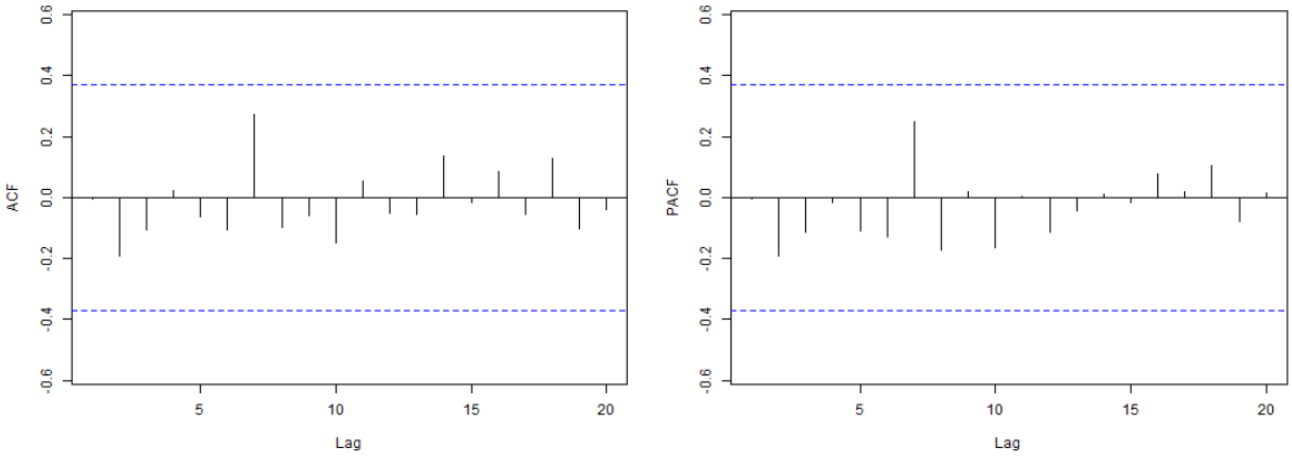


Figure 2.4.7 – ACF and PACF plots of the third model for DGS

For the fourth model, the number of regular AR coefficients are decreased by 1. Resulting model is SARIMA (0,1,0)(1,1,0), which performed better than the third model. Output of the fourth model and corresponding ACF and PACF plots are:

```
AIC=616.23    AICc=616.83    BIC=618.5
```

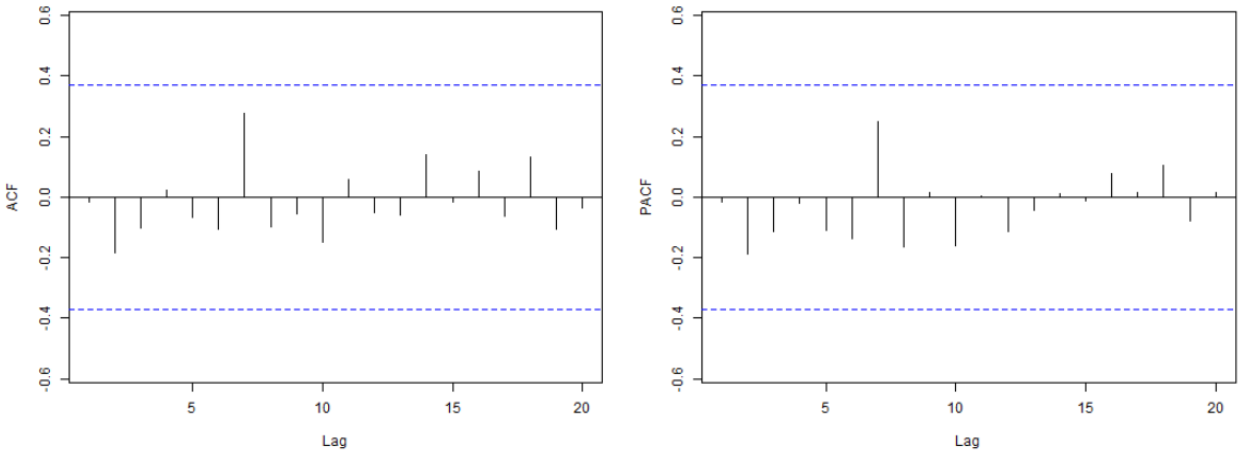


Figure 2.4.8 – ACF and PACF plots of the fourth model for DGS

For the fifth model, the number of regular MA coefficients are increased by 1 compared to the fourth model. Resulting model is SARIMA (0,1,1)(1,1,0), which performed worse than the fourth model. Output of the fifth model and corresponding ACF and PACF plots are:

```
AIC=618.21    AICc=619.47    BIC=621.62
```

