STAT 463 Final Report

Data set 1:Australian total wine sales

1. Introduction

The data set wineind describes the Australian total wine sales by winemakers in bottles less than liter during Jan 1980 to Aug 1994. The units are counts of the number of sales and there are a total of 176 observations.

2. Methodology

We impute all the required libraries needed for analysis which are require(forecast),require(astsa),require(TSA),library(dplyr),library(readxl),library(tseries), library(stats),library(readxl),library(lmtest). And, we use SARIMA to create our model, AIC & BIC to select the best model, and carry out residual analysis for our chosen model.

3. Model Specification

3a. Reading data

Using head() to display the data

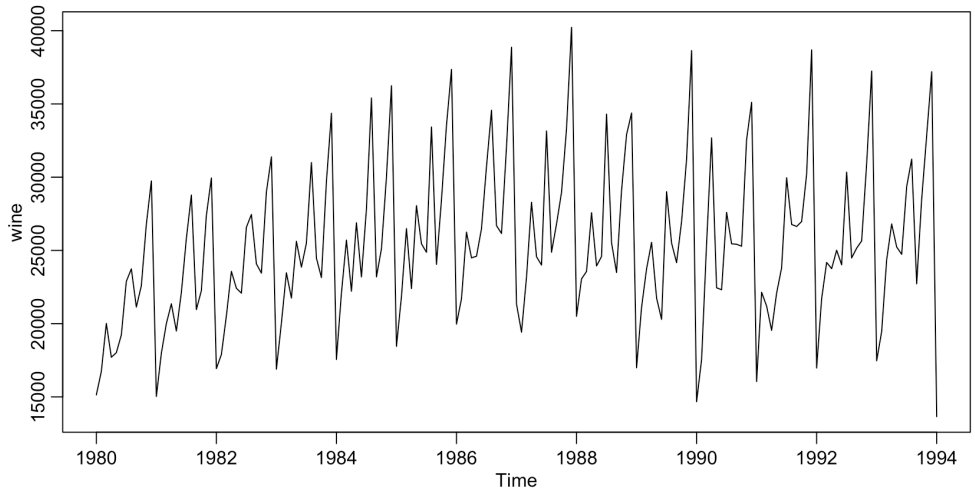
```
> head(wine)
       Jan   Feb   Mar   Apr   May   Jun
1980 15136 16733 20016 17708 18019 19227
```

3b. Analyzing the data

As we convert the data frame into time series data and generate a plot to draw conclusions regarding trends, seasonality, and behavior.

-Trend: There is a strong upward trend.

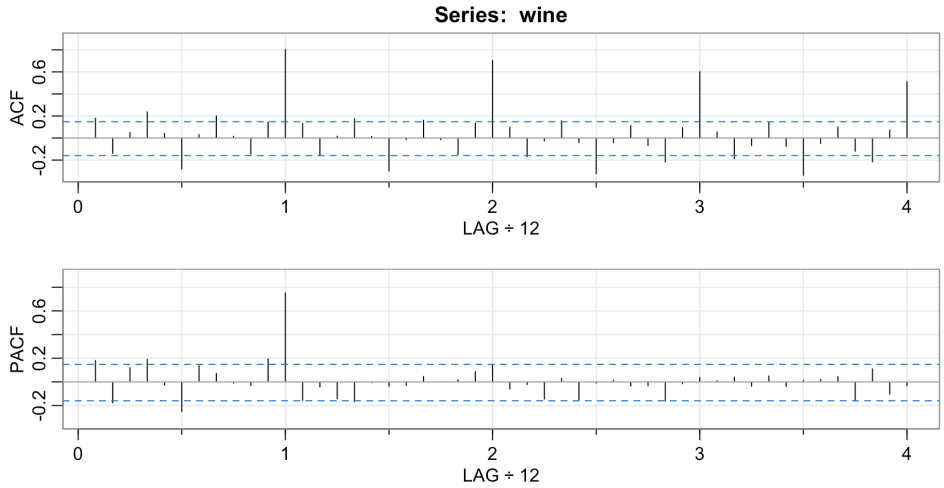
-Seasonality: A seasonal pattern is a rise and fall in data values that repeat after regular intervals.



Time series plot of Australian total wine sales Jan 1980 to Aug 1994.

3b.1 ACF & PACF plots

We have a strong correlation at lags 12,24,36 and so on we consider the existence of a seasonal auto correlation relationship. The PACF below shows 1 seasonal lag at 12.



