Assignment 3

Shivakanth Thudi

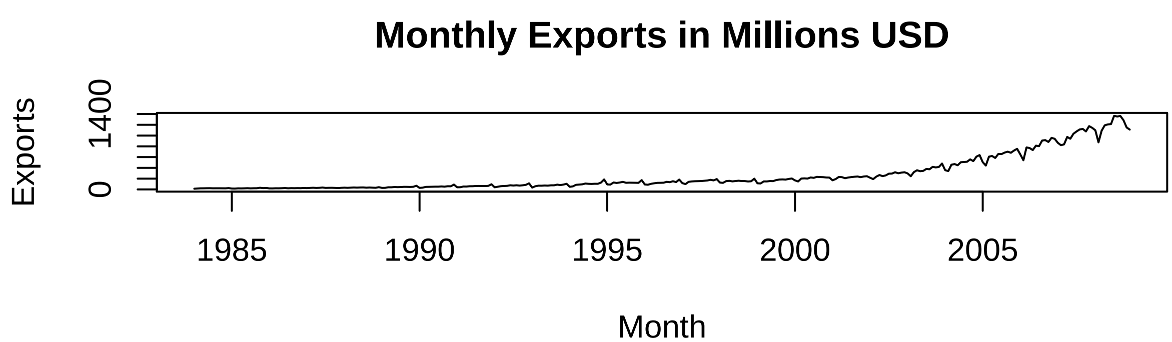
12/05/2016

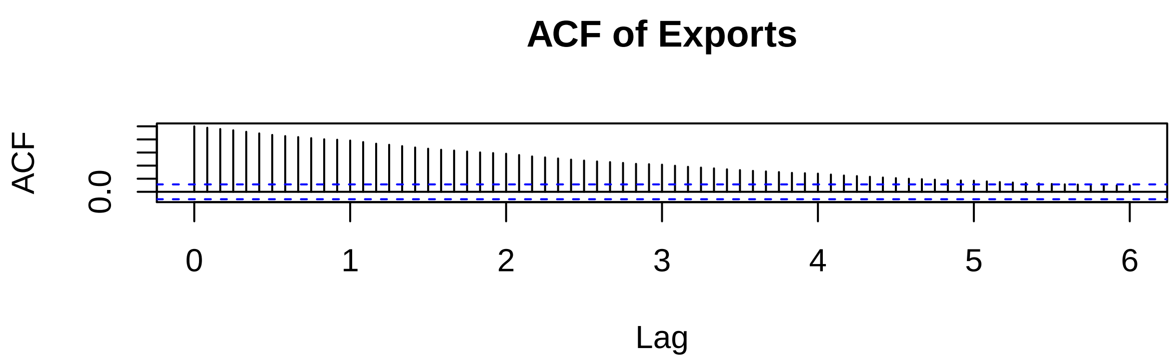
Introduction

We read in the data and observe that both the export and import time series exhibit non-stationarity and that there is an increasing trend over the years. The data exhibits seasonality as well, though this is more apparent in the later years. We also observe that the variability in the import and export variables increases with time.

Since our primary objective is to build a model that forecasts well, we split the data into training and testing sets. The training set consists of the first 19 years of data, with observations from January 1984 - December 2002. The test set includes observations for the next six years, from January 2003 - December 2008.

In each method, we consider whether we need to log-transform the data or not, and fit models accordingly.





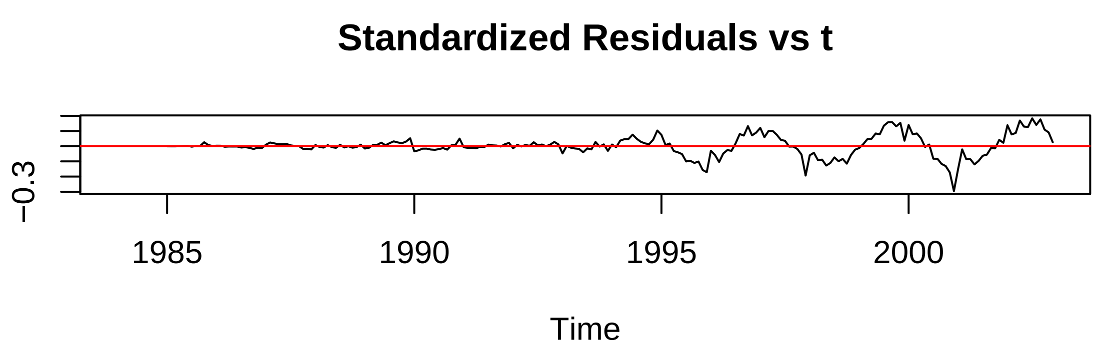
Part (a) - Holt-Winters

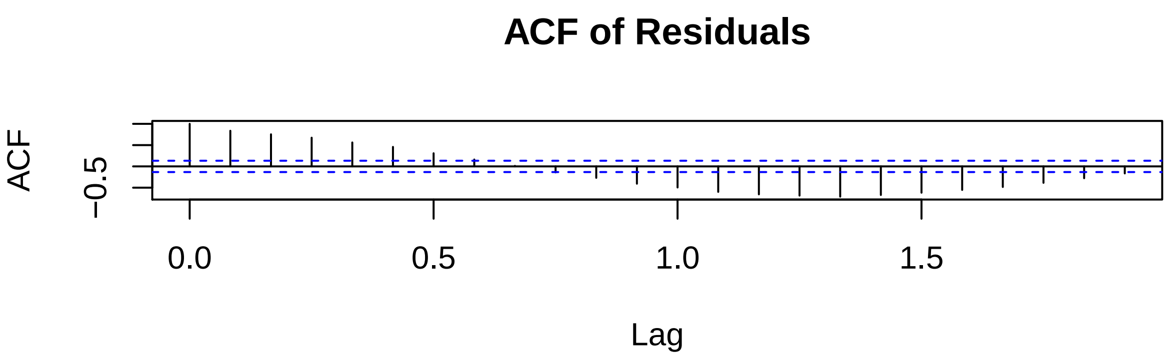
Since the data was non-stationary and exhibited both trend and seasonality, our intuition is that a triple exponential smoothing approach would work best. Since the variability increases with time, a multiplicative version of triple exponential smoothing should work better than the additive version. However, we will look at all possibilities (Single, Double, and Triple Exponential smoothing with both additive and multiplicative seasonality), and determine which method has the best predictive root mean squared error on the test set.

For the level, trend, and seasonal (α, β, γ) parameters, we check all possible values from 0.0 to 1.0 with a step size of 0.01, and pick the best parameters in each smoothing method that correspond to the lowest RMSE on the test set.

Model	RMSE on the test set	Optimal α	Optimal β	Optimal γ
Single Exponential Smoothing (SES)	529.8	1.0	N/A	N/A
Double Exponential Smoothing (DES)	77.63	0.04	0.82	N/A
Triple Exponential Smoothing (TES) with Additive Seasonality	**61.89**	0.14	0.94	0.55
Triple Exponential Smoothing (TES) with Multiplicative Seasonality	62.4	0.03	0.99	0.46

We observe that the Triple Exponential Smoothing method with additive seasonality works best here, since it has the lowest RMSE - 61.89, on the test set. The version with multiplicative seasonality performs slightly worse. **However, in both triple exponential smoothing methods, we observe from the residual diagnostics that the residuals are correlated.** In particular, note the following plot of the standardized residuals and the ACF for the triple exponential smoothing method with multiplicative seasonality:





We see increasing variability in the standardized residuals with time, and there are many significant spikes in the ACF plot. So while these triple exponential smoothing methods have low RMSE on the testing set, the model assumptions are being violated.

This prompts us to log-transform the data and consider Triple Exponential Smoothing with additive seasonality, since logging will convert multiplicative seasonal patterns to additive patterns. We perform exponential smoothing on the log-transformed exports data.

