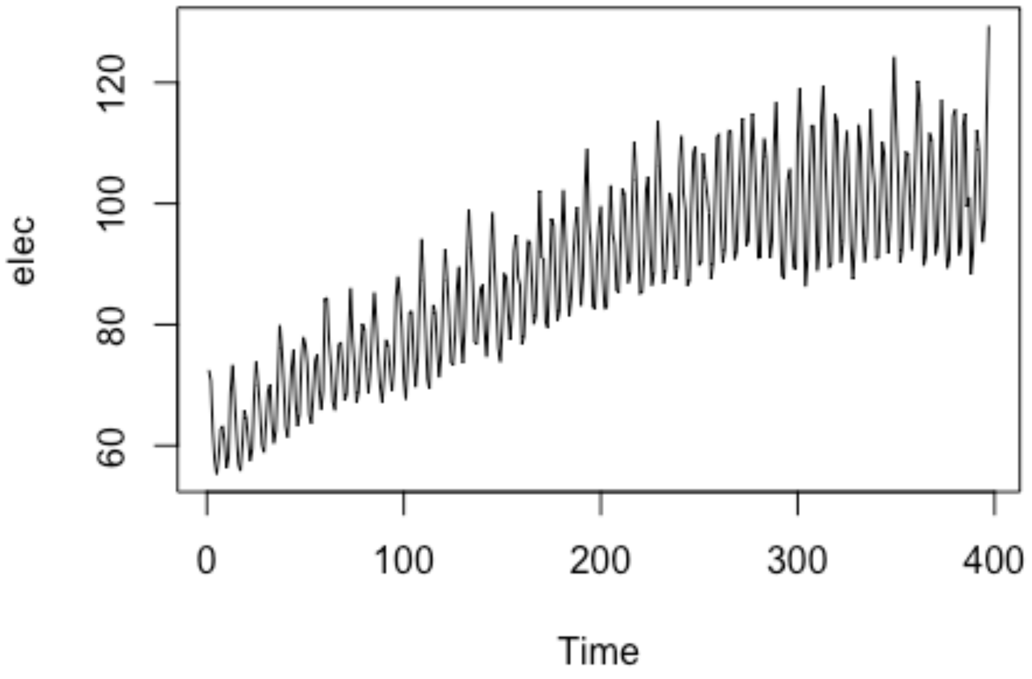
Time Series Project

Farida Simaika, Joyce Wassef, Katia Gabriel, Marina Guindy

5/21/2023

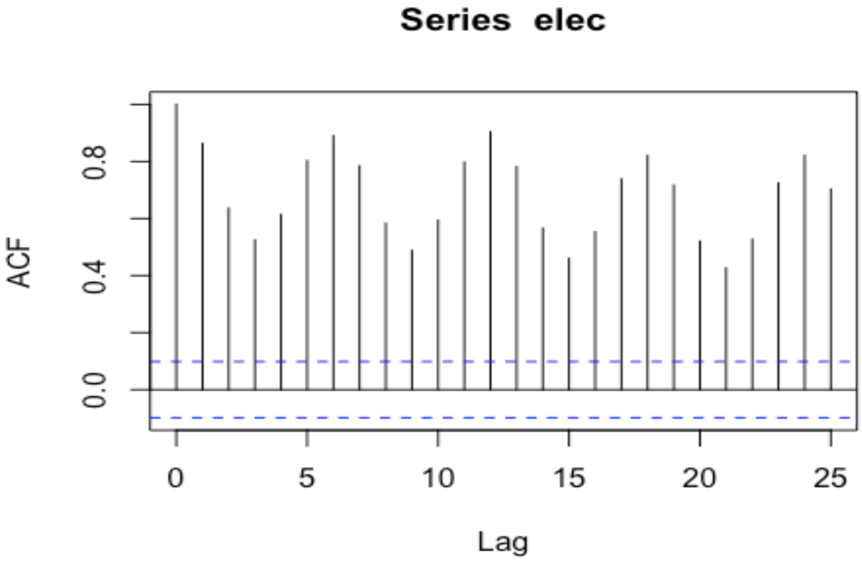
Our focus is on understanding and predicting electricity consumption patterns. This project holds significance as it aims to empower the factory with insights to optimize energy use, reduce costs, and validate key assumptions in our forecasting models.

```r
elec<-scan("electricity.csv",skip=1)
plot.ts(elec)
```