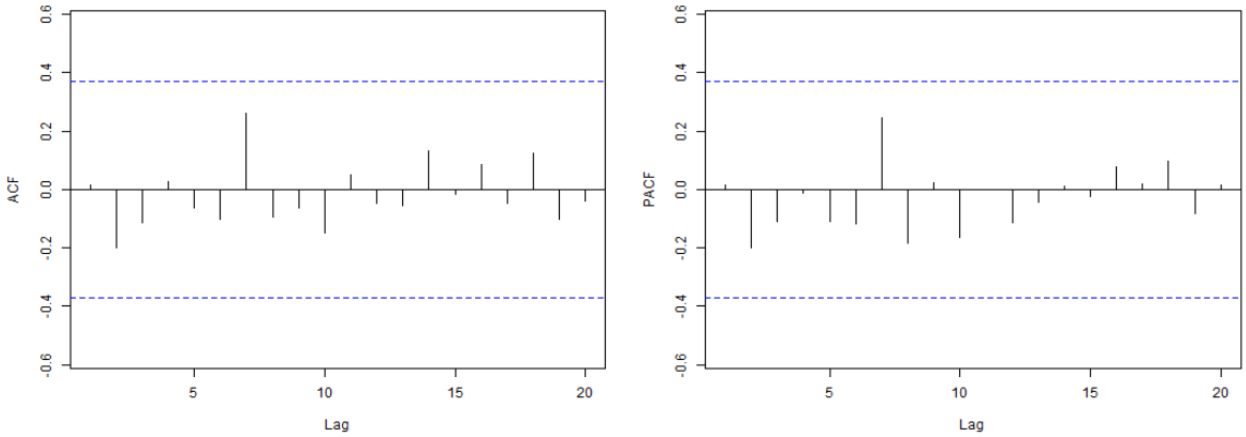


Figure 2.4.9 – ACF and PACF plots of the fifth model for DGS

For the sixth model, the number of seasonal MA coefficients are increased by 1 compared to the fourth model. Resulting model is SARIMA (0,1,0)(1,1,1), which performed worse than the fourth model. Output of the sixth model and corresponding ACF and PACF plots are:

```
AIC=618.17   AICc=619.43   BIC=621.57
```

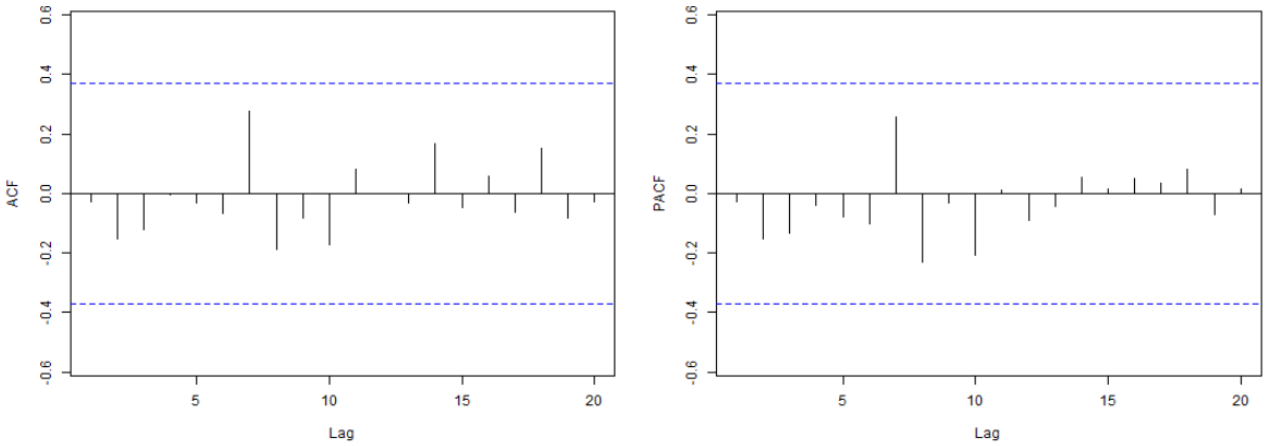


Figure 2.4.10 – ACF and PACF plots of the sixth model for DGS

Finally, the auto.arima function is used for finding a suitable model for DGS. Resulting auto.arima model is SARIMA(1,0,0)(1,1,0) with:

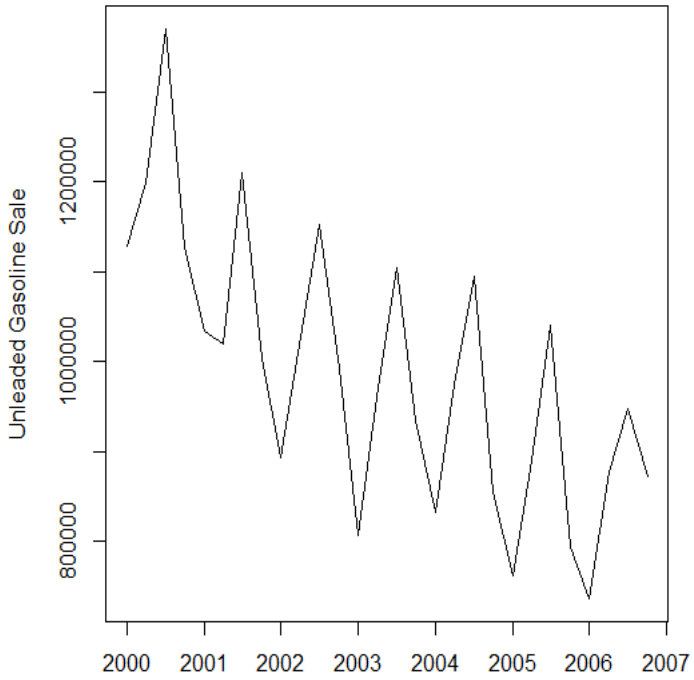
```
AIC=641.62   AICc=643.72   BIC=646.3
```

which performed worse than all of the models created. This also indicates that regular differencing is very important for DGS.