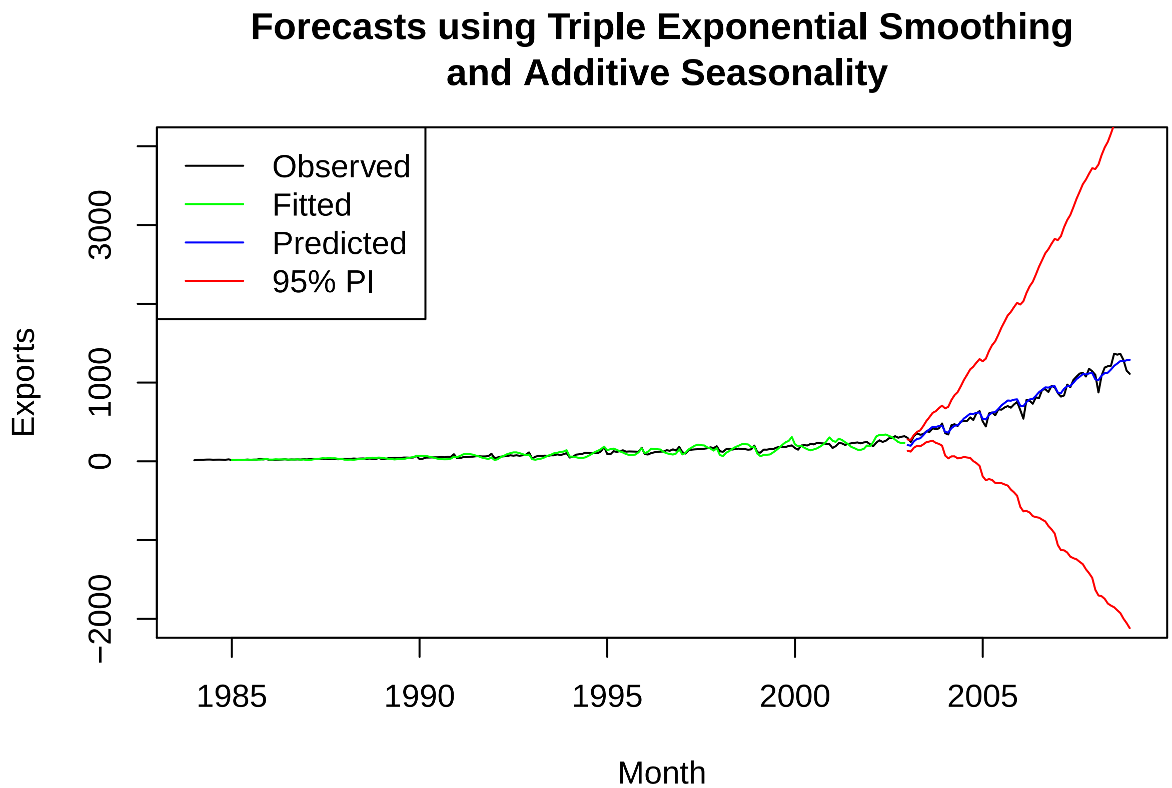
Model	RMSE on the test set after **Log-Transformation**	Optimal α	Optimal β	Optimal γ
Single Exponential Smoothing (SES)	529.8	1.0	N/A	N/A
Double Exponential Smoothing (DES)	78.35	0.23	0.87	N/A
Triple Exponential Smoothing (TES) with Additive Seasonality	65.53	0.1	0.86	0.62

We observe that after log-transforming the exports data, the triple exponential smoothing method with additive seasonality works best here, since it has the lowest RMSE - 65.53, on the test set. However, **we note that the model assumption of zero-correlation is still not met, after we log-transform and perform smoothing**. The residual diagnostics are attached in the appendix.

When we tune the smoothing parameters and build models that minimize the testing RMSE - regardless of whether we log transform or not - we find that the residuals are correlated. This casts doubt on the validity of the model to make predictions. So while we provide the optimal model's forecasts below ("optimal" according to test set RMSE), **it is our recommendation that the other models in this report be used in lieu of these exponential smoothing methods since the other models satisfy all model assumptions.**

The model with the lowest RMSE was the triple exponential smoothing model with additive seasonality (without log-transformation). Below, we show the model's forecasts and their associated 95% prediction intervals in both a tabular and graphical format. For the tabular format, we show only the first year's forecasts; in the graph, we show all the forecasts from January 2003 to December 2008.

```
         Forecast    Lower    Upper
Jan 2003 205.5591 133.4399 277.6784
Feb 2003 197.9695 123.2376 272.7015
Mar 2003 253.4345 173.2446 333.6244
Apr 2003 285.1637 196.1804 374.1469
May 2003 291.1879 190.0553 392.3205
Jun 2003 331.1590 214.7977 447.5202
Jul 2003 377.3673 243.0744 511.6602
Aug 2003 406.7847 252.2147 561.3547
Sep 2003 438.5611 261.6657 615.4565
Oct 2003 435.4711 234.4362 636.5061
Nov 2003 448.6065 221.7999 675.4132
Dec 2003 453.3626 199.2943 707.4309
```



The forecasts are made for the period between January 2003 to December 2008.

Residual Diagnostics

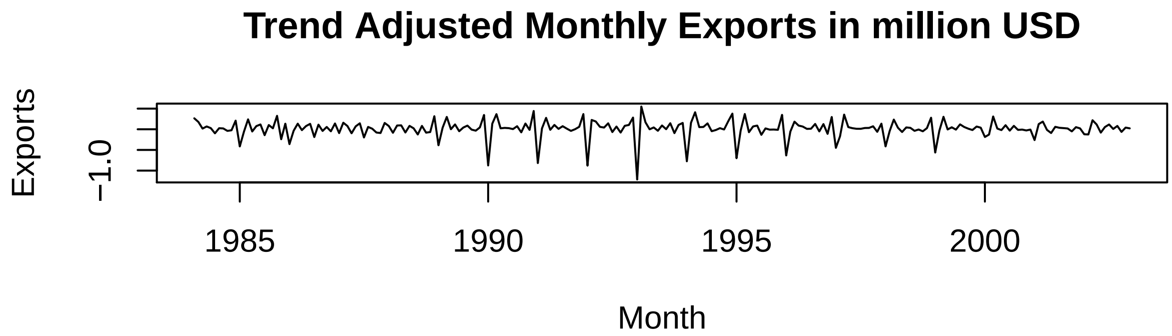
The residual diagnostics are shown in the appendix. The only assumption that is met is the assumption of zero-mean. The assumptions of zero-correlation and homoscedasticity are both not met. The assumption of normality is also not met, however, this is not a requirement in the case of exponential smoothing methods.

Part (b) - SARIMA

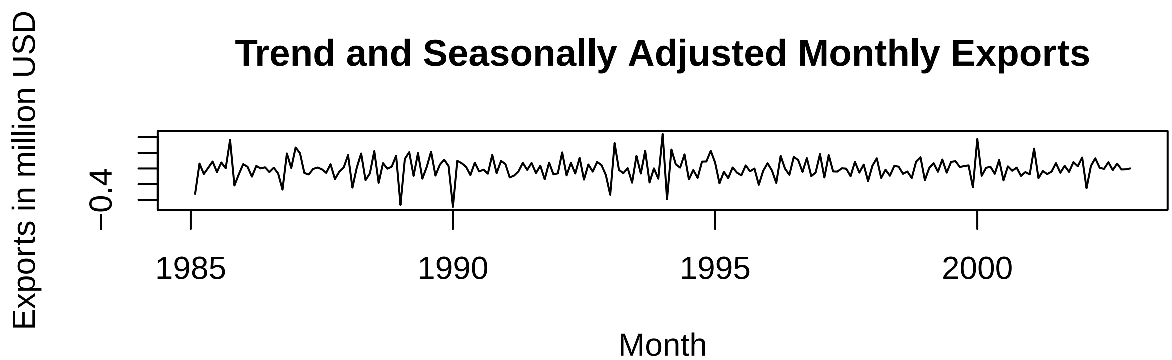
Since the variability increases with time, we log-transform the export data to stabilize the variance.

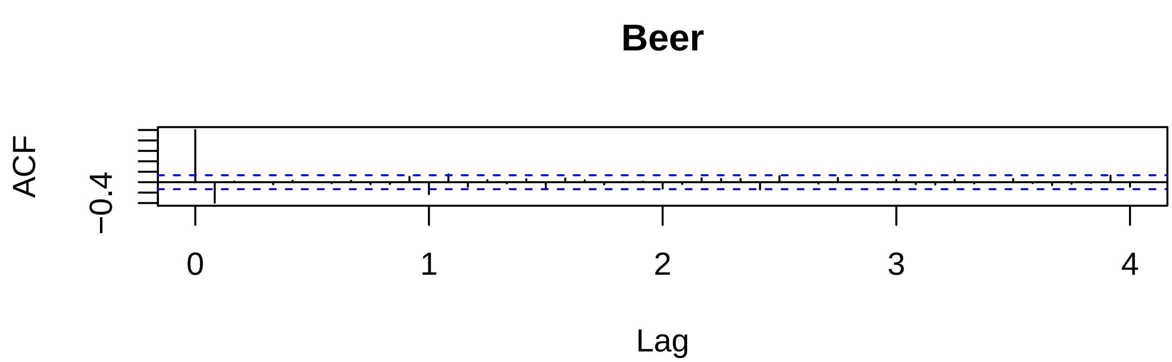
Choosing d and D

Since the time series was not stationary, we will need to do some differencing. We difference once (ordinary) and look at the transformed time-series and ACF plot afterwards:



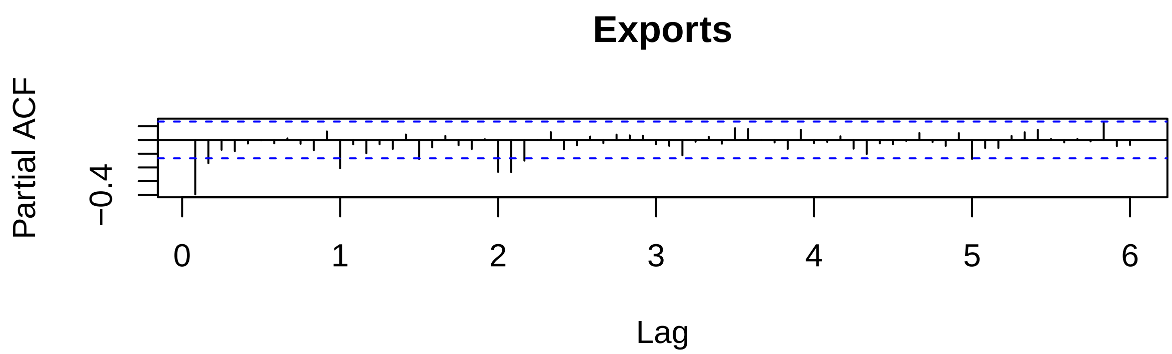
We observe that after ordinary differencing once, the observed time series looks flat. The ACF plot depicts rapid decay rather than slow decay. The ACF plot also indicates that there is seasonality present, due to the recurring nature of the peaks when the lag is 12. So we do seasonal differencing as well.