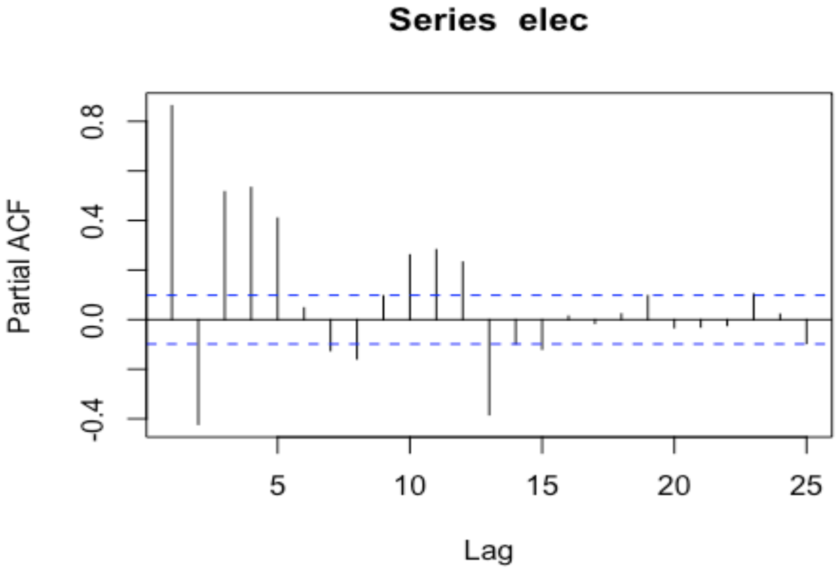


We first started by examining the series to see if there is trend or a pattern in time. After plotting the intial series, it can be observed that there is an upward trend, which suggests that the series may not be stationary.

```
acf(elec)
```



```
pacf(elec)
```



We additionally observed the ACF and PACF to further support the initial statement we mentioned. From the ACF we can see that it is slowly decaying, and they do not die quickly, and this also suggest that our first statement might be true and that this series is non-stationary. However human eye can be deceiving therefore we decided to perform the Dickey Fuller's test to confirm our initial statement.

```
library(urca)
library(tseries)
elec1<-as.matrix(elec)
df<-ur.df(elec1,type="trend",lag=1)
adf.test(elec1,k=1)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  elec1
## Dickey-Fuller = -22.937, Lag order = 1, p-value = 0.01
## alternative hypothesis: stationary
```