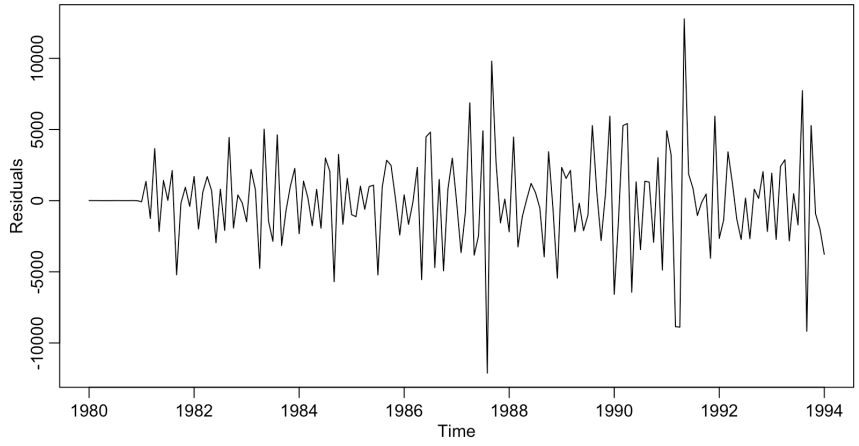
ACF and PACF plot for the data

3c. Seasonality

3c.1 Residual Approach:

1. **Specification of the seasonal part (D=1)**

We do so by fitting the ARIMA(0,0,0)x(0,1,0) model and plotting the graphs. Although the general upward trend is resolved, we plot the ACF & PACF.

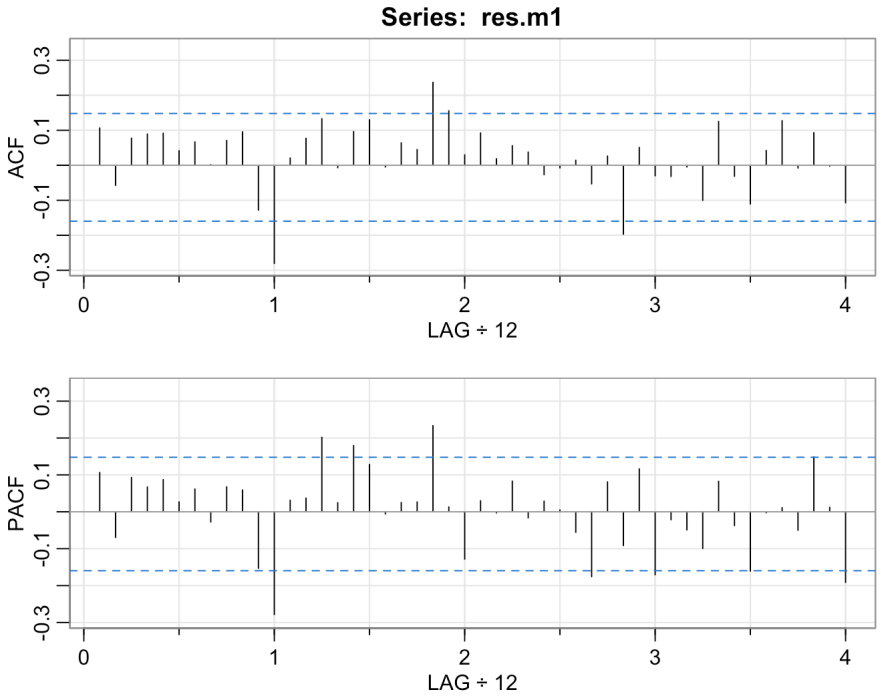


Time series plot of 1st seasonal difference

ACF & PACF plot

The ACF/PACF plots show the seasonal trend is filtered out.

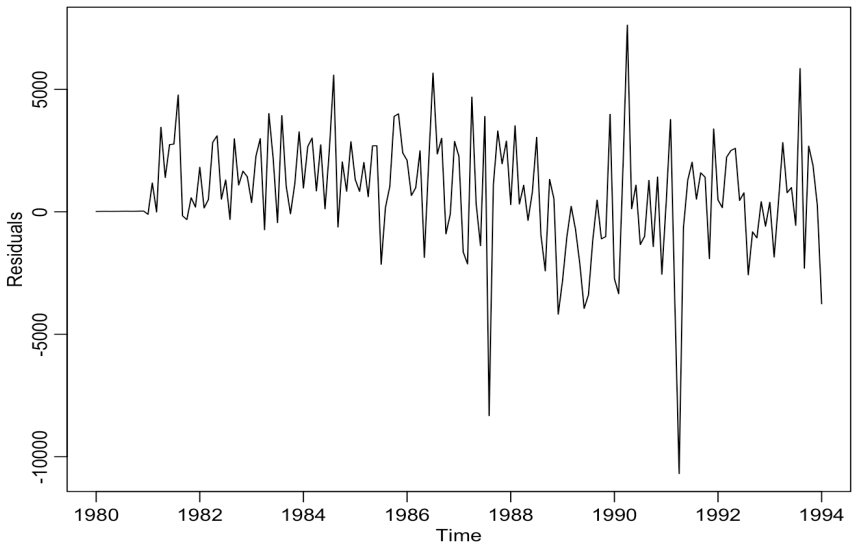
- ACF & PACF shows 1 seasonal lag , indicates SARMA(1,1)