Comparing all six models and the auto.arima model according to their AIC, AICc and BIC values; the best model to forecast future prices for DGS is **Model 4** since it has the lowest values for these criteria.

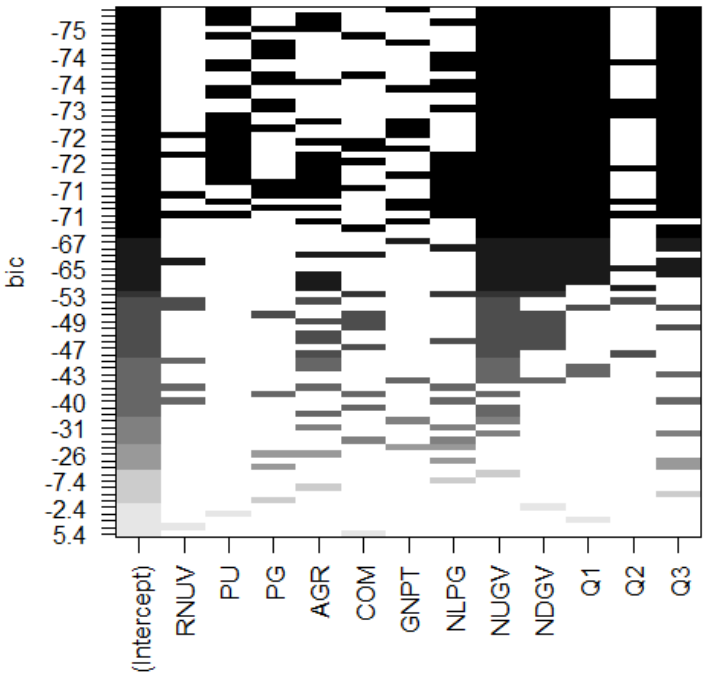
3. Method B: Forecasting with Regression

For Unleaded Gasoline Sale(UGS):



On the left side (Figure 3.1), "*Unleaded Gasoline Sales*" are given. When this data is analyzed, the variance seems to be steady. Thus, there is no need to take the logarithm of the data. Also there seems to be a linear decreasing trend. So, to fit a linear regression model, we do not need to take further differences. To take care of quarterly seasonality, 3 new variables are added which are "Q1, Q2 and Q3". The variable Q1 takes value 1 if the data is collected at Q1 and 0 otherwise. Q2 and Q3 are also defined in a similar manner.

Figure 3.1 – UGS Data



To find the best regression model different fits are sorted according to their Bayesian Information Criterion (BIC). The lowest BIC value model is considered to be the best fit. In our case the best fit has these independent variables:

- **PU:** Price of Unleaded Gasoline
- **GNPT:** GNP Total
- **NUGV:** # of Unleaded Gasoline Vehicles
- **NDGV:** # of Diesel Gasoline Vehicles
- **Q1** and **Q3**

Figure 3.2 – Variables and BIC Values

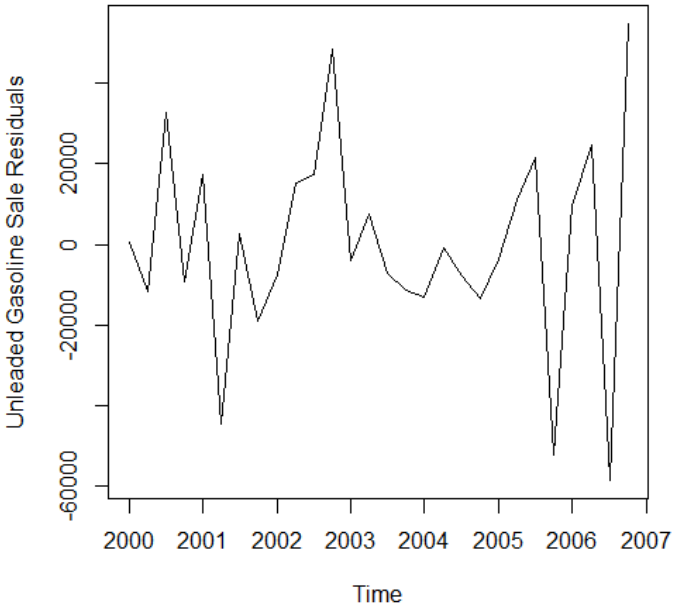
Validity of the Fit:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.488e+06  1.158e+05  21.489 8.88e-16 ***
PU          -5.185e+02  1.580e+02  -3.282  0.00355 **
GNPT        -7.772e-03  3.594e-03  -2.163  0.04226 *
NUGV        -6.628e-01  6.506e-02 -10.188 1.40e-09 ***
NDGV         8.007e+03  1.020e+03   7.849 1.12e-07 ***
Q1          -1.536e+05  2.033e+04  -7.553 2.05e-07 ***
Q3           2.245e+05  2.845e+04   7.891 1.03e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 29220 on 21 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.9709,  Adjusted R-squared:  0.9626
F-statistic: 116.7 on 6 and 21 DF,  p-value: 5.09e-15
```

Comments:

- All the p-values of coefficients are smaller than 0.05. Thus, they are all significant.
- Multiple R-squared (0.9709) and Adjusted R-squared (0.9626) values are high. This means that the model can explain a significant part of the variation.
- The F-statistic’s p-value is 5.09e-15. This means that the overall model is significant.



On the left side (Figure 3.3) the residuals of the fit are plotted. When we check the residuals, there exists no visible pattern. To further test if there exists any autocorrelation, the Durbin-Watson statistic is used:

Figure 3.3 – UGS Fit Residuals Plot