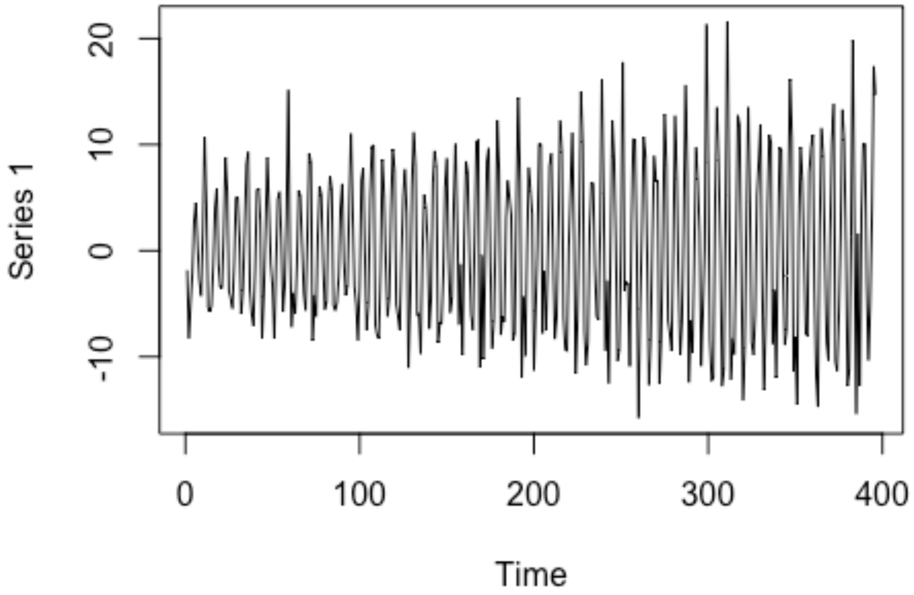
```
summary(df)
```

```
##
## ###############################################
## # Augmented Dickey-Fuller Test Unit Root Test #
## ###############################################
##
## Test regression trend
##
##
## Call:
## lm(formula = z.diff ~ z.lag.1 + 1 + tt + z.diff.lag)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.2844  -3.2133  -0.0126   3.2279  14.1703
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 48.195510   2.156320   22.35   <2e-16 ***
## z.lag.1     -0.723885   0.031560  -22.94   <2e-16 ***
## tt           0.081131   0.004078   19.89   <2e-16 ***
## z.diff.lag   0.739536   0.034538   21.41   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 4.721 on 391 degrees of freedom
## Multiple R-squared:  0.6341, Adjusted R-squared:  0.6312
## F-statistic: 225.8 on 3 and 391 DF,  p-value: < 2.2e-16
##
##
## Value of test-statistic is: -22.9369 175.4757 263.1096
##
## Critical values for test statistics:
##       1pct  5pct 10pct
## tau3 -3.98 -3.42 -3.13
## phi2  6.15  4.71  4.05
## phi3  8.34  6.30  5.36
```

In general a series might not be stationary for two reasons; a pattern in time or a trend. The DF test statistic is a test statistic that helps us know whether the series is non-stationary or stationary. We can observe z.lag.1 to see whether the series is stationary or not, and from tt we can observe whether there is a trend or not. Since the p-value is less than alpha for both z.lag.1 and tt, this suggests that we should reject H0 for both test statistics, which states that the series is stationary (it has no pattern in time). but there is a trend, therefore we need to take the first difference.

```
elec_diff1<-diff(elec1,differences=1)
plot.ts(elec_diff1)
```