sample ACF/PACF plot of the residuals

2. Specification of the seasonal part (D=1, P=1, Q=1)

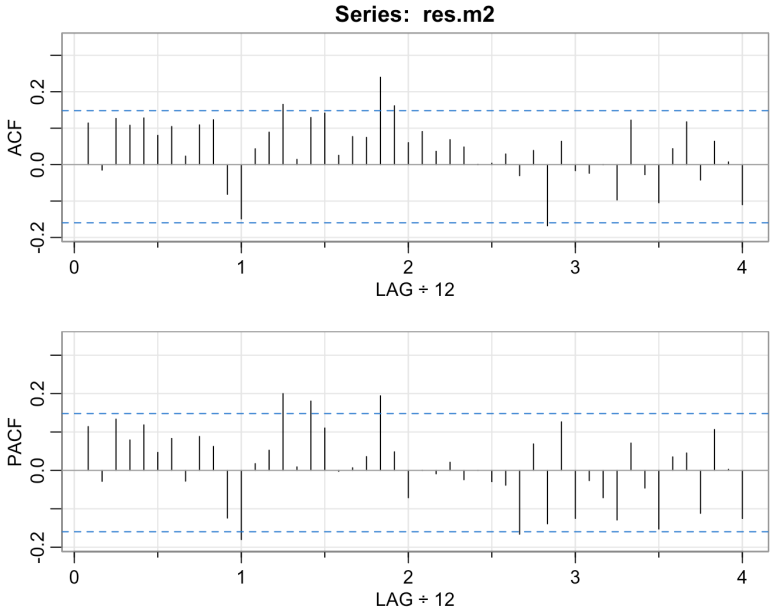
The general upward trend is no longer seen. We plot the residuals ACF & PACF plots.



Time series plot with seasonal AR & MA coefficient (P=1, Q=1)

ACF & PACF plots

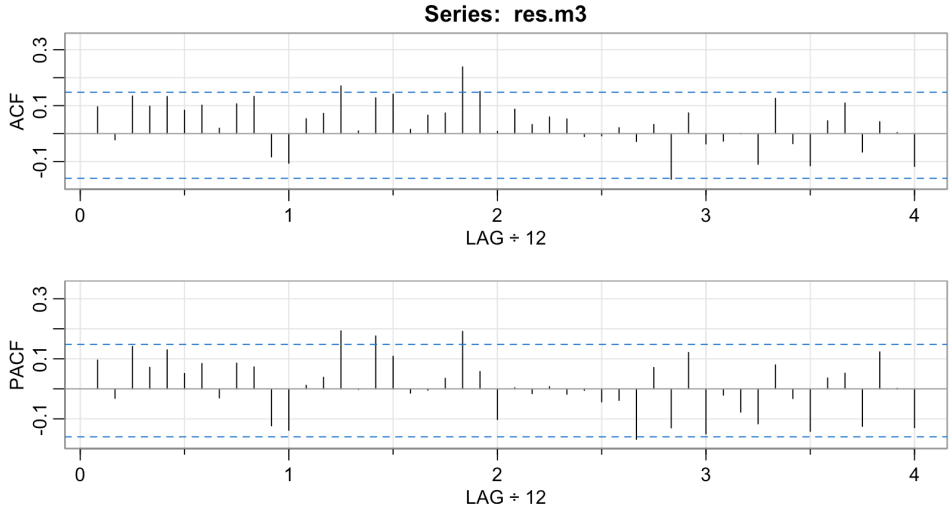
- Autocorrelation is still present on the seasonal lags. Therefore we need to repeat the process with a higher number of Q until filtering out seasonality.



sample ACF/PACF plot of the residuals with AR & MA coefficient (P=1,D=1,Q=1)