```
Durbin-Watson test

data:  reg_bic_1
DW = 2.5414, p-value = 0.8184
alternative hypothesis: true autocorrelation is greater
than 0
```

According to the test results, we do not reject the null hypothesis and the true autocorrelation may be 0.

For Diesel Gasoline Sales(DGS):

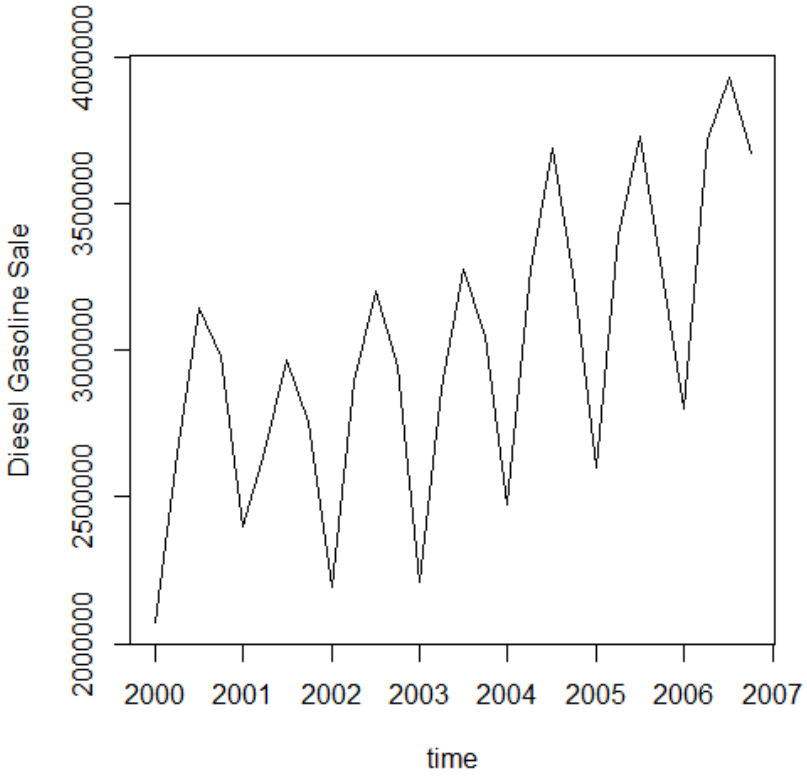
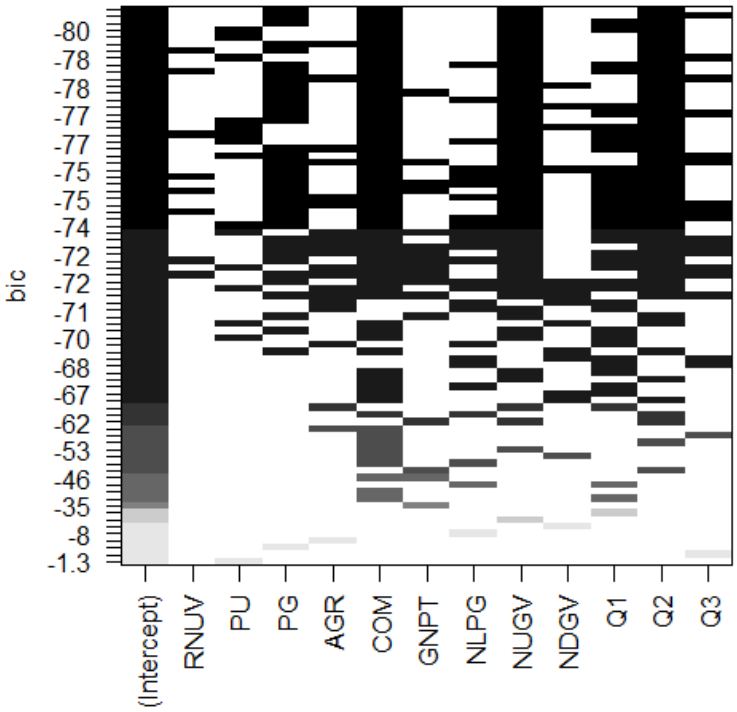


Figure 3.4 – DGS Plot

On the left side (Figure 3.4), "Diesel Gasoline Sales" are given. When this data is analyzed, the variance seems to be steady. Thus, there is no need to take the logarithm of the data. Also there seems to be a linear increasing trend. So, to fit a linear regression model, we do not need to take further differences. To take care of quarterly seasonality, 3 new variables are added which are "Q1, Q2 and Q3". The variable Q1 takes value 1 if the data is collected at Q1 and 0 otherwise. Q2 and Q3 are also defined in a similar manner.



To find the best regression model different fits are sorted according to their Bayesian Information Criterion (BIC). The lowest BIC value model is considered to be the best fit. In our case the best fit has these independent variables:

- **PG:** Price of Diesel Gasoline
- **COM:** GNP Commerce
- **NUGV:** # of Unleaded Gasoline Vehicles
- **Q2**

Figure 3.5 – Variables and BIC Values

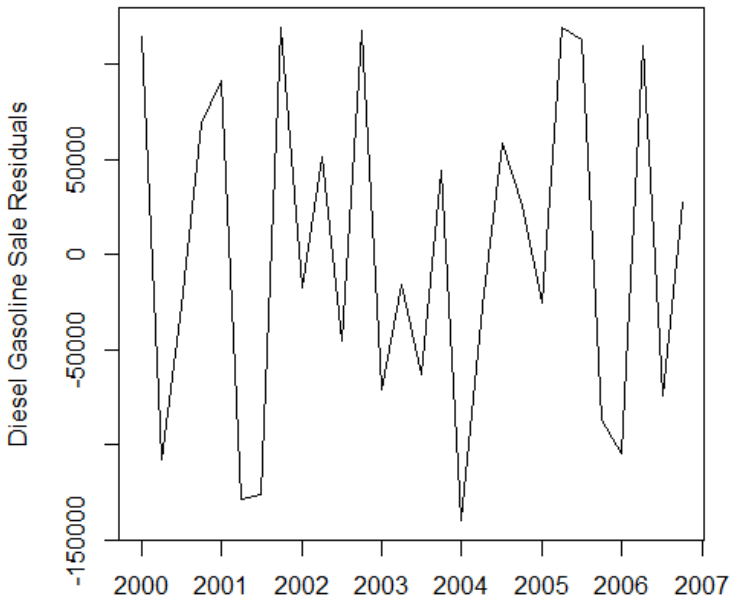
Validity of the Fit:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.974e+06  3.657e+05  -5.397 1.75e-05 ***
PG          -2.857e+03  6.592e+02  -4.334 0.000245 ***
COM          4.052e-01  2.090e-02  19.390 9.49e-16 ***
NUGV         7.607e-01  1.051e-01   7.240 2.27e-07 ***
Q2           2.732e+05  4.203e+04   6.501 1.24e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 94330 on 23 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.9692,  Adjusted R-squared:  0.9638
F-statistic: 180.8 on 4 and 23 DF,  p-value: < 2.2e-16
```