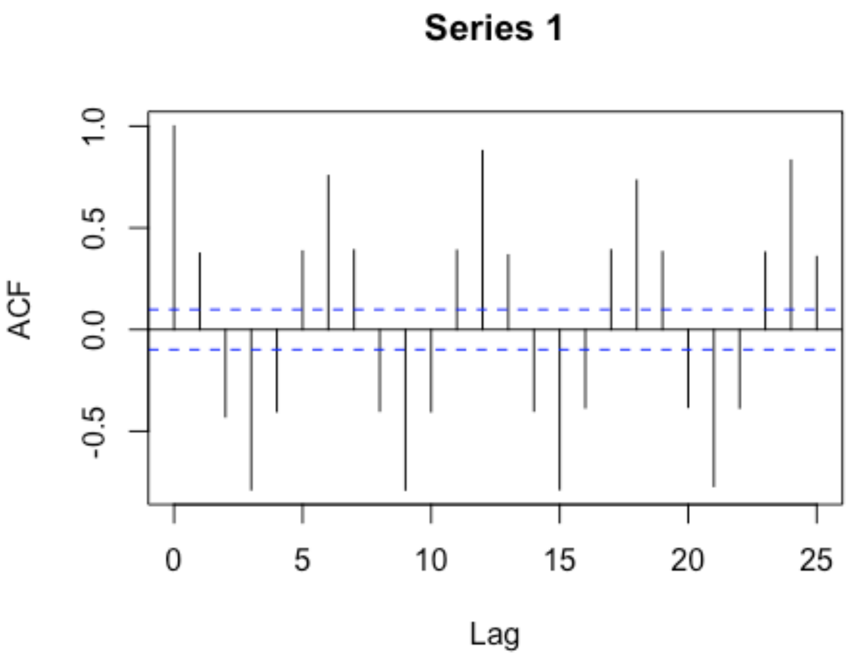
```
df1<-ur.df(elec_diff1,type="trend",lag=1)
summary(df1)
```

```
## 
## ############################################### 
## # Augmented Dickey-Fuller Test Unit Root Test # 
## ############################################### 
## 
## Test regression trend 
## 
## 
## Call:
## lm(formula = z.diff ~ z.lag.1 + 1 + tt + z.diff.lag)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -24.1099  -2.3346   0.6136   3.1852  17.3903 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  0.111095   0.546902   0.203    0.839    
## z.lag.1     -1.046409   0.042589 -24.570   <2e-16 ***
## tt           0.000158   0.002391   0.066    0.947    
## z.diff.lag   0.669615   0.038049  17.599   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

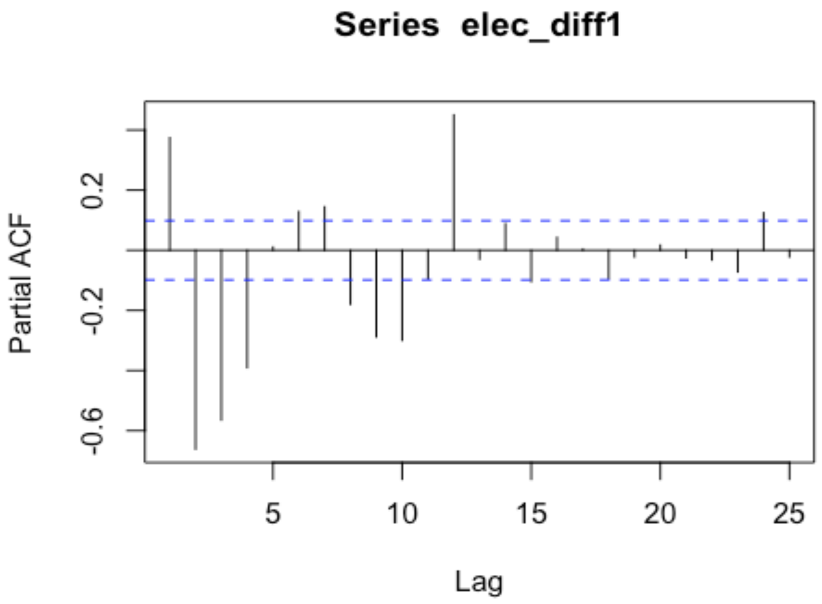
```
##
## Residual standard error: 5.397 on 390 degrees of freedom
## Multiple R-squared:  0.6154, Adjusted R-squared:  0.6124
## F-statistic:   208 on 3 and 390 DF,  p-value: < 2.2e-16
##
##
## Value of test-statistic is: -24.5701 201.2438 301.8479
##
## Critical values for test statistics:
##       1pct  5pct 10pct
## tau3 -3.98 -3.42 -3.13
## phi2  6.15  4.71  4.05
## phi3  8.34  6.30  5.36
```

After plotting the series again we can see that the plot now suggests that it became stationary, and from the DF test we can see that the p-value for z.lag.1 is less than alpha therefore we are going to reject the null hypothesis and conclude that the series is stationary, and for the tt the p-value is greater than alpha, therefore we will fail to reject the null hypothesis and conclude that the series has no trend.

```
acf(elec_diff1)
```



```
pacf(elec_diff1)
```



Since the series is now stationary, we can plot the ACF and PACF. After plotting both graphs, we can observe that both plots are decaying which suggests that the suitable model for this series is ARMA, and since we took the first difference we're going to try different ARIMA Models and choose the one with the lowest AIC. The AIC (Akaike's Information Criterion) is a measure of loss of information due to additional parameters. Therefore, a lower AIC means that the model was able to fit the model adequately using fewer parameters.