3. Specification of the seasonal part (D=1, P=1, Q=2)

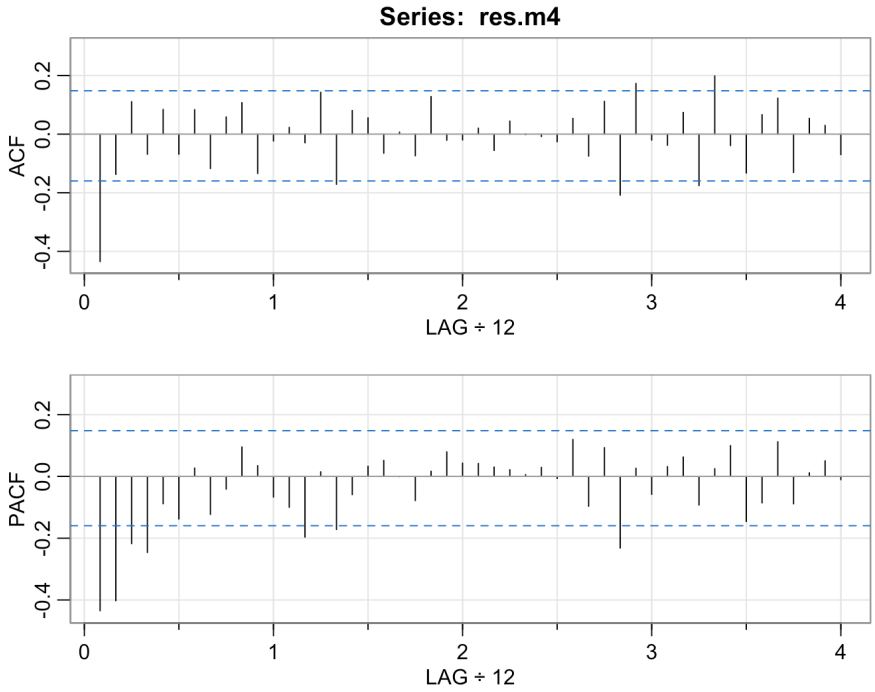
- Now no seasonal lags are present. Hence seasonality specification completes at P=1, D=1, Q=2. We also see that there are multiple lags in the ordinary pane (i.e. lags before 1st seasonal lag).



sample ACF/PACF plot of the residuals with P=1, D=1, Q=2

3d. Non-seasonal Differencing

Here the main aim is to come up with a set of possible models. So let's start with d=1. As we don't see any decaying pattern, we start with ordinary differencing d=1, to get rid of the remaining trend and correlation in ACF/PACF plots. Differencing gives us something. It helps us come up with a model.



sample ACF/PACF plot of the residuals with ordinary differencing

- We get a high correlation at 1st lag and also we observe a few significant lags. All this is due to the intervention point.
- We come up with AR(4) from ACF & PACF plots.
- Also, there is no evidence for an ordinary trend. we can still apply the ADF test on residuals to make sure.

```
> adf.test(res.m4)

        Augmented Dickey-Fuller Test

data:  res.m4
Dickey-Fuller = -9.1311, Lag order = 5, p-value = 0.01
alternative hypothesis: stationary
```

EACF

Now we use EACF on the residuals of the previous step (i.e. res.m4), to see information about AR and MA parts lefts in residuals. Our candidates for ARMA part are ARMA(1,2), ARMA(1,3) & ARMA (2,2)

```
> eacf(res.m4) #
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x x o o o o o o o o o  x  o  o
1 x x o o o o o o o x o  x  o  o
2 x x o o o o o o o o o  x  o  o
3 x x o o o o o o o o o  x  o  o
4 x x x o o o o o o o o  x  o  o
5 x x x o o o o o o o o  x  o  o
6 x x o o o o x o o o o  x  o  o
7 x x x o o o o o o o o  x  o  o
```

Therefore, our tentative models are specified as

- SARIMA (4,1,0)x(1,1,2) which we fit as a model 5
- SARIMA (1,1,2)x(1,1,2) which we fit as a model 6
- SARIMA (1,1,3)x(1,1,2) which we fit as a model 7
- SARIMA (2,1,2)x(1,1,2) which we fit as a model 8

4.Model Fitting

From the given set of models, we start fitting one by one and we eventually get that model7 which is **SARIMA (1,1,3)x(1,1,2)** can fit our data the best

- AIC/BIC function gives us model7 i.e. SARIMA(1,1,3)x(1,1,2) as best model.

```
> sort.score(sc.AIC, score = "aic")
         df      AIC
m7.wine  8 3002.718
m8.wine  8 3003.715
m6.wine  7 3006.545
m5.wine  8 3019.191
> sort.score(sc.BIC, score = "aic")
         df      AIC
m7.wine  8 3027.758
m6.wine  7 3028.455
m8.wine  8 3028.754
m5.wine  8 3044.230
```