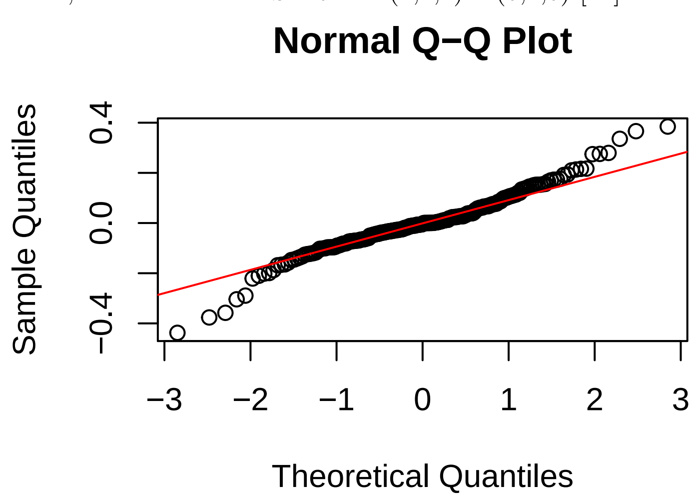
After seasonally differencing once, the ACF plot does not indicate that any more seasonal differencing is required. **So we choose the order of the ordinary differencing, d as 1 and the order of the seasonal differencing, D as 1 too.**

Choosing p,q,P,and Q



We observe that there are are 2 significant spikes in the PACF. Considering seasonal lags separated by 12 (at 1s, 2s, 3s, etc. with s=12), we see 2 spikes. **So p might be less than/equal to 2 and P less than/equal to 2.** In the ACF, we see 1 significant spike, and with seasonal lags separated by 12 we see 1 significant spike. **So q might be less than 1, and Q might be less than 1.**

So with 36 different combinations of p,P,q, and Q, we fit SARIMA models on the log-transformed data using Maximum Likelihood. However, when we pick an optimal model based on test set RMSE and look at the residual diagnostics, we observe that the normality assumption is not met. In particular, observe the qqplot below, which is for the SARIMA (1,1,1) x (0,1,0) [12] model:



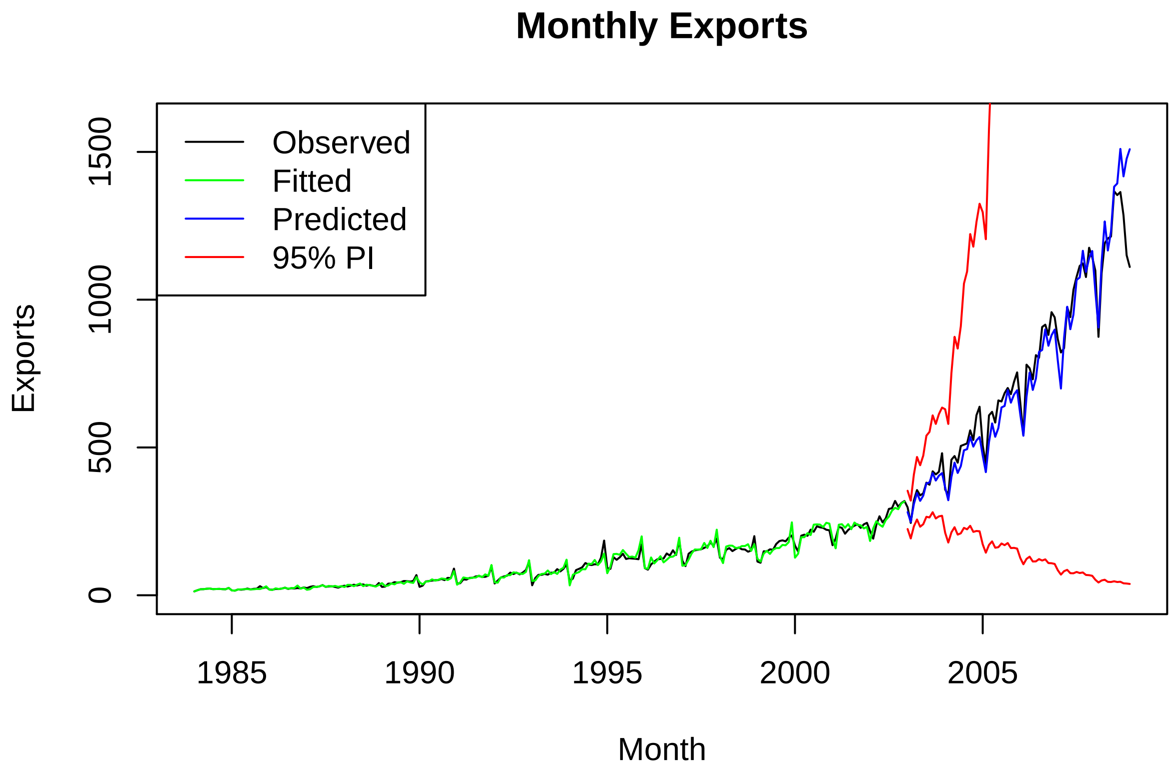
Since the normality assumption is not met, we fit the models using the method of Least Squares instead. The following table shows the top 5 optimal choices for p,P,q, and Q based on test set RMSE when fitting with least squares:

```
                                 Model     RMSE
31 SARIMA ( 2 1 1 ) x( 0 1 0 )[12] 79.98316
19 SARIMA ( 1 1 1 ) x( 0 1 0 )[12] 84.95750
7  SARIMA ( 0 1 1 ) x( 0 1 0 )[12] 85.06099
13 SARIMA ( 1 1 0 ) x( 0 1 0 )[12] 85.77685
1  SARIMA ( 0 1 0 ) x( 0 1 0 )[12] 86.18874
```

We pick as our optimal model the one with the lowest test RMSE, which is the SARIMA(2,1,1)x(0,1,0)[12] model here. The test set RMSE was 79.98.

Below, we show the model's forecasts and their associated 95% prediction intervals in both a tabular and graphical format. For the tabular format, we show only the first year's forecasts; in the graph, we show all the forecasts from January 2003 to December 2008.

```
         Forecast    Lower    Upper
Jan 2003 281.4010 223.9657 353.5655
Feb 2003 248.0462 192.1569 320.1910
Mar 2003 308.6399 233.1587 408.5568
Apr 2003 346.2803 256.2597 467.9239
May 2003 319.3219 231.8434 439.8074
Jun 2003 337.0589 240.3899 472.6018
Jul 2003 378.4828 265.4004 539.7477
Aug 2003 381.4239 263.1771 552.7996
Sep 2003 413.4802 280.9142 608.6054
Oct 2003 388.0191 259.7228 579.6904
Nov 2003 404.3842 266.8212 612.8694
Dec 2003 413.1951 268.8801 634.9679
```



Residual Diagnostics

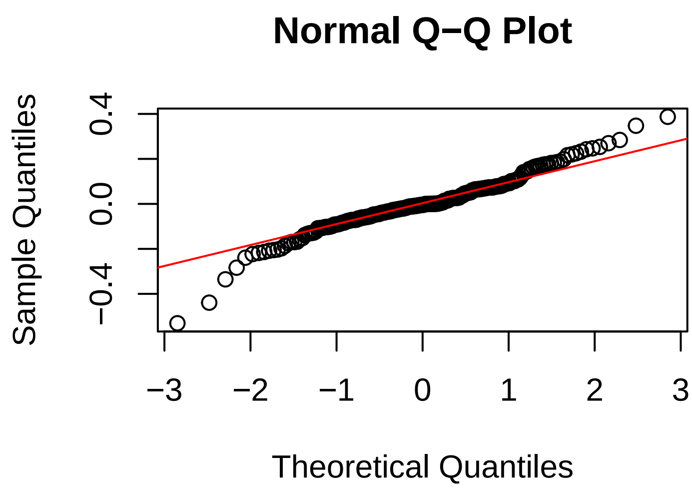
The residual diagnostics are attached in the appendix. In particular, the assumptions of zero-mean, zero-correlation, and homoscedasticity are all met. The assumption of normality was not met; however, since we use the method of least squares, we do not need to meet this requirement.

Part (c) - SARIMA + exogenous variable (i.e., the "import" time series)

Choosing the orders and parameters

We pick the same orders for the ordinary and seasonal differencing as we did in part (b), i.e. we pick d as 1 and D as 1 too. The same ranges for p,q,P,and Q are maintained. We also fit the SARIMAX model on the log-transformed export data. **Here, the exogenous variable is the log-transformed import time series.**

As in part (b), when we pick an optimal model based on test set RMSE and look at the residual diagnostics, we observe that the normality assumption is not met. In particular, observe the qqplot below, which is for the SARIMAX (0,1,0) x (0,1,0) [12] model:



Since the normality assumption is not met, we fit the models using the method of Least Squares instead. The following table shows the top 5 optimal choices for p,P,q, and Q based on test set RMSE when fitting SARIMAX models with least squares:

	Model	RMSE
1	SARIMA (0 1 0) x (0 1 0)[12]	45.66432
13	SARIMA (1 1 0) x (0 1 0)[12]	49.97851
25	SARIMA (2 1 0) x (0 1 0)[12]	50.04115
7	SARIMA (0 1 1) x (0 1 0)[12]	51.12772
31	SARIMA (2 1 1) x (0 1 0)[12]	51.27860

The top 5 choices for p,P, q and Q are shown above, based on the test set RMSE. We pick as our optimal model the one with the lowest test RMSE, which is the SARIMAX (0,1,0)x(0,1,0)[12] model here. The test set RMSE was 45.66.