```
### Different models
m0<-arima(elec,order = c(0,1,1))
m1<-arima(elec, order=c(0,1,4))
#m2<-arima(elec, order=c(3,1,4))
m3<-arima(elec, order=c(3,1,0))
m4<-arima(elec, order=c(1,1,1))
m5<-arima(elec, order=c(1,1,3))
m6<-arima(elec, order=c(4,1,0))
m7<-arima(elec, order=c(4,1,1))
#m8<-arima(elec, order=c(4,1,2))
#m9<-arima(elec, order=c(4,1,3))
#m10<-arima(elec, order=c(4,1,4))

cat("ARIMA (0,1,1)")
```

```
## ARIMA (0,1,1)
```

```
m0
```

```
##
## Call:
## arima(x = elec, order = c(0, 1, 1))
##
## Coefficients:
##          ma1
##       0.5854
## s.e.  0.0361
##
## sigma^2 estimated as 43.53:  log likelihood = -1309.28,  aic = 2622.55
```

```
cat("ARIMA (0,1,4)")
```

```
## ARIMA (0,1,4)
```

```
m1
```

```
##
## Call:
## arima(x = elec, order = c(0, 1, 4))
##
## Coefficients:
##          ma1      ma2      ma3      ma4
##       0.3225  -0.2905  -0.5203  -0.3563
## s.e.  0.0430   0.0556   0.0355   0.0611
##
## sigma^2 estimated as 27.04:  log likelihood = -1215.71,  aic = 2441.43
```

```
cat("ARIMA (3,1,0)")
```

```
## ARIMA (3,1,0)
```

```
m3
```

```
##
## Call:
## arima(x = elec, order = c(3, 1, 0))
##
## Coefficients:
##          ar1      ar2      ar3
##       0.2344  -0.3077  -0.5854
## s.e.  0.0408   0.0396   0.0409
##
## sigma^2 estimated as 18.96:  log likelihood = -1145.81,  aic = 2299.62
```

```
cat("ARIMA (1,1,1)")
```

```
## ARIMA (1,1,1)
```

```
m4
```

```
##
## Call:
## arima(x = elec, order = c(1, 1, 1))
##
## Coefficients:
##          ar1     ma1
##       0.0732  0.5450
## s.e.  0.0733  0.0559
##
## sigma^2 estimated as 43.43:  log likelihood = -1308.79,  aic = 2623.59
```

```
cat("ARIMA (1,1,3)")
```

```
## ARIMA (1,1,3)
```

```
m5
```

```
##
## Call:
## arima(x = elec, order = c(1, 1, 3))
##
## Coefficients:
##          ar1     ma1      ma2      ma3
##       0.1226  0.0700  -0.4631  -0.4799
## s.e.  0.0771  0.0609   0.0448   0.0569
##
## sigma^2 estimated as 28.37:  log likelihood = -1225.19,  aic = 2460.38
```

```
cat("ARIMA (4,1,0)")
```

```
## ARIMA (4,1,0)
```

```
m6
```

```
##
## Call:
## arima(x = elec, order = c(4, 1, 0))
##
## Coefficients:
##           ar1      ar2      ar3      ar4
##       -0.0298  -0.4448  -0.4847  -0.4494
## s.e.   0.0452   0.0381   0.0379   0.0453
##
## sigma^2 estimated as 15.17:  log likelihood = -1102.09,  aic = 2214.18
```

```
cat("ARIMA (4,1,1)")
```

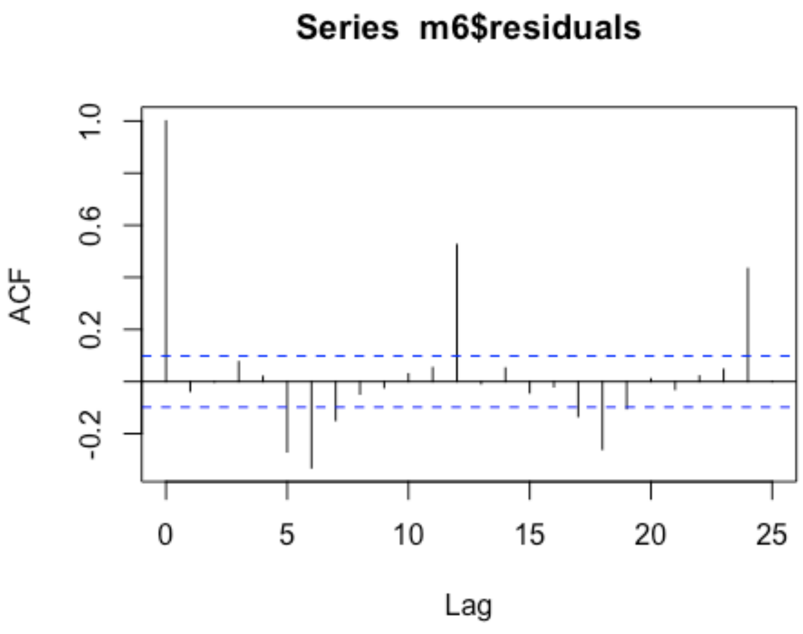
```
## ARIMA (4,1,1)
```