- Here is the coefficient teat for SARIMA (1,1,3)x(1,1,2)

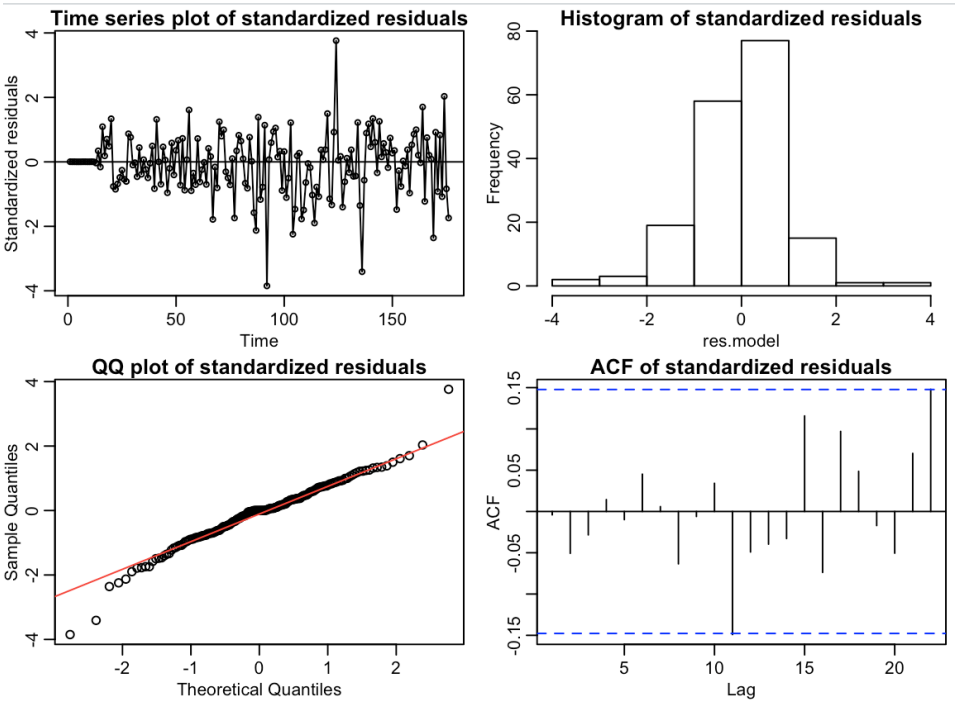
```
> coeftest(m7.wine) #

z test of coefficients:

      Estimate Std. Error  z value  Pr(>|z|)
ar1  -0.840208   0.055637 -15.1015 < 2.2e-16 ***
ma1  -0.118755   0.151072  -0.7861   0.43182
ma2  -0.897081   0.131936  -6.7994 1.051e-11 ***
ma3   0.221603   0.100868   2.1970   0.02802 *
sar1  0.642515   0.577263   1.1130   0.26569
sma1 -1.280187   0.616343  -2.0771   0.03780 *
sma2  0.362726   0.419613   0.8644   0.38735
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

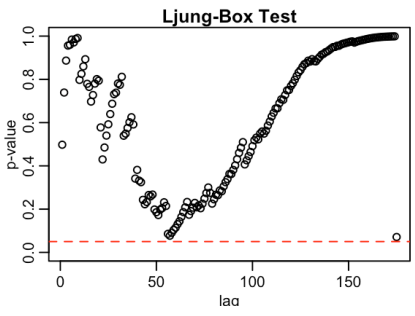
5. Model Diagnostics

5a. Residual Analysis

From the residual analysis of the SARIMA (1,1,3)x(1,1,2) model we draw the following conclusions:

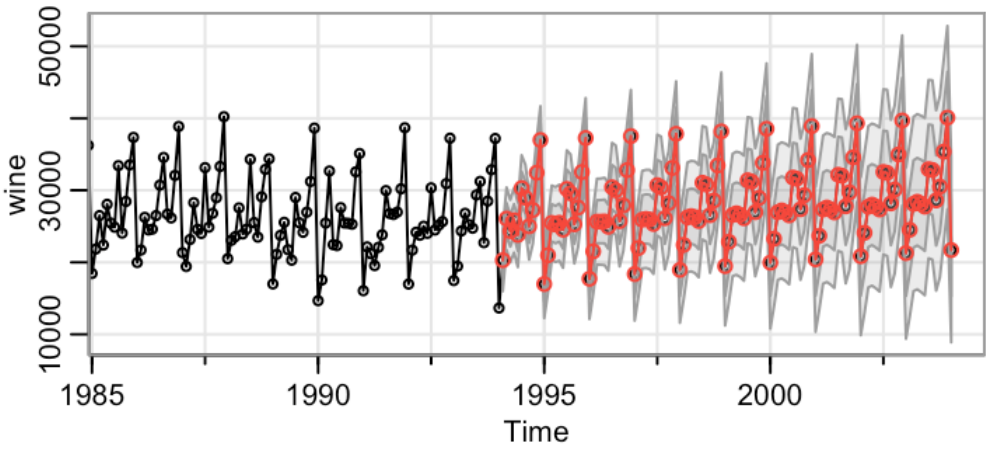
1. The histogram shows a normal distribution of the residuals.
2. ACF plot shows the presence of white noise.
3. Ljung Box test shows good residuals.
4. QQplot shows normality with 2 outliers at the head and 1 at the tail.
5. TS plots do not show any trend (as seen earlier) except an intervention point.