However, when we check the residual diagnostics for this model, we find that the the ACF plot has one significant spike and the Ljung-Box plot indicates that the residuals are correlated since the p-values are all below the threshold. (See Appendix)

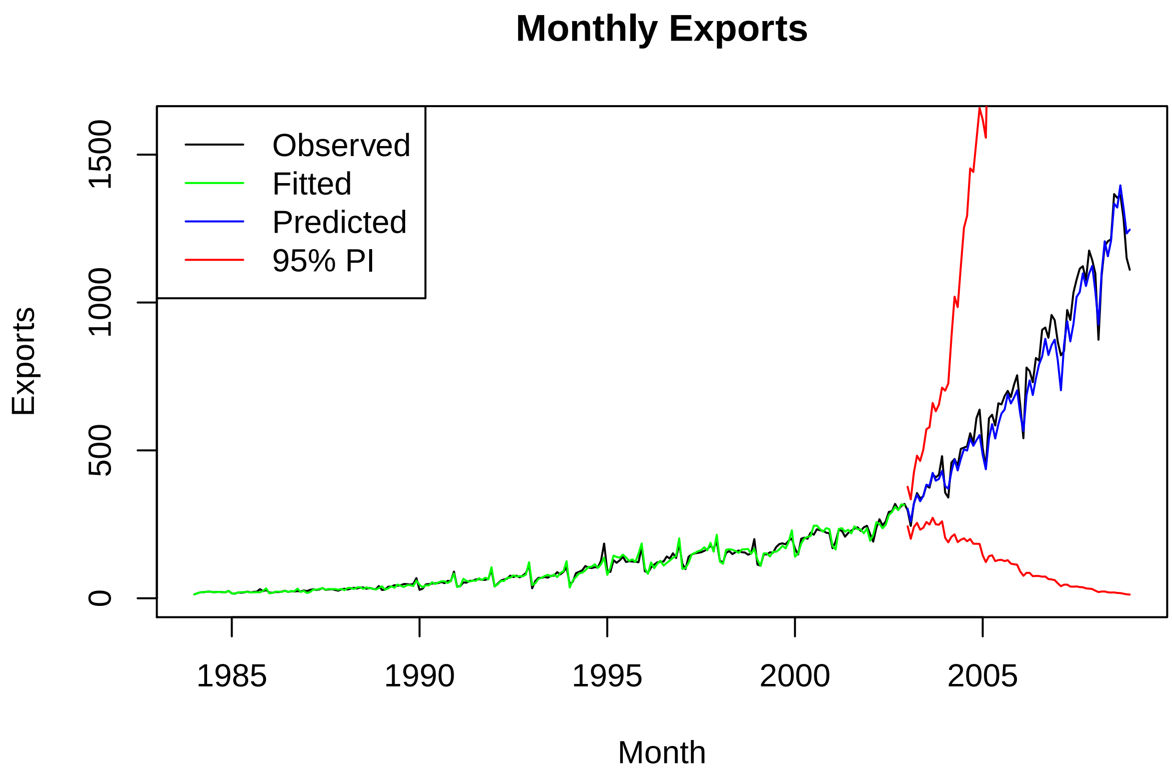
This prompts us to choose the next best model, which is the SARIMAX (1,1,0)x(0,1,0)[12] model. However, we run into the same problem - the ACF plot still has one significant spike and the Ljung-Box plot also indicates that the residuals are correlated since there are some p-values that are below the threshold. **This prompts us to move on to the next best model, which is the SARIMAX (2,1,0)x(0,1,0)[12] model.**

When we check the residual diagnostics for the SARIMAX (2,1,0)x(0,1,0)[12] model, we find that the ACF plot does not have any significant spikes and the Ljung-Box plot confirms that the error residuals are zero-correlated/uncorrelated since the p-values are all above the threshold. A plot of the residuals indicates that they have zero mean. The residuals also display constant variance. The normality assumption is not met, but this is not required since we fit the model using least squares.

All the residual diagostics discussed above are detailed in the appendix.

Below, we show the model's forecasts and their associated 95% prediction intervals in both a tabular and graphical format. For the tabular format, we show only the first year's forecasts; in the graph, we show all the forecasts from January 2003 to December 2008.

	Forecast	Lower	Upper
Jan 2003	303.0150	243.5723	376.9645
Feb 2003	259.1336	200.8613	334.3114
Mar 2003	319.9178	240.8052	425.0217
Apr 2003	350.5768	254.9585	482.0555
May 2003	328.0947	231.9059	464.1803
Jun 2003	345.7677	238.1226	502.0745
Jul 2003	384.0620	258.0086	571.7003
Aug 2003	379.6705	249.2034	578.4419
Sep 2003	423.9737	272.2335	660.2923
Oct 2003	397.5462	249.9665	632.2567
Nov 2003	403.5033	248.6742	654.7318
Dec 2003	430.6011	260.3134	712.2851



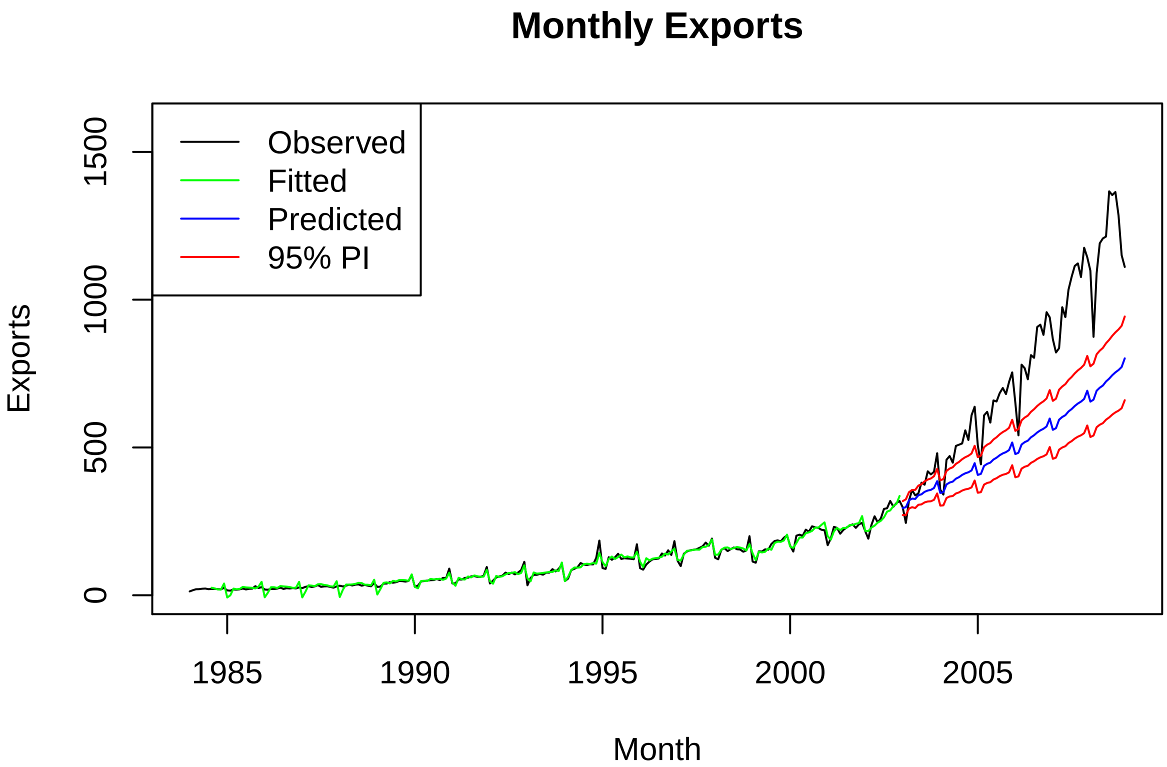
Part (d) - VAR + seasonal indicator (where both “export”" and “import”" time series are treated as endogenous variables)

We consider the import and export time series to be endogenous variables and use Vector Autoregression to fit a model. We choose the order p of the VAR model by evaluating the test set RMSE for each possible value of p and picking the value of p that corresponds to the lowest testing RMSE. Here, we consider a range of 1 to 10 for the value of p. Also, for the seasonal indicator, we specify that the period is 12 since the data corresponds to monthly observations.

```
        Model     RMSE
7   VAR ( 7 ) 295.4727
5   VAR ( 5 ) 296.4001
10 VAR ( 10 ) 299.3785
6   VAR ( 6 ) 300.2649
9   VAR ( 9 ) 302.2728
```

We observe that the best value of p is 7, based on test set RMSE. We pick the VAR(7) model as our optimal choice and show the model’s forecasts and their associated 95% prediction intervals in both a tabular and graphical format. For the tabular format, we show only the first year’s forecasts; in the graph, we show all the forecasts from January 2003 to December 2008.

```
         Forecast    Lower    Upper
Jan 2003 294.8311 271.1761 318.4861
Feb 2003 297.8715 271.2244 324.5187
Mar 2003 321.0228 293.3704 348.6752
Apr 2003 327.0805 297.9652 356.1958
May 2003 325.7919 295.1162 356.4676
Jun 2003 338.4424 305.9804 370.9044
Jul 2003 341.6085 307.4744 375.7426
Aug 2003 349.8311 314.1245 385.5377
Sep 2003 354.3213 317.1928 391.4498
Oct 2003 356.4974 317.9409 395.0539
Nov 2003 362.8487 322.8311 402.8663
Dec 2003 386.0434 344.5801 427.5067
```



Residual Diagnostics

The residual diagnostics are displayed in the appendix. The assumptions of zero-mean, zero-correlation, and homoscedasticity are satisfied. The assumption of normality is not satisfied, but this is not a requirement when using Vector AutoRegression (VAR).