```
m7
```

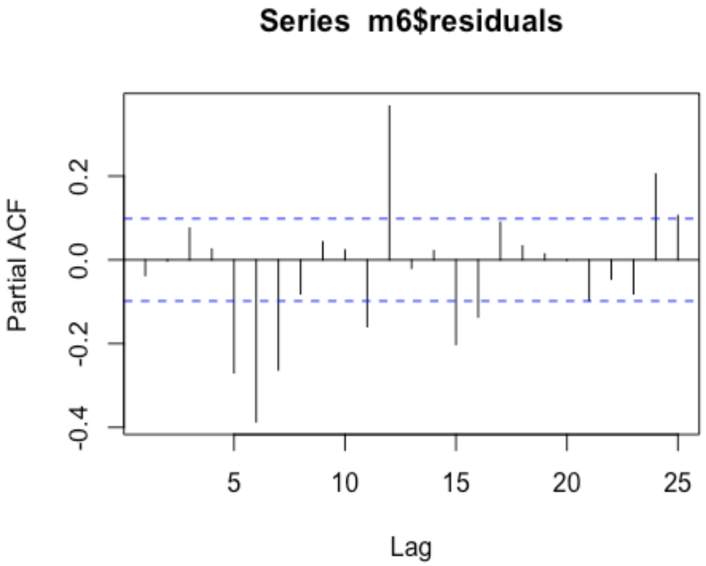
```
## 
## Call:
## arima(x = elec, order = c(4, 1, 1))
## 
## Coefficients:
##          ar1      ar2      ar3      ar4      ma1
##       0.0640  -0.4679  -0.4552  -0.3955  -0.1192
## s.e.  0.0882   0.0437   0.0458   0.0657   0.0917
## 
## sigma^2 estimated as 15.1:  log likelihood = -1101.27,  aic = 2214.54
```

According to the AIC, it is suggested that m6 is the best model for this series which is ARIMA(4,1,0), as it has the lowest AIC. Since we knew the order of the model we now need to validate the NICE assumptions which are; Normality of residuals, Independence of residuals, residuals have Constant variance, and Expectation of the residuals are equal to 0.