Comments:

- All the p-values of coefficients are smaller than 0.001. Thus, they are all significant.
- Multiple R-squared (0.9692) and Adjusted R-squared (0.9638) values are high. This means that the model can explain a significant part of the variation.
- The F-statistic's p-value is smaller than 2.2e-16. This means that the overall model is significant.



On the left side (Figure 3.6), the residuals of the fit are plotted. When we check the residuals, there exists no visible pattern. To further test if there exists any autocorrelation, the Durbin-Watson statistic is used.

Figure 3.6 – DGS Fit Residuals Plot

```
Durbin-Watson test

data:  reg_bic_2
DW = 2.424, p-value = 0.7824
alternative hypothesis: true autocorrelation is greater
than 0
```

According to the test results, we do not reject the null hypothesis and the true autocorrelation may be 0.

4. Comparison of Methods A and B

To compare forecasting methods, Method A (time series) and Method B (regression) were used to forecast the next four quarters for UGS and DGS. The performance evaluation of each method is based on RMSE (Root Mean Squared Error) and MAPE (Mean Absolute Percent Error) metrics.

RMSE was chosen as the metric because it provides a measure of the mean forecast error in the same units as the original data. A lower RMSE indicates higher accuracy, as it reflects how close the predicted value is to the actual value. In addition, MAPE was used because it expressed the mean percentage difference between the predicted value and the actual value. MAPE is useful for evaluating the accuracy of predictions in terms of relative performance, regardless of the magnitude of the predicted values.

For the UGS, Method A yielded an RMSE of 43830.2 and a MAPE of 0.0359, while Method B achieved a lower RMSE of 25308.8 and a lower MAPE of 0.0196. This demonstrates that Method B provided more accurate forecasts for the UGS compared to Method A.

Similarly, for the DGS, Method A resulted in an RMSE of 129519.5 and a MAPE of 0.0326, whereas Method B exhibited a lower RMSE of 85493.8 and a lower MAPE of 0.0263. These results indicate that Method B outperformed Method A in terms of forecast accuracy for the DGS.