Conclusion

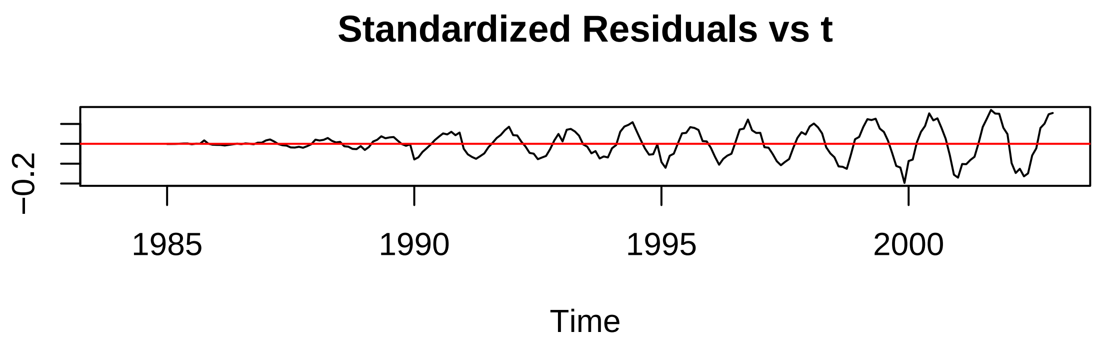
The SARIMAX (2,1,0)x(0,1,0)[12] model performed the best among all four models and had an RMSE of 50.04 on the test set. All the model assumptions were satisifed as well - zero-mean, zero-correlation, and homoscedasticity. Since the model was fit with least squares, the assumption of normality did not need to be satisifed. The residual diagnostics are attached in the appendix.

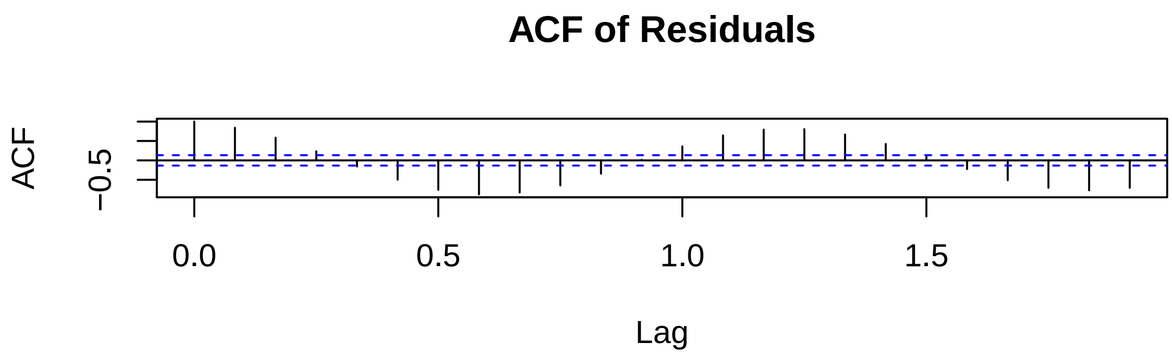
Our recommendation is to use the SARIMAX (2,1,0)x(0,1,0)[12] model for forecasting.

Appendix

Residual Diagnostics for Part (a) - Triple Exponential Smoothing with Additive Seasonality

1) Zero Mean and Zero-Correlation

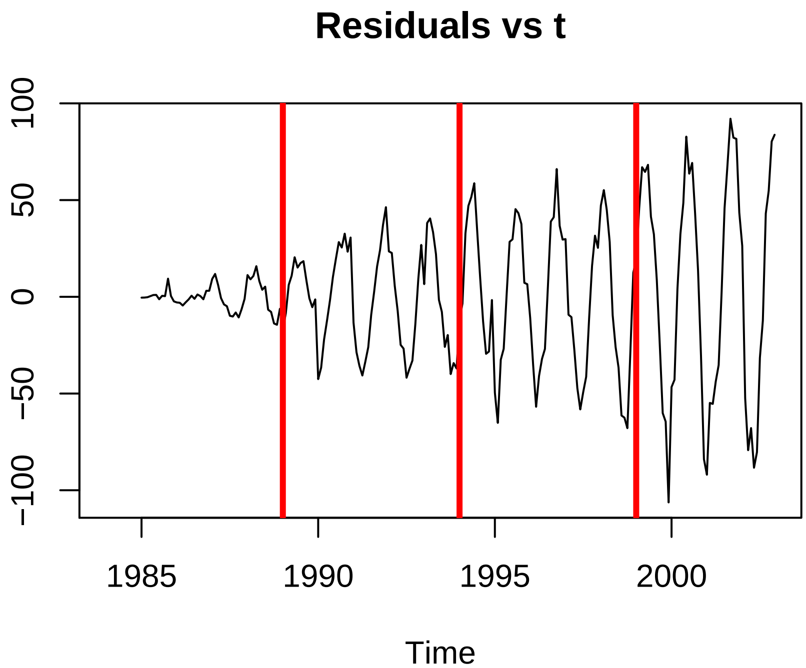




The plot of the residuals indicates that the error terms have a zero mean. However, the ACF plot indicates that the residuals are correlated. This casts doubts on the model's ability to make reliable predictions, even though it minimizes test set RMSE.

2) Homoscedasticity

```
par(mfrow = c(1, 1))
plot(e, main = "Residuals vs t", ylab = "")
abline(v = c(1989, 1994, 1999), lwd = 3, col = "red")
```




```r
group <- c(rep(1, 57), rep(2, 57), rep(3, 57), rep(4, 57))
levene.test(e, group)  #Levene
```

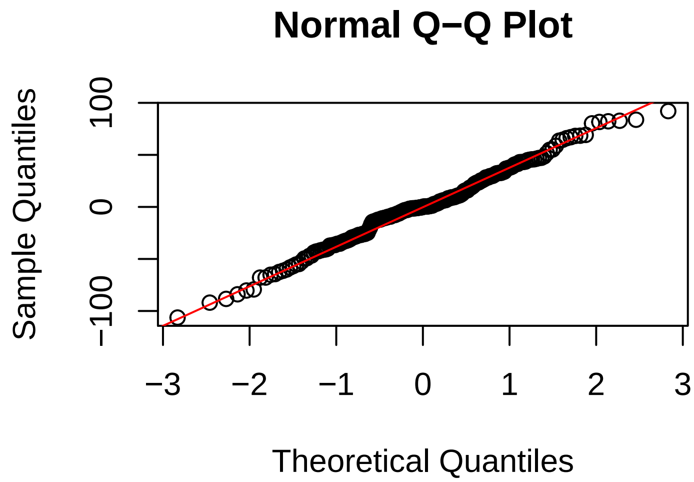
```
    modified robust Brown-Forsythe Levene-type test based on the
    absolute deviations from the median

data:  e
Test Statistic = 54.416, p-value < 2.2e-16
```

```r
bartlett.test(e, group)  #Bartlett
```

```
    Bartlett test of homogeneity of variances

data:  e and group
Bartlett's K-squared = 158.28, df = 3, p-value < 2.2e-16
```

The error terms are clearly heteroscedastic. We also confirm this with a Levene test and Bartlett test, which reject the null hypothesis that the error terms are homoscedastic. This is a result of tuning the smoothing parameters to minimize the test set RMSE, which results in a model that does well on the testing data but which fails to meet model assumptions.

3) Normality

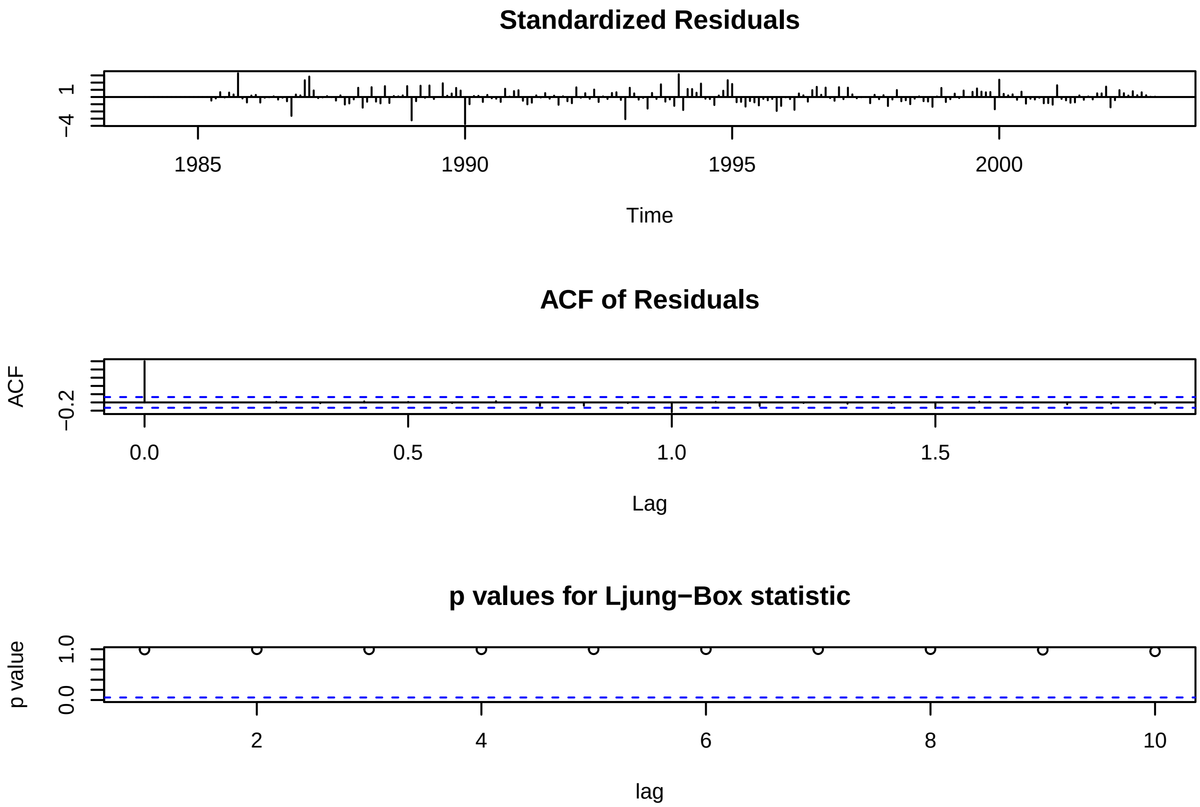



```
    Shapiro-Wilk normality test

data:  e
W = 0.99217, p-value = 0.3039
```