```
acf(m6$residuals)
```

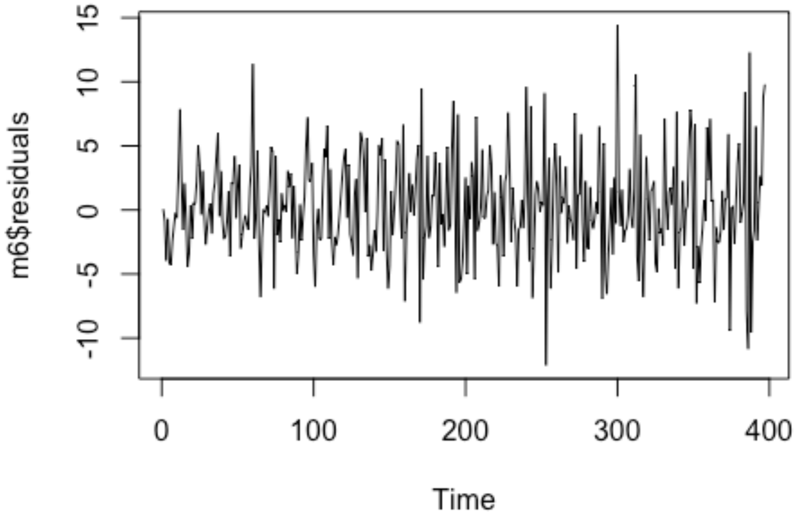


```
pacf(m6$residuals)
```

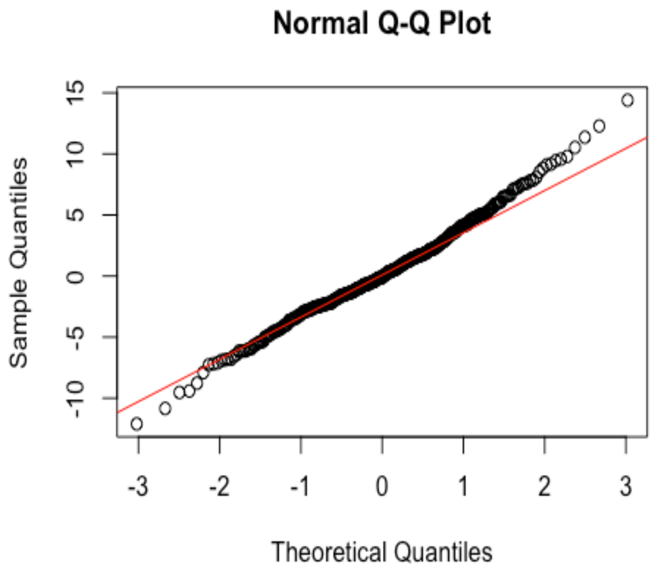


According to the ACF of residuals, it can be noticed that there are significant autocorrelation coefficients at lags 5 and 6, which might indicate dependence of residuals.

```
plot.ts(m6$residuals)
```



```
qqnorm(m6$residuals)
qqline(m6$residuals, col="red")
```



The time series of residuals shows that they have 0 mean, which validates the assumption of zero expectation. Additionally, the plot does not show any pattern for the variance of residuals, which validates the constant varince assumption. The Normal QQ-plot of residuals shows that the residuals are very close to the straight line, which indicates that they are normally distributed.

```
Box.test(m6$residuals,lag=20,fitdf=16)

##
##  Box-Pierce test
##
## data:  m6$residuals
## X-squared = 236.58, df = 4, p-value < 2.2e-16
```

To check whether our conclusion about the independence of residuals assumption is valid or not, we can use the Lung-Box-Pierce test. This test basically tests the null hypothesis: all autocorrelation coefficients of residuals are equal to 0 versus the alternative hypothesis: at least one autocorrelation coefficient is not equal to 0. According to Lung-Box-Pierce Test, the null hypothesis of independent residuals can be rejected, which agrees with the conclusion that was previously made from the ACF and PACF plots, which is that residuals are not independent.

```
library(forecast)
forecast=elec[c(394,395,396,397)]
series=elec[-c(394,395,396,397)]
m6<-arima(series, order=c(4,1,0))
predict(m6,n.ahead=4)
```

```
## $pred
## Time Series:
## Start = 394
## End = 397
## Frequency = 1
## [1]  91.04241  92.93096 102.65825 109.77594
##
## $se
## Time Series:
## Start = 394
## End = 397
## Frequency = 1
## [1] 3.853718 5.337419 5.700223 5.706730
```

```
forecast
```

```
## [1]  93.6137  97.3359 114.7212 129.4048
```

Now that the residuals' assumptions have been checked, we can use our model for forecasting. Accordingly, we can remove the last 4 observations and check whether the chosen model is able to provide accurate forecasts for these observations. The output of the R function "predict" gives forecasts that are close enough to the observed values. It can be noticed that the standard error or the uncertainty increases as the lags increase, or as we go forward in time, which means that the prediction becomes less accurate. This can also be noticed by the increase in the difference between the forecasted value and the observed value. Another reason why the forecast may not be very accurate is that the assumption of independent residuals was not validated.