6. Forecasting

After getting one best model SARIMA(1,1,3)x(1,1,2) we will consider it for the prediction of future realizations of time series. The below figure shows forecasts for a lead time of 10 years for the model that we fit.



7. Conclusion

We have finally forecasted the total wine sales for the next 10 years using the model we found out that SARIMA(1,1,3)x(1,1,2) is the best out of the set of possible models. The forecast shows that the sales of wine fluctuate in a similar pattern for every 1 year

Data set 2: CO2 Reading in Hawaii

1. Introduction

The data set CO2Hawaii describes the Monthly carbon dioxide readings at Mauna Loa, Hawaii.

2. Required Library

We impute all the required libraries needed for analysis which are require(forecast), require(astsa),require(TSA), library(dplyr), library(readxl), library(tseries). And, we use SARIMA to create our model, AIC & BIC to select the best model, and carry out residual analysis for our chosen model.

3. Model Specification

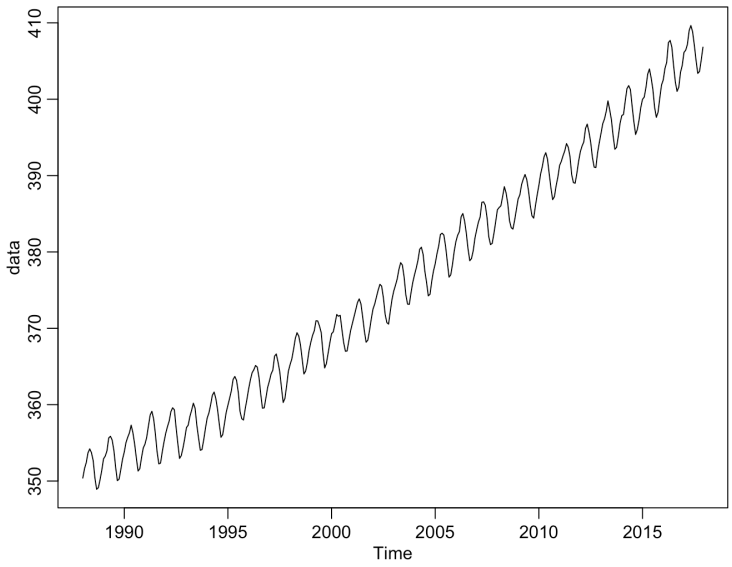
3a. Reading data

Using head() to display the data