The qqplot indicates normality of the error terms. The Shapiro Test fails to reject the null hypothesis that the residuals are normally distributed. However, since we fit the model using an exponential smoothing method, we do not need to satisfy the assumption of normality.

Residual Diagnostics for Part (b) - SARIMA(2,1,1)x(0,1,0)[12]

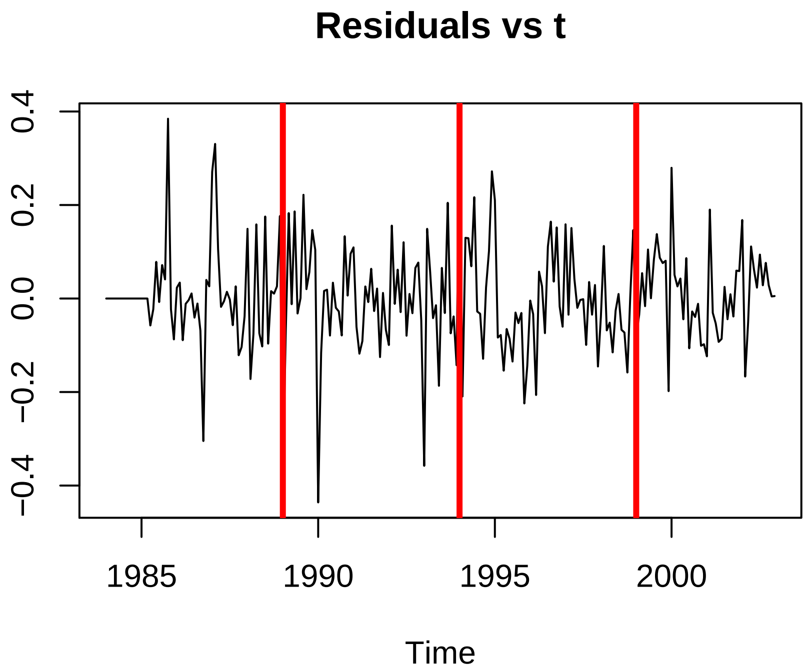
1) Zero Mean and Zero-Correlation



The plot of the residuals indicates that the error terms have a zero mean. The ACF plot does not have any significant spikes and the Ljung-Box plot confirms that the error residuals are zero-correlated/uncorrelated since the p-values are all above the threshold.

2) Homoscedasticity

```
e <- best_model$residuals  # residuals
r <- e/sqrt(best_model$sigma2)  # standardized residuals

par(mfrow = c(1, 1))
plot(e, main = "Residuals vs t", ylab = "")
abline(v = c(1989, 1994, 1999), lwd = 3, col = "red")
```




```
group <- c(rep(1, 57), rep(2, 57), rep(3, 57), rep(4, 57))
levene.test(e, group)  #Levene
```

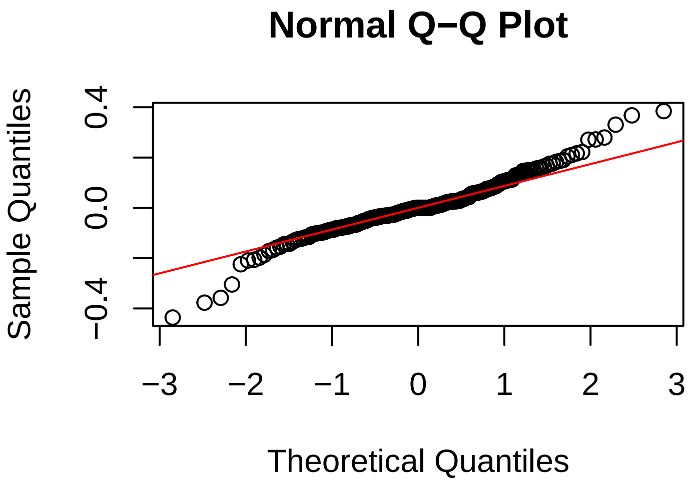
```
    modified robust Brown-Forsythe Levene-type test based on the
    absolute deviations from the median

data:  e
Test Statistic = 1.7716, p-value = 0.1534
```

```
bartlett.test(e, group)  #Bartlett
```

```
    Bartlett test of homogeneity of variances

data:  e and group
Bartlett's K-squared = 6.7899, df = 3, p-value = 0.0789
```

The error terms look homoscedastic.We confirm this with a Levene test and Bartlett test, which fails to reject the null hypothesis that the error terms are homoscedastic. Hence, the assumption of homoscedasticity is met.

3) Normality

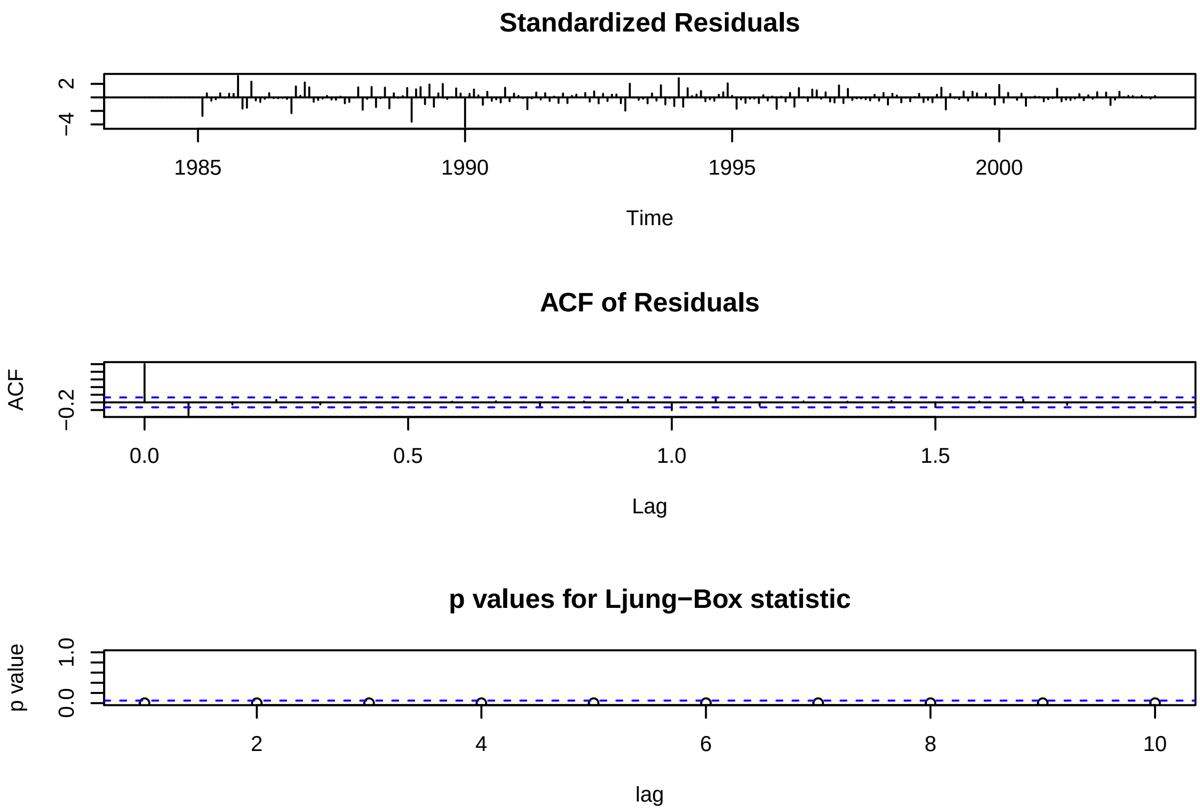



```
    Shapiro-Wilk normality test

data:  e
W = 0.96101, p-value = 7.024e-06
```

The qqplot indicates non-normality of the error terms. The Shapiro Test also indicates that the error terms are not normally distributed. However, since we fit the model using Least Squares, we do not need to satisfy the assumption of normality.

