4/7/2020

Advanced topics in statistics

Time Series

Brikena Kokalari

ASSIGNMENT I

Contents

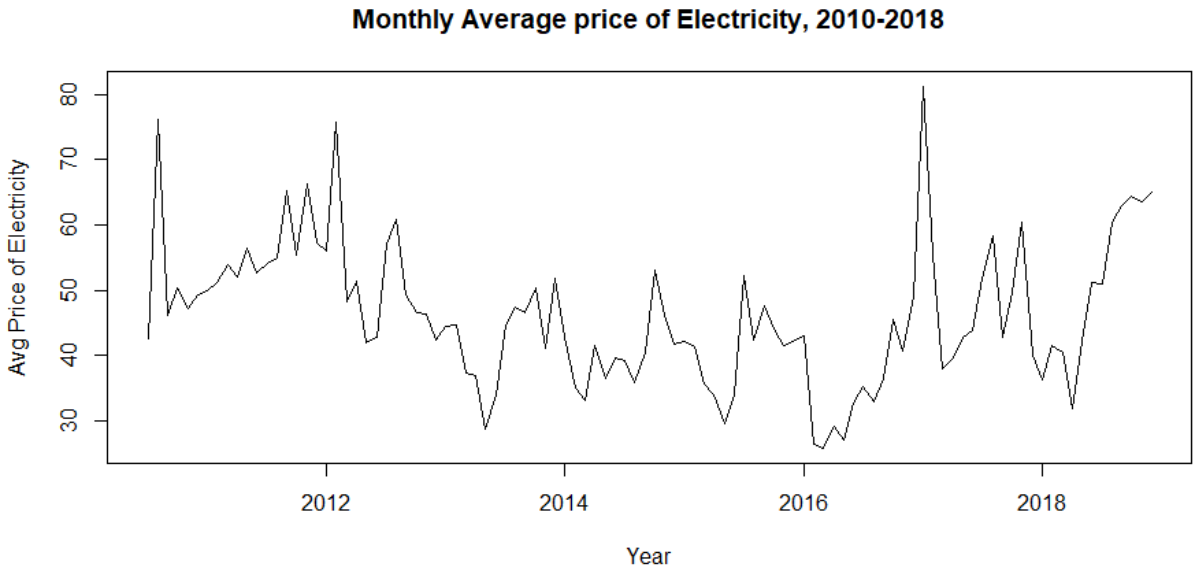
Explanatory Data Analysis 2
Identification step - Autocorrelation and Partial-Autocorrelation 5
ARIMA Model 6
Estimation step - Best Model Selection 7
Diagnostic checking step 8
Test of Autocorrelation of Residuals and Zero mean 8
Normality Test of residuals 9
Test of Heteroskedasticity of Residuals 9
Forecasting step 10

Explanatory Data Analysis

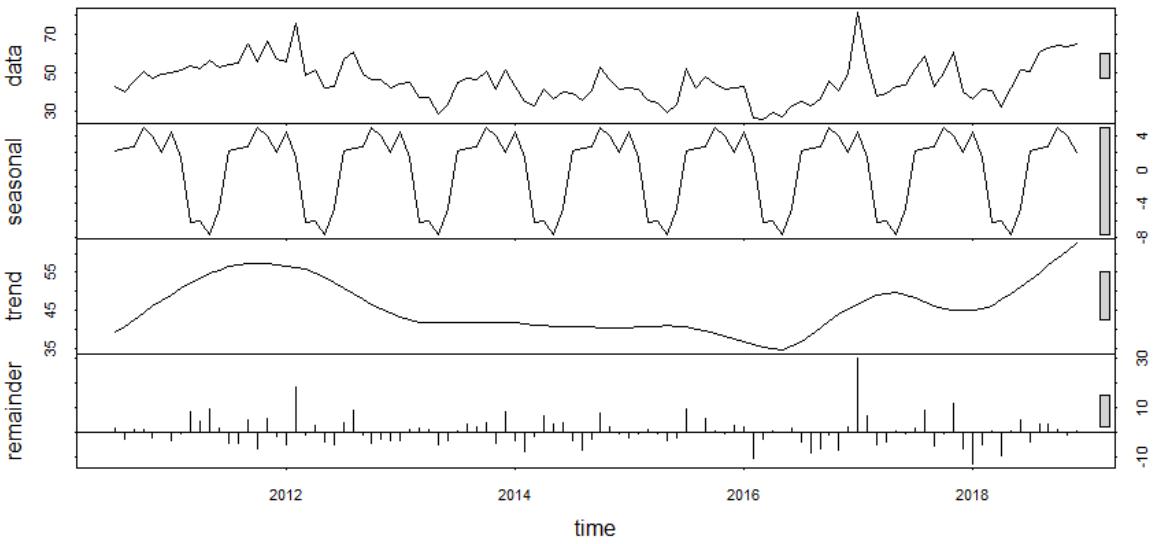
The data used for this time series analysis concern daily prices of electricity from 21/07/2010 to 31/12/2018. The objective of this analysis and modelling is to review time series theory and experiment with R packages in order to predict next six months prices of electricity.