```
> head(Hawaii)
# A tibble: 6 × 3
   Year Month   CO2
  <dbl> <dbl> <dbl>
1  1988     1  350.
2  1988     2  352.
3  1988     3  352.
4  1988     4  354.
5  1988     5  354.
6  1988     6  354.
```

3b. Analyzing the data

As we convert the data frame into time series data and generate a plot, we know that the data has a strong upward trend and has a seasonal pattern is a rise and fall in data values that repeat after regular intervals.

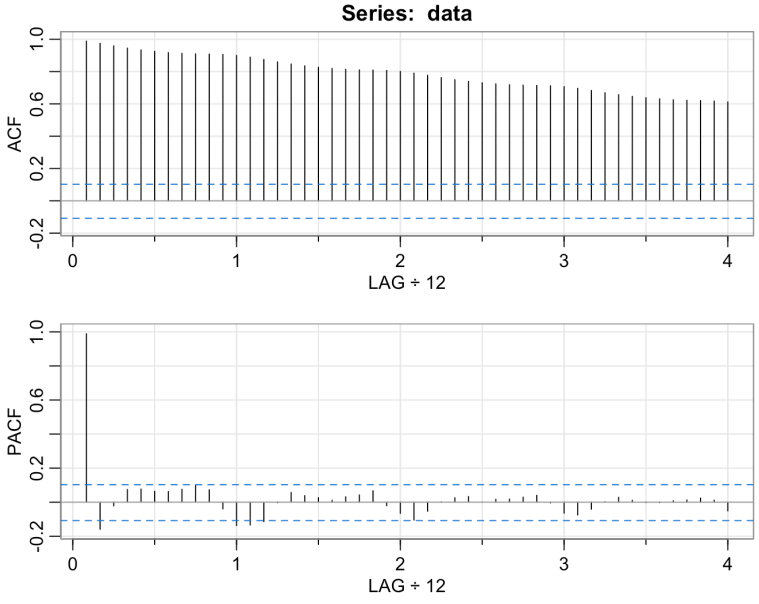


Time series plot of carbon dioxide readings at Mauna Loa, Hawaii

3b.1 ACF & PACF plots

From the have a strong correlation at all lags and so on we consider the existence of a seasonal auto correlation relationship. The PACF below shows 1 seasonal lag at 12

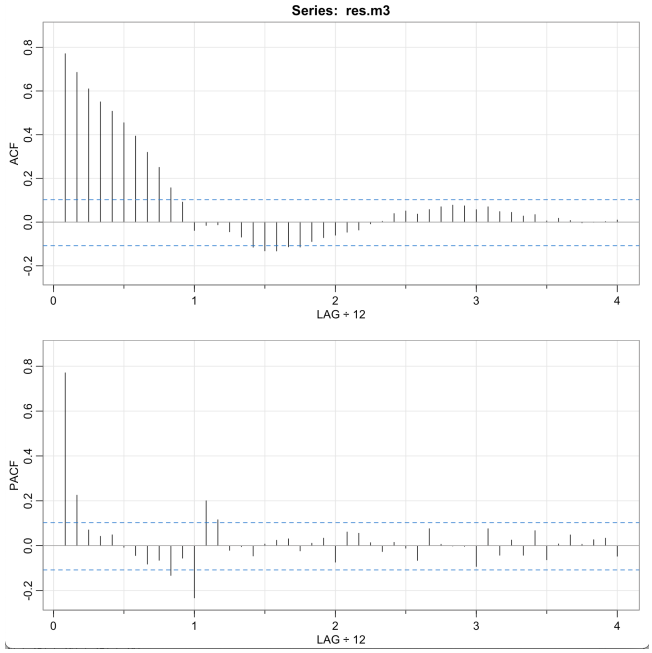
ACF and PACF plot for the data



3c. Seasonality

3c.1 Residual Approach:

In this process, we first did the specification of the seasonal part (D=1), seasonal trend is still present hence with took (D=2) then we then did the specification of the seasonal part (P=1, D=2, Q=1) and (P=1, D=2, Q=2) in order. We decide to choose the seasonal part (D=1, P=2, Q=2), and here is the ACF and PACF plot of SARIMA(0,1,0)x(1,2,2). No seasonal lag is present.



sample ACF/PACF plot of the residuals with P=1, D=2, Q=2

3d. Non-seasonal Differencing

Here the main aim is to come up with a set of possible models. So let's start with d=1. As we don't see any decaying pattern, we start with ordinary differencing d=1, to get rid of the remaining trend.

EACF

Our candidates for ARMA part are ARMA(1,2), ARMA(1,1) & ARMA (2,2)