In summary, the comparison of the forecasting methods using RMSE and MAPE metrics highlights that Method B (Regression) consistently delivered more accurate forecasts than Method A (Time Series) for both the UGS and DGS.

Method A(Time Series)	**Method B(Regression)**
rmse_UGS = 43830.2	rmse_UGS = 25308.84
mape_UGS = 0.0358731	mape_UGS = 0.01962965
rmse_DGS = 129519.5	rmse_DGS = 85493.82
mape_DGS = 0.03259292	mape_DGS = 0.02628542

5. Forecasts for UGS and DGS for the year 2007

5.1. Forecasting with Time Series(Method A)

To forecast UGS prices**, Model 5** (as discussed in section 2.4) with SARIMA(1,1,0)(0,1,1) is used. Resulting forecasts for all quarters of 2007 are as below:

	Point Forecast	Lo 80	Hi 80
2007 Q1	687583.6	622688.1	752479.0
2007 Q2	879879.0	810940.7	948817.3
2007 Q3	953991.6	868857.4	1039125.9
2007 Q4	815443.7	724199.9	906687.5

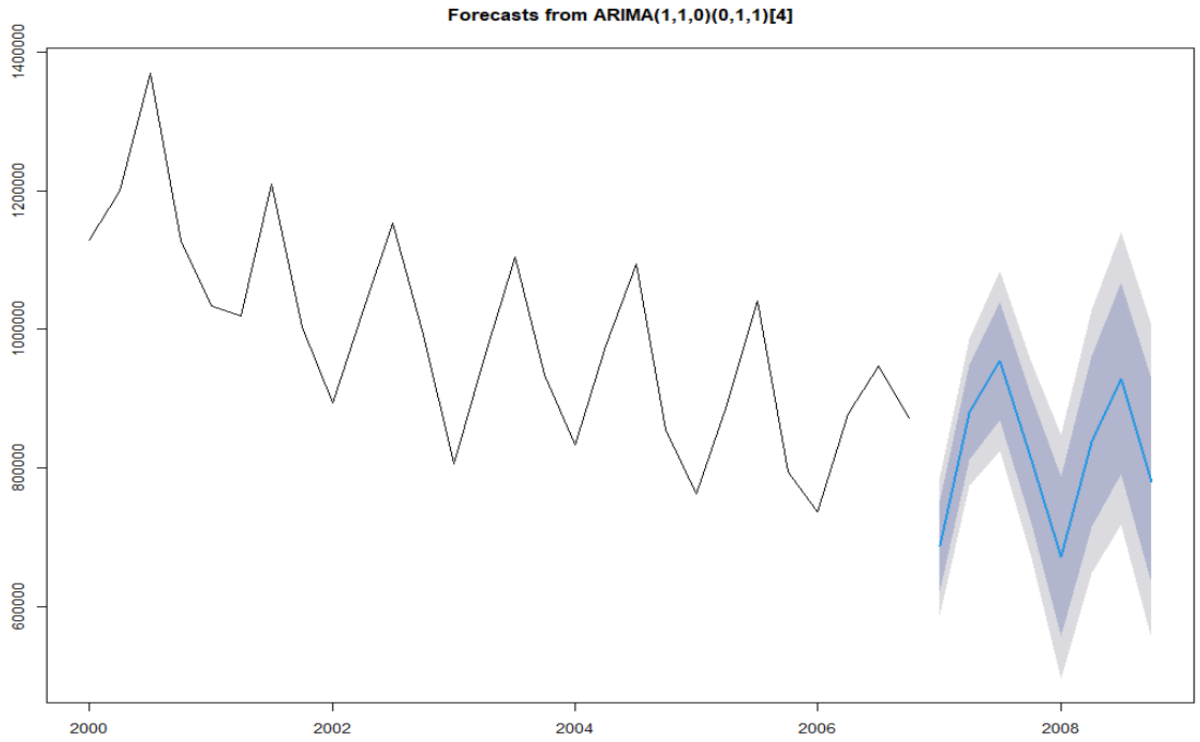


Figure 5.1.1 – Forecasts for UGS with Time Series

To forecast DGS prices**, Model 4** (as discussed in section 2.4) with SARIMA(0,1,0)(1,1,0) is used. Resulting forecasts for all quarters of 2007 are as below:

	Point Forecast	Lo 80	Hi 80
2007 Q1	3135792	2498535	3323049
2007 Q2	3993450	3728628	4258272
2007 Q3	4266170	3941831	4590509
2007 Q4	3909955	3535440	4284470