Residual Diagnostics for Part (c) - SARIMAX models

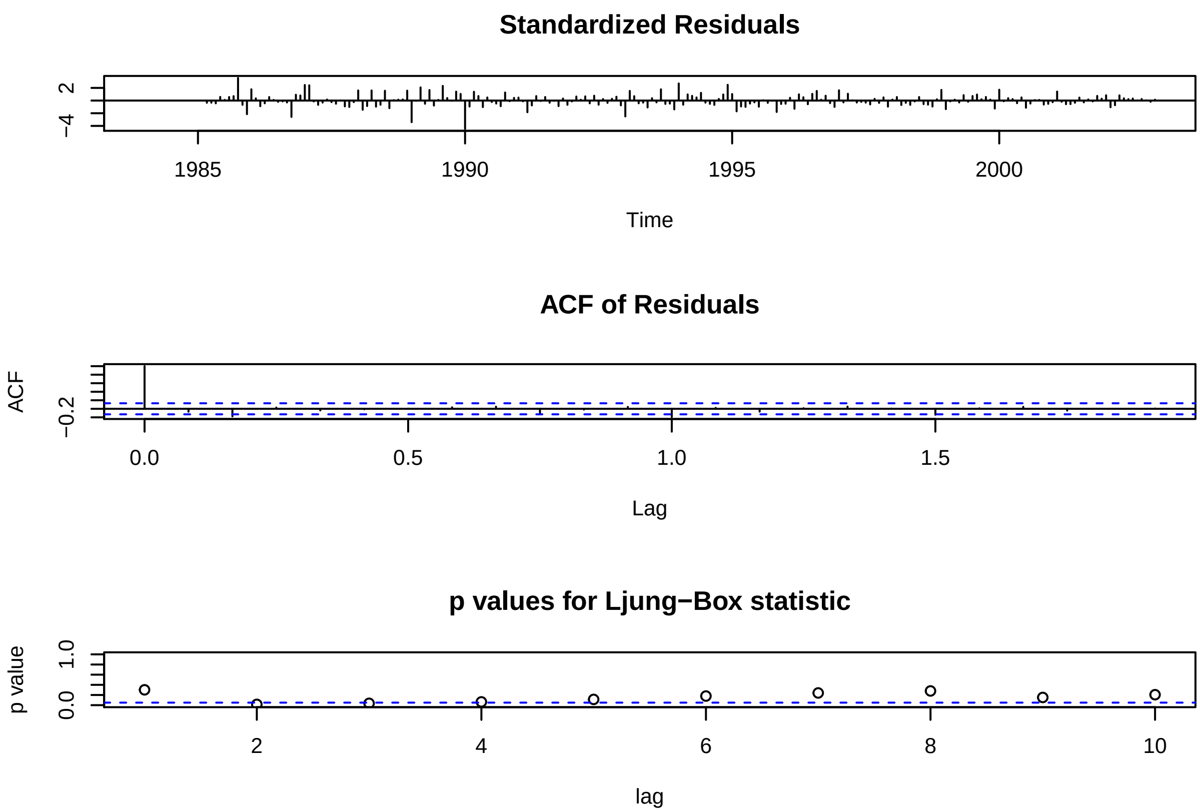
Zero Mean and Zero-Correlation for the SARIMAX (0,1,0)x(0,1,0)[12] model



The plot of the residuals indicates that the error terms have a zero mean. However, the ACF plot has one significant spike and the Ljung-Box plot seems to indicate that the residuals are correlated since the p-values are all below the threshold.

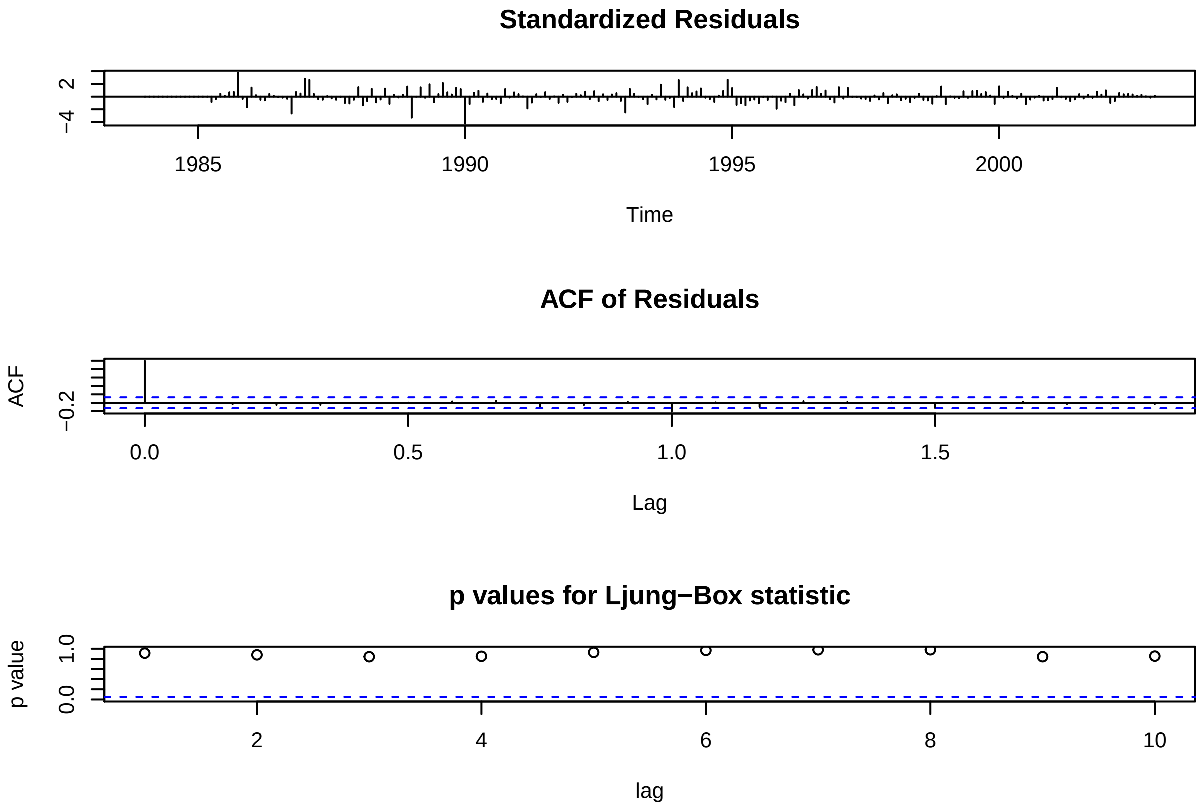
This prompts us to choose the next best model, which was the SARIMAX (1,1,0)x(0,1,0)[12] model. We check the assumption of uncorrelated residuals for this model now:

Zero Mean and Zero-Correlation for the SARIMAX (1,1,0)x(0,1,0)[12] model



The ACF plot has one significant spike and the Ljung-Box plot seems to indicate that the residuals are correlated since there are some p-values that are below the threshold. **This prompts us to move on to the next best model, which is the SARIMAX (2,1,0)x(0,1,0)[12] model.** We check the diagnostics for this model below.

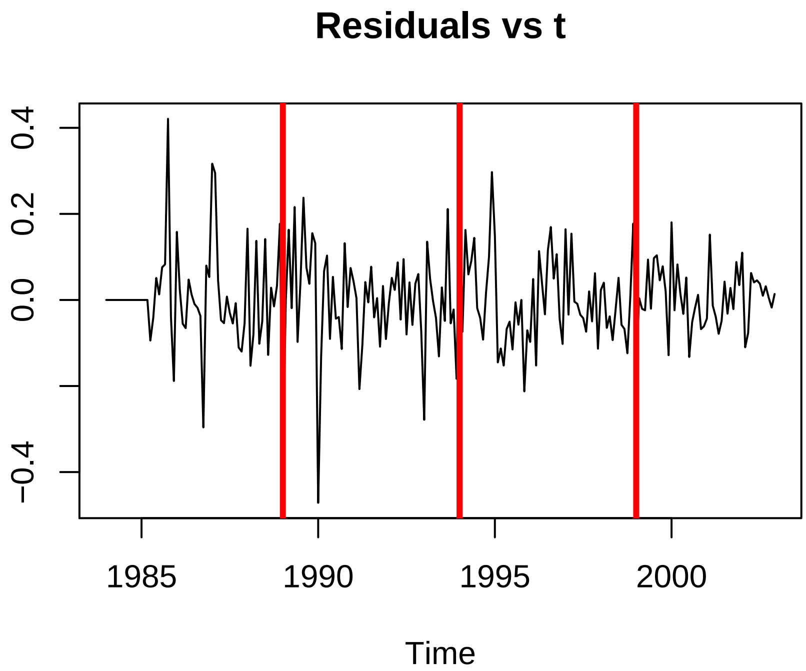
1) Zero Mean and Zero-Correlation for the SARIMAX (2,1,0)x(0,1,0)[12] model



The ACF plot does not have any significant spikes and the Ljung-Box plot confirms that the error residuals are zero-correlated/uncorrelated since the p-values are all above the threshold. Hence, we pick this model as our optimal choice.