```
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x o o o o o o o o o o  o  o  o
1 x o o o o o o o o o o  o  o  o
2 x x o o o o o o o o o  o  o  o
3 x x o x o o o o o o o  o  o  o
4 x x x o o o o o o o o  o  o  o
5 x x x x o o o o o o o  o  o  o
6 x o o x o o o o o o o  o  o  o
7 x x o x x o o o o o o  o  o  o
```

Therefore, our tentative models are specified as

- SARIMA (1,1,2)x(1,2,2) which we fit as a model 5
- SARIMA (1,1,1)x(1,2,2) which we fit as a model 6
- SARIMA (2,1,2)x(1,2,2) which we fit as a model 7

4.Model Fitting

From the given set of models, we start fitting one by one and we eventually get that model7 which is **SARIMA (1,1,1)x(1,2,2)** can fit our data the best

- AIC/BIC function gives us model7 i.e. SARIMA(1,1,1)x(1,2,2) as best model.

```
> sort.score(sc.AIC, score = "aic")
           df      AIC
hawaii.m6  6 321.6286
hawaii.m5  7 323.5567
hawaii.m7  8 325.3018
> sort.score(sc.BIC, score = "aic")
           df      AIC
hawaii.m6  6 344.9452
hawaii.m5  7 350.7594
hawaii.m7  8 356.3906
```

- Here is the coefficient test for SARIMA (1,1,1)x(1,2,2)

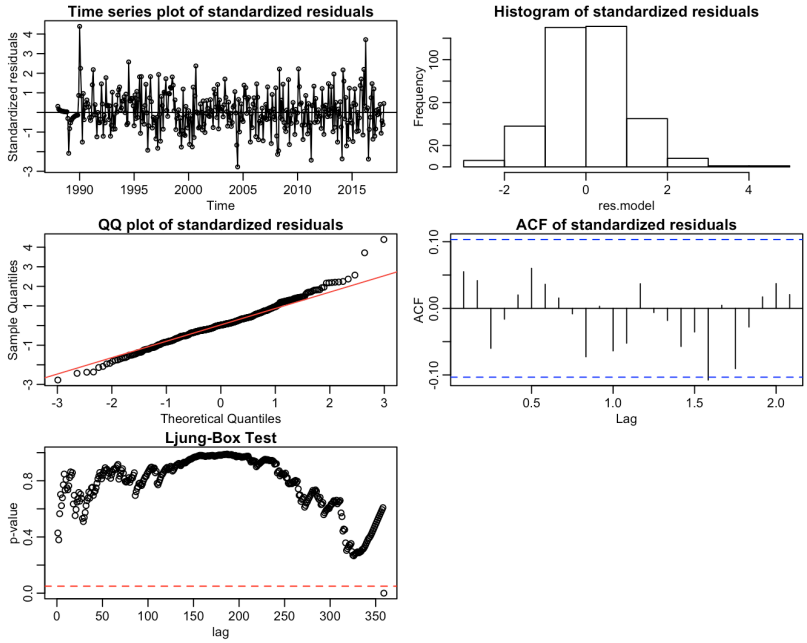
```
z test of coefficients:

      Estimate Std. Error  z value  Pr(>|z|)
ar1   0.192452   0.126096   1.5262    0.1270
ma1  -0.573590   0.104431  -5.4926 3.962e-08 ***
sar1  0.033581   0.076520   0.4389    0.6608
sma1 -1.844570   0.070341 -26.2232 < 2.2e-16 ***
sma2  0.865363   0.072756  11.8940 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

5. Model Diagnostics

5a. Residual Analysis

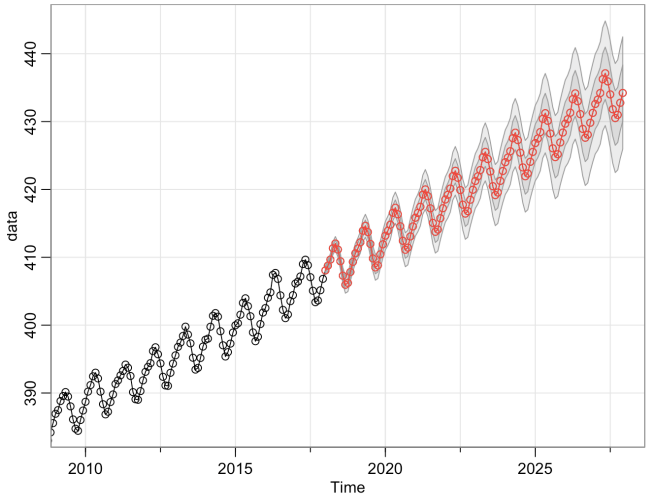
From the residual analysis of the SARIMA (1,1,1)x(1,2,2) model we draw the following conclusions:

1. The histogram shows a normal distribution of the residuals.
2. ACF plot shows the presence of white noise.
3. Ljung Box test shows good residuals.
4. QQplot shows normality with 2 outliers at the head and 1 at the tail.
5. TS plots do not show any trend (as seen earlier) except an intervention point.



6. Forecasting

After getting one best model SARIMA(1,1,1)x(1,2,2) we will consider it for the prediction of future realizations of time series. The below figure shows forecasts for a lead time of 10 years for the model that we fit.



7. Conclusion

We have finally forecasted the carbon dioxide readings at Mauna Loa, Hawaii for the next 10 years using the model we found out that SARIMA(1,1,1)x(1,2,2) is the best out of the set of possible models. The forecast shows an upward trend which indeed sounds true as CO2 reading rate possibly goes high every year due to global warming and many reasons