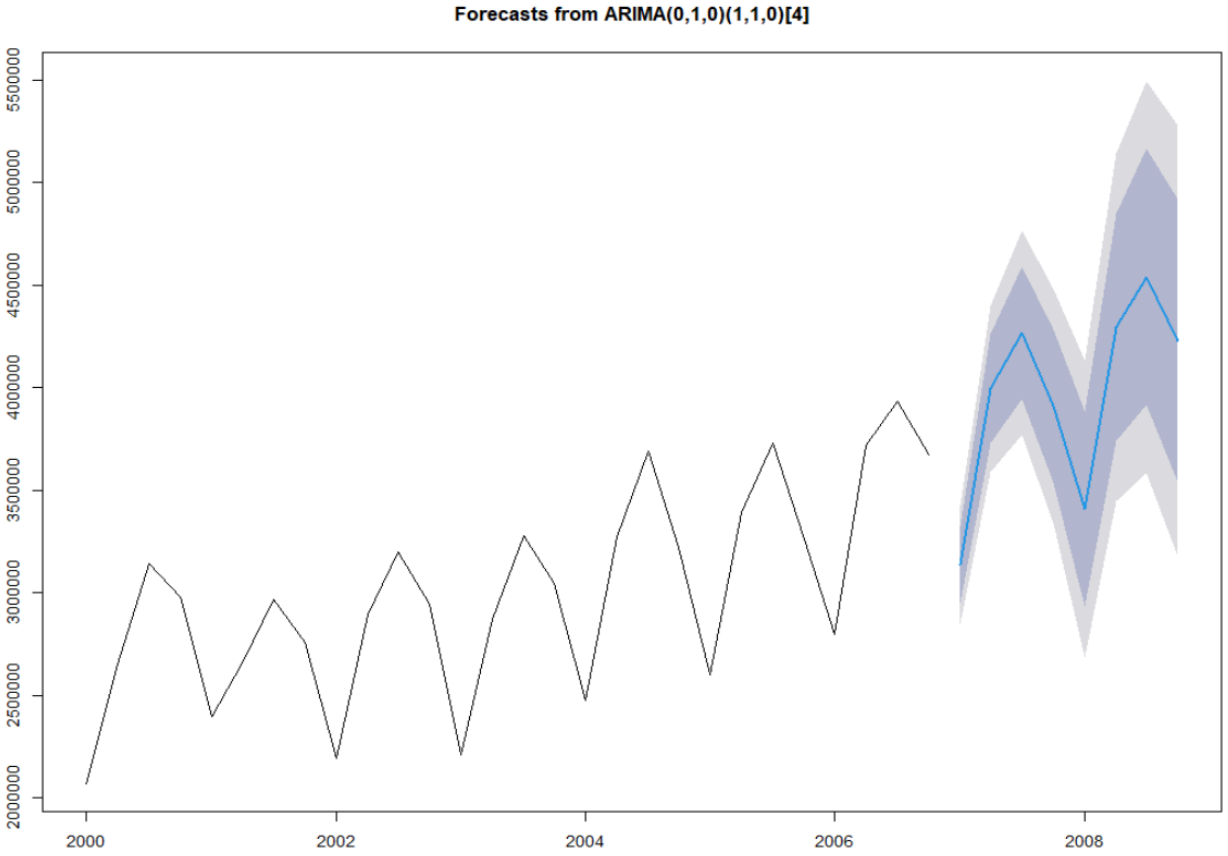


Figure 5.1.2 – Forecasts for DGS with Time Series

5.2. Forecasting with Regression(Method B)

For UGS:

The prediction obtained from the best fit (as discussed in section 3) for the four quarters of 2007 are given below:

```
     Q1       Q2       Q3       Q4
701019.7 826691.1 976435.1 791323.6
```