2) Homoscedasticity

```r
e <- best_model$residuals  # residuals
r <- e/sqrt(best_model$sigma2)  # standardized residuals

par(mfrow = c(1, 1))
plot(e, main = "Residuals vs t", ylab = "")
abline(v = c(1989, 1994, 1999), lwd = 3, col = "red")
```




```r
group <- c(rep(1, 57), rep(2, 57), rep(3, 57), rep(4, 57))
levene.test(e, group)  #Levene
```

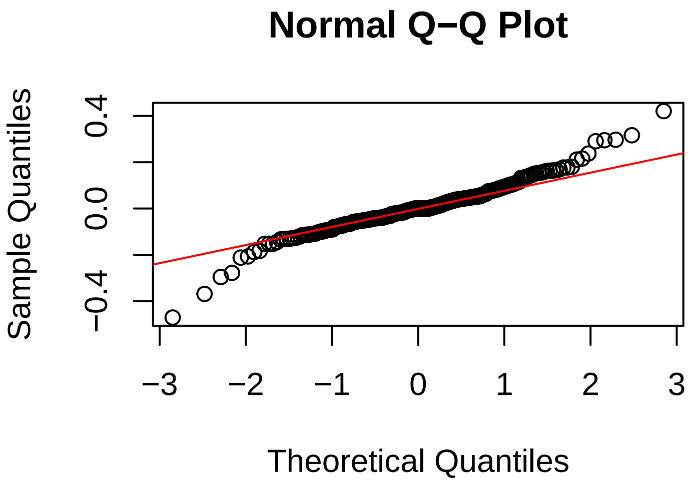
```
    modified robust Brown-Forsythe Levene-type test based on the
    absolute deviations from the median

data:  e
Test Statistic = 2.2564, p-value = 0.08269
```

```r
bartlett.test(e, group)  #Bartlett
```

```
    Bartlett test of homogeneity of variances

data:  e and group
Bartlett's K-squared = 16.788, df = 3, p-value = 0.0007814
```

The error terms look homoscedastic. We confirm this with a Levene test which fails to reject the null hypothesis that the error terms are homoscedastic. The Bartlett test, on the other hand, rejects the null hypothesis. This might be attributed to the fact that the Bartlett test is sensitive to departures from normality. Hence, we can still conclude that the assumption of homoscedasticity is met.

3) Normality

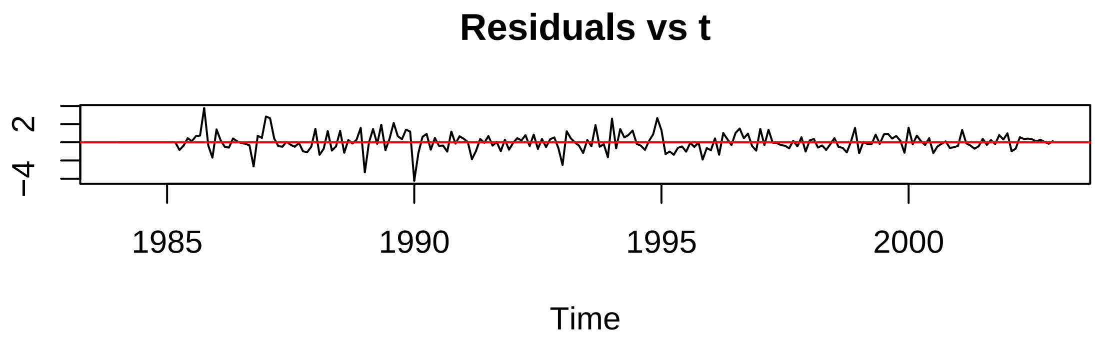


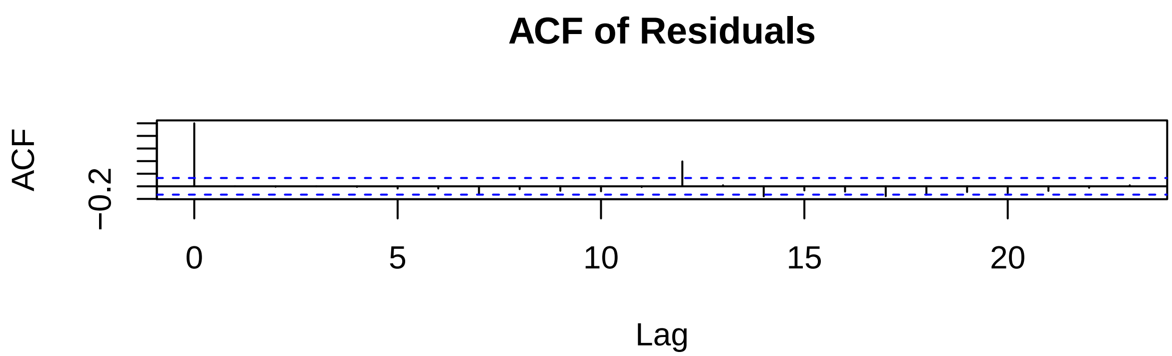

```
    Shapiro-Wilk normality test

data:  e
W = 0.95802, p-value = 3.181e-06
```

The qqplot indicates non-normality of the error terms. The Shapiro Test also indicates that the error terms are not normally distributed. However, since we fit the model using Least Squares, we do not need to satisfy the assumption of normality.

Residual Diagnostics for Part (d) - VAR(7) + seasonal indicator

1) Zero Mean and Zero-Correlation

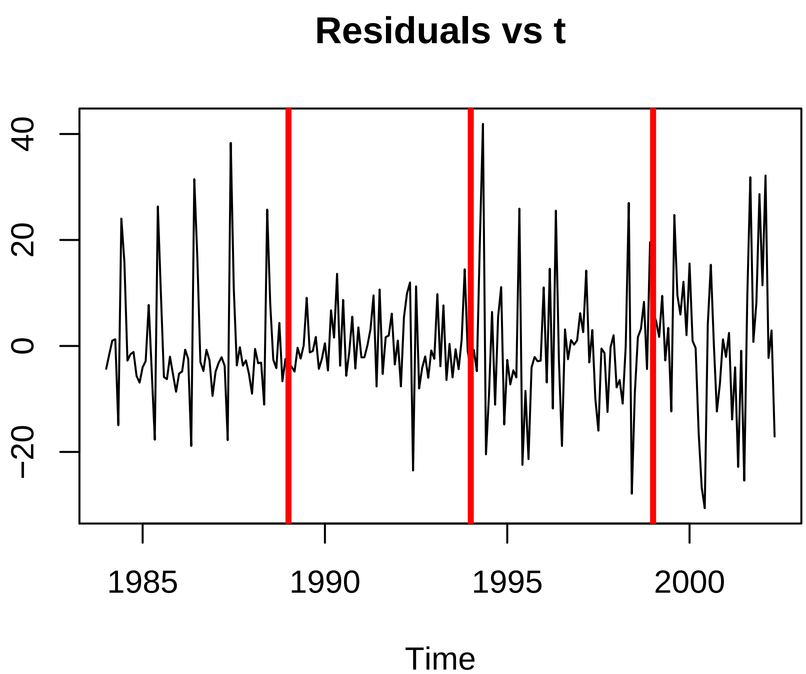




The plot of the residuals indicates that the error terms have a zero mean. The ACF plot does not show any significant spikes. The model assumptions of zero mean and zero correlation are both satisfied.

2) Homoscedasticity

```
par(mfrow = c(1, 1))
plot(ts(e, start = 1984, frequency = 12), main = "Residuals vs t",
    ylab = "")
abline(v = c(1989, 1994, 1999), lwd = 3, col = "red")
```




```
group <- c(rep(1, 57), rep(2, 57), rep(3, 57), rep(4, 50))
levene.test(e, group)  #Levene
```

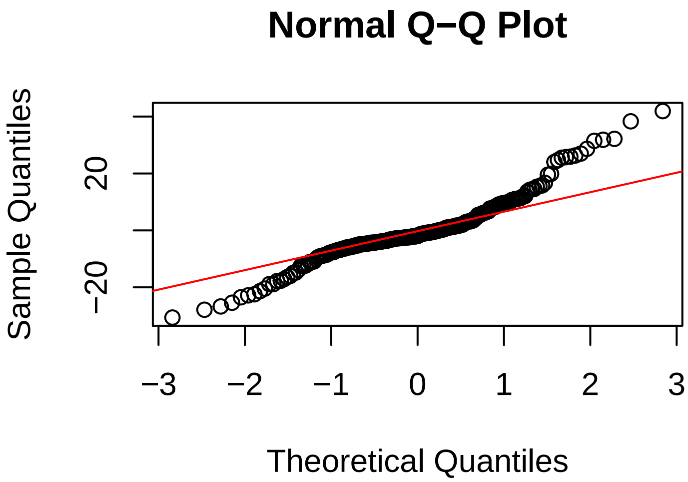
```

    modified robust Brown-Forsythe Levene-type test based on the
    absolute deviations from the median

data:  e
Test Statistic = 4.3633, p-value = 0.005236
```

```
bartlett.test(e, group)  #Bartlett
```

```

    Bartlett test of homogeneity of variances

data:  e and group
Bartlett's K-squared = 32.491, df = 3, p-value = 4.123e-07
```

The error terms look mostly homoscedastic; however, in the second partition, there seems to be lesser variance than in other regions. The Levene test and Bartlett test both reject the null hypothesis that the error terms are homoscedastic. However, we will state that the model assumption of homoscedasticity is still maintained, despite some slight departures.

3) Normality




```
    Shapiro-Wilk normality test

data:  e
W = 0.93821, p-value = 4.703e-08
```

The qqplot indicates non-normality of the error terms. The Shapiro Test rejects the null hypothesis that the residuals are normally distributed. However, since we fit the model using Vector AutoRegression (VAR) , we do not need to satisfy the assumption of normality.