After loading the data and having a summary of it, we can observe that the maximum value of the price of electricity is 1147,96, while the price in the 3rd quantile seems to be 53,74. We considered that the maximum value of price is an outlier and eliminated it. Moreover, the price has 3 negative values which means that the supply of electricity those days was bigger than the demand. It may imply that the cost of storage was too big.

Since we have daily data and we want to forecast the month prices of next six months, we converted the data to monthly by calculating the average price. Another alternative would have been analyzing the daily data and then making predictions. That would bigger variability since monthly data are just the average. It is essential to analyze the trends prior to building any kind of time series model. The details we are interested in pertains to any kind of trend, seasonality or random behavior in the series. What better way to do so than visualize the Time Series?



Decomposing a time series means separating it into its constituent components, which are often a trend component and a random component, and if the data is seasonal, a seasonal component. Seasonal trend decomposition using Loess(STL) is an algorithm that was developed to help to divide up a time series into three components namely: the trend, seasonality and remainder. The four graphs are the original data, seasonal component, trend component and the remainder and this shows the periodic seasonal pattern extracted out from the original data and the trend that moves around between 35 and 65. There is a bar at the right hand side of each graph to allow a relative comparison of the magnitudes of each component.



So on the upper panel, we might consider the bar as 1 unit of variation. The bar on the trend panel is only slightly larger than that on the data panel, indicating that the trend signal is large relative to the variation in the data. In other words, if we shrunk the trend panel such that the box became the same size as that in the data panel, the range of variation on the trend panel would be similar to but slightly smaller than that on the data panel.

Now considering the seasonal panel; the grey box is now much larger than either of the ones on the data or trend panel, indicating the variation attributed to the seasonality is much smaller than the trend component and consequently only a small part of the variation in the data series. The variation attributed to the seasonality is considerably smaller than the stochastic component (the remainders). As such, we can deduce that these data do not exhibit seasonality.

In order to proceed with further analysis, we need the time series data to be stationary. A stationary time series has the conditions that the mean, variance and covariance are not functions of time. A time series is said to be stationary if it holds the following conditions true: The mean value of time-series is constant over time, which implies, the trend component is nullified, and the variance does not increase over time.

We will use the Augmented Dickey-Fuller Test to test the stationarity from the tseries R package.

First set the hypothesis test:

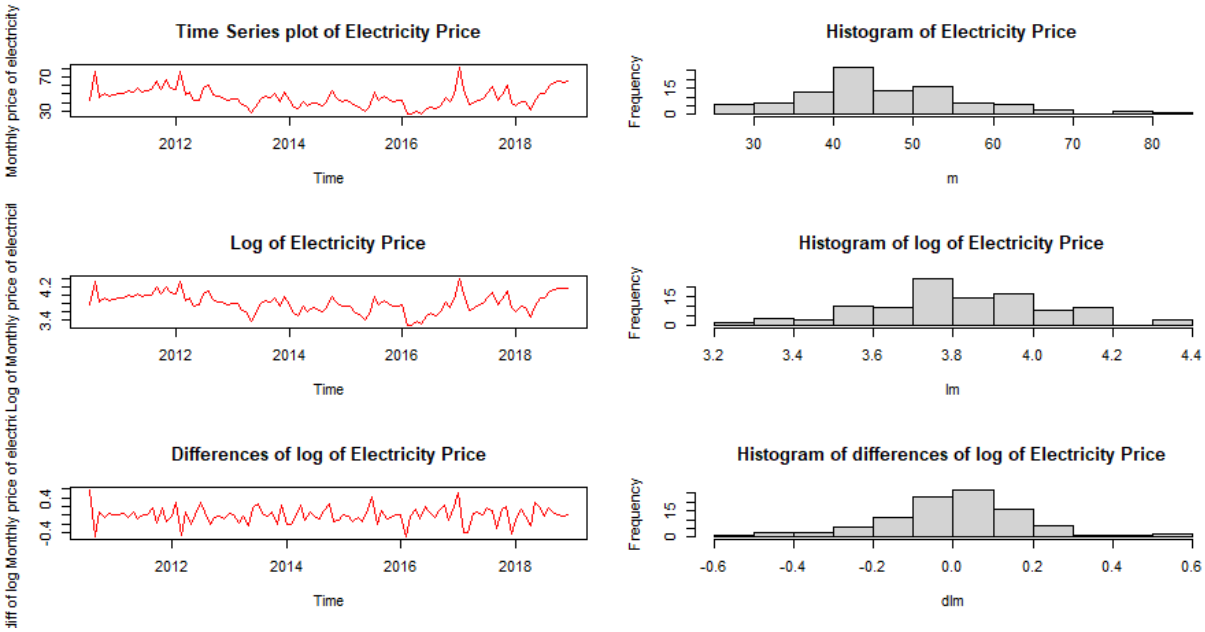
- The null hypothesis: time series is non stationary
- The alternative hypothesis: time series is stationary

As a rule of thumb, where the p-value is less than 0.05, we strong evidence against the null hypothesis, so we reject the null hypothesis. Augmented Dickey-Fuller Test gives a p-value of 0.3867, so we do have enough evidence to reject null hypothesis of non-stationarity. From the above p-value, we concluded that the time series is non-stationary.

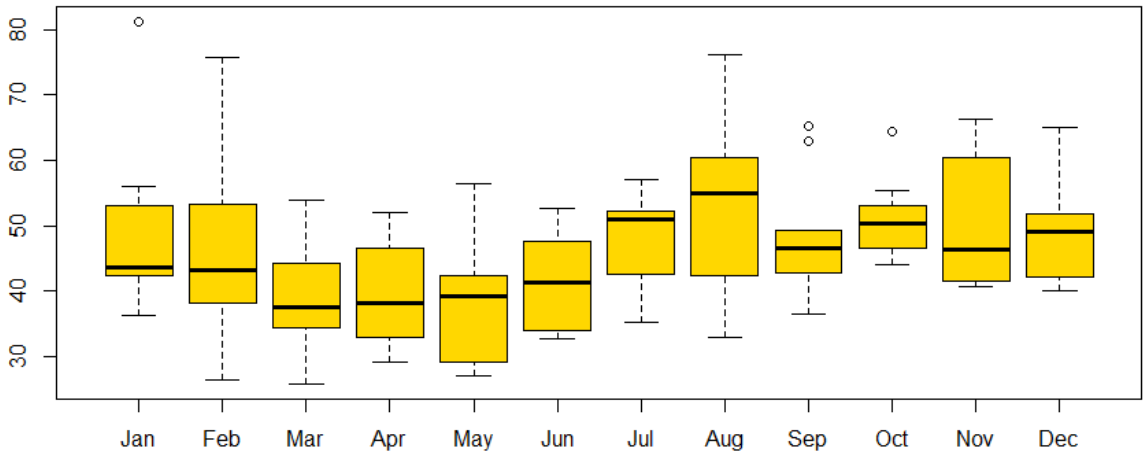
Sometimes, a non-stationary time series can be transformed to stationarity, by taking the first differences. Differencing a time series means, to subtract each data point in the series from its successor. It is commonly used to make a time series stationary. For most time series patterns, 1 or 2 differencing is

necessary to make it a stationary series. But, how to know how many differencing is needed? The ndiffs from forecast package can help find out how many regular differencing is needed to make the series stationary. Another trick before differencing is logging the data. Functions such as the log difference are helpful for making non-stationary data stationary, since by logging we make the numbers lower and fixing the variance (removing unequal variance).

In the below plot, we observe that by taking the first differences of log data, the series seems to be stationary.



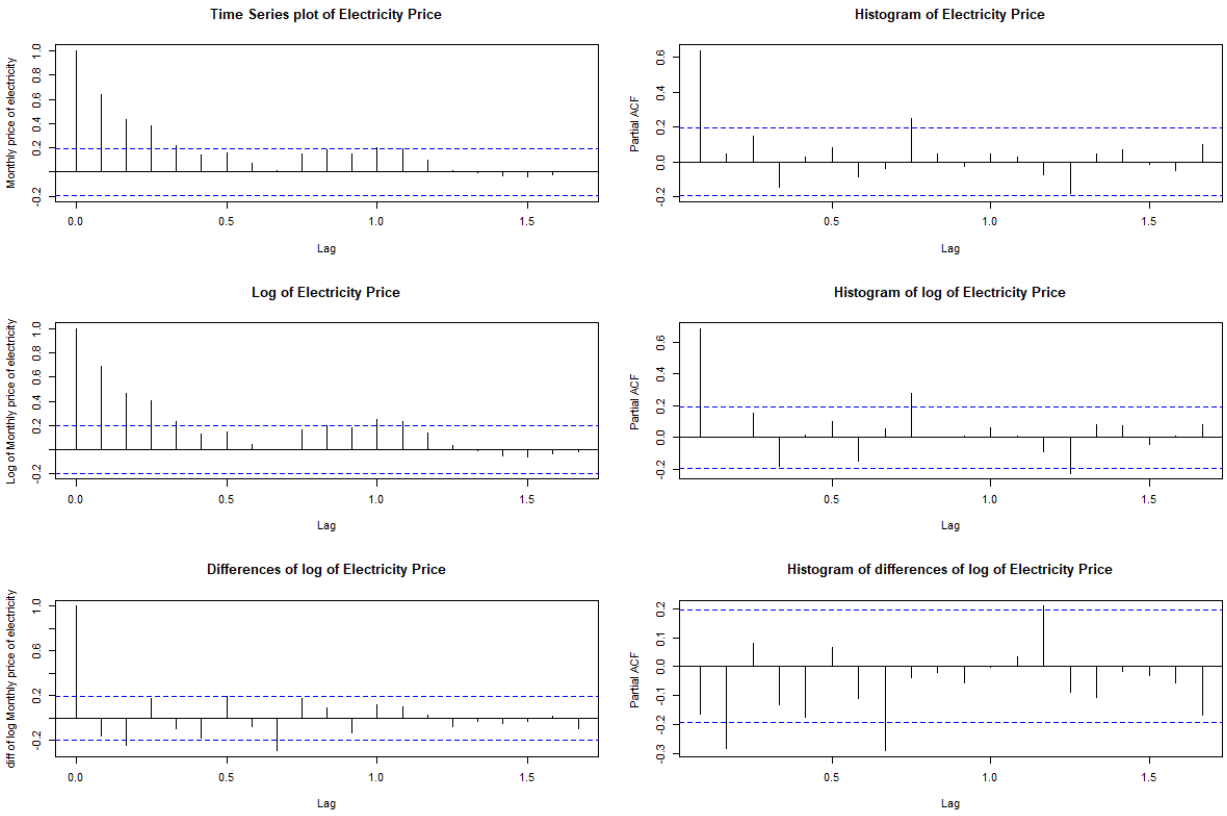
In the boxplot below, we can see in spring (March, April, May) the price of electricity is lower, and it gradually increases in summer. The rationale for this could be more people using more electricity in summer months due to the hot weather or the increase in the business activity.



Identification step - Autocorrelation and Partial-Autocorrelation

Autocorrelation is the correlation of a Time Series with lags of itself. This is a significant metric because, it shows if the previous states (lagged observations) of the time series has an influence on the current state. In the autocorrelation chart, if the autocorrelation crosses the dashed blue line, it means that specific lag is significantly correlated with current series. For example, in autocorrelation chart of initial data of electricity price - the top-left chart (below), there is significant autocorrelation for the first 4 lags shown on x-axis.

Autocorrelation is used commonly to determine if the time series is stationary or not. A stationary time series will have the autocorrelation fall to zero fairly quickly but for a non-stationary series it drops gradually.



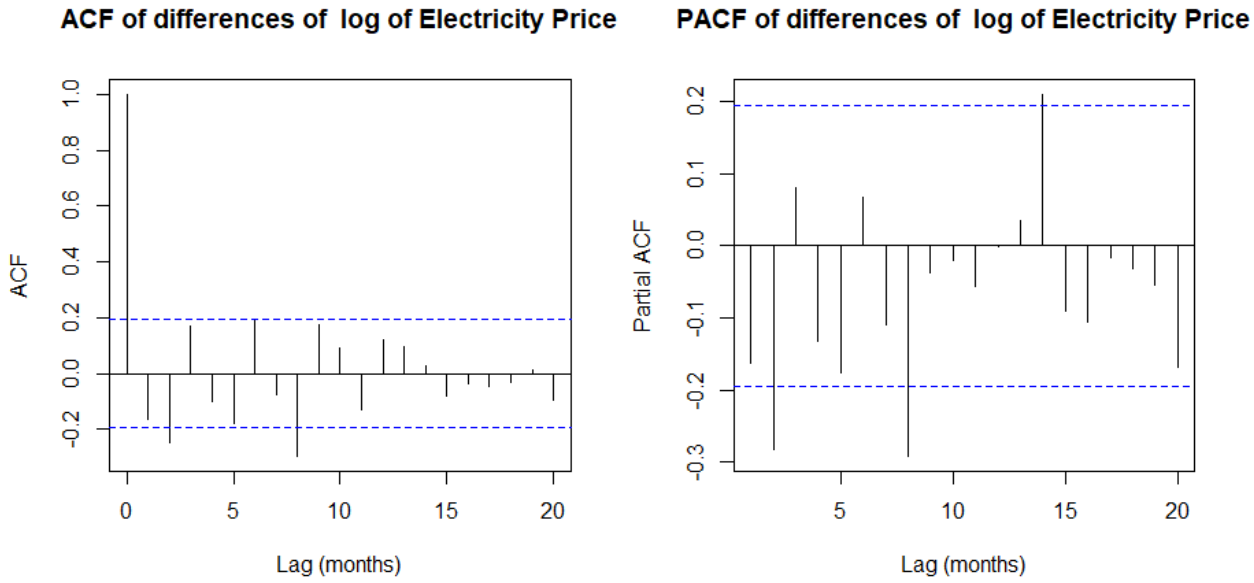
Partial Autocorrelation is partial correlation is a conditional correlation. It is the correlation between two variables under the assumption that we know and take into account the values of some other set of variables.

Following we plotted the original time series, the log and the first difference of log. The original plot shows a clear non stationarity. The ACF does not cuts off, rather it shows a slow decrease. On the other hand, the first difference data looks much stationary. The ACF cuts off after some lags. We can perform the previous test (Augmented Dickey-Fuller Test) to test the stationarity of the differenced data. But here from plot of the data and the ACF plot provide good indication that the differenced data is stationary.

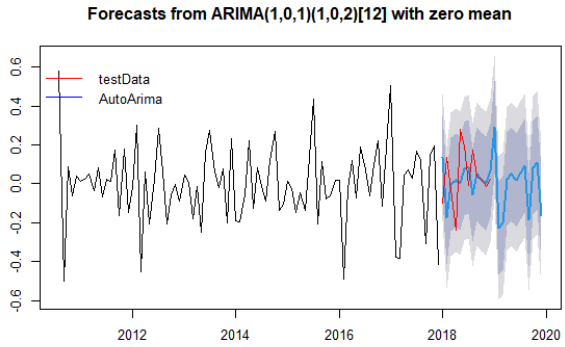
ARIMA Model

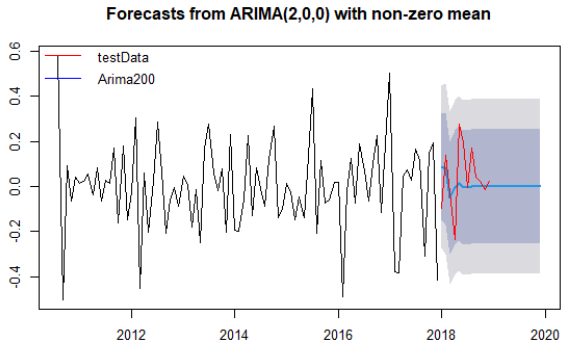
ARIMA (autoregressive integrated moving average) is a commonly used technique utilized to fit time series data and forecasting. It is a generalized version of ARMA (autoregressive moving average) process, where the ARMA process is applied for a differenced version of the data rather than original. Three numbers p, d and q specify ARIMA model and the ARIMA model is said to be of order (p,d,q). Here p, d and q are the orders of AR part, Difference and the MA part respectively. Since we took one differencing to get stationarity, here d=1.

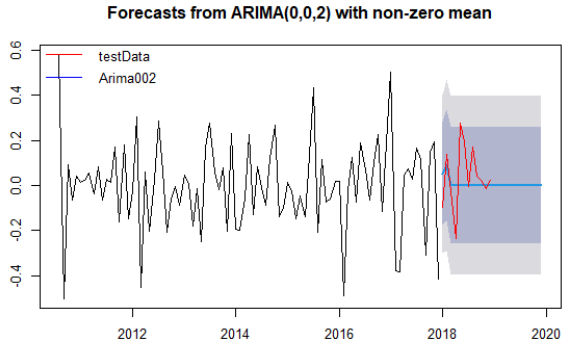
Now that the data is stationary, the next step is to get the values of p and q, the order of AR and MA part by making some guesses about p & q. To do so, we plotted the sample ACF and PACF of the first log differenced data. We see that the ACF cuts off after lag 2, having a significant spike at lag 8, and the PACF cuts off after lag 2 having also having a significant spike at lag 8. So, we propose some ARMA models for the differenced data: ARMA(0,1), ARMA(2,0) ,ARMA(2,1) ect... That is, for the original time series, we propose three ARIMA models, ARIMA(0,1,1), ARIMA(2,1,0) , ARIMA(2,1,1)…

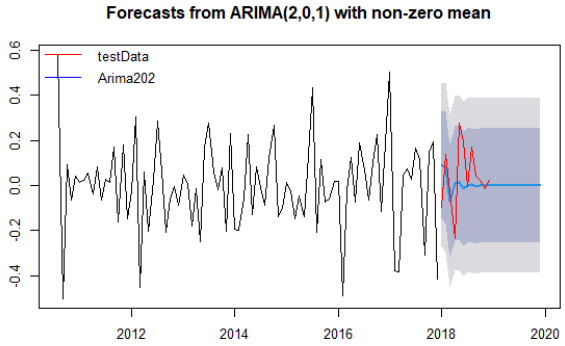


Let us make three ARIMA model with orders as proposed earlier. We retain last year (2018) observations for forecasting and use first years (2010-2017) observations to fit the models. In the preceding part we manually examined different aspects of the time series in order to find the correct specification in R. But we can do it in R using the auto.arima() function of the forecast package.