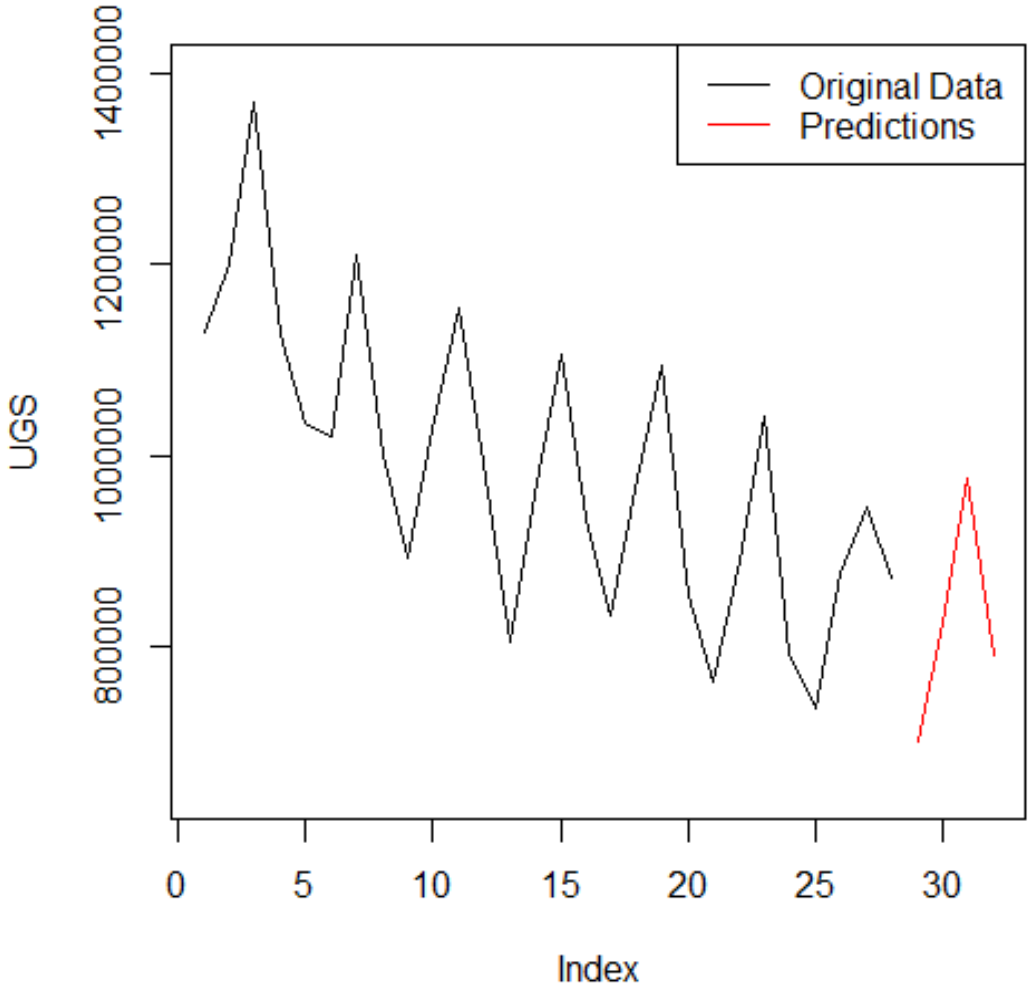


Figure 5.2.1 - Forecasts for UGS with Regression

For DGS:

The prediction obtained from the best fit (as discussed in section 3) for the four quarters of 2007 are given below:

```
     Q1       Q2       Q3       Q4
3175548  3899426  4331507  3949309
```

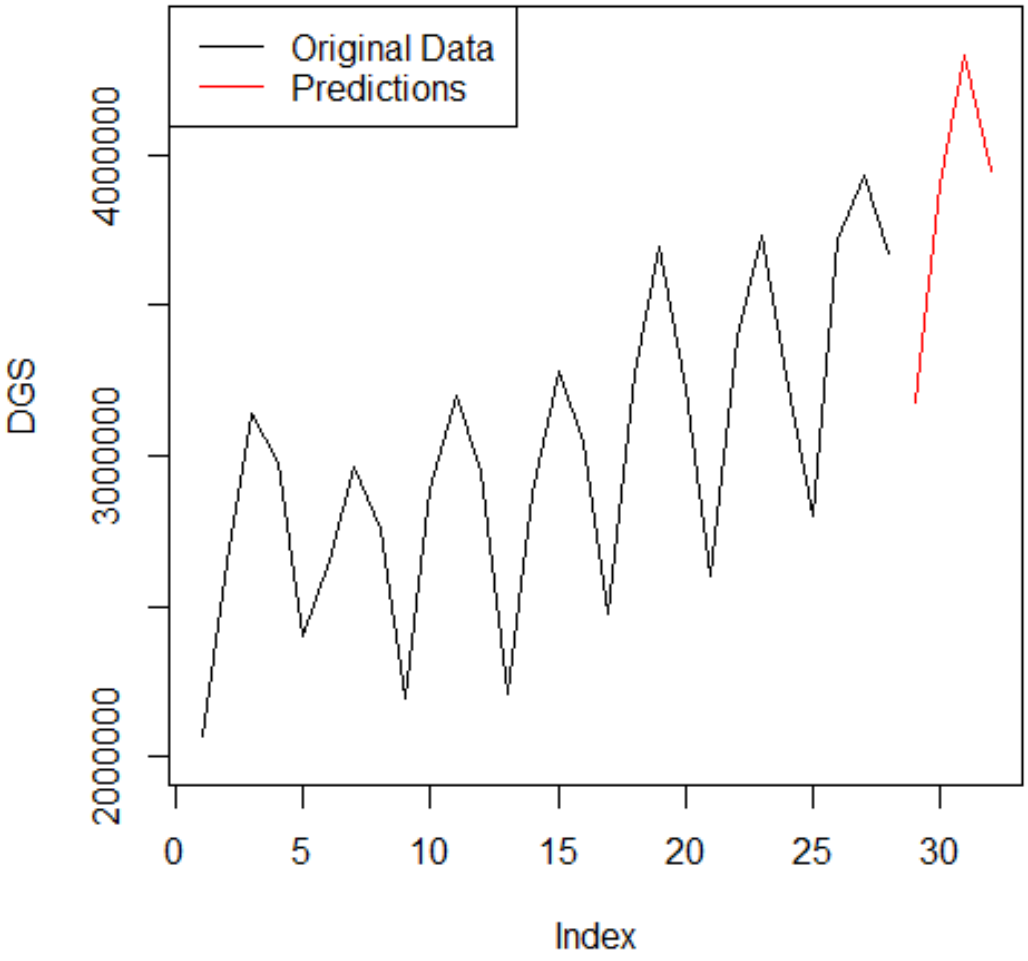


Figure 5.2.2 - Forecasts for DGS with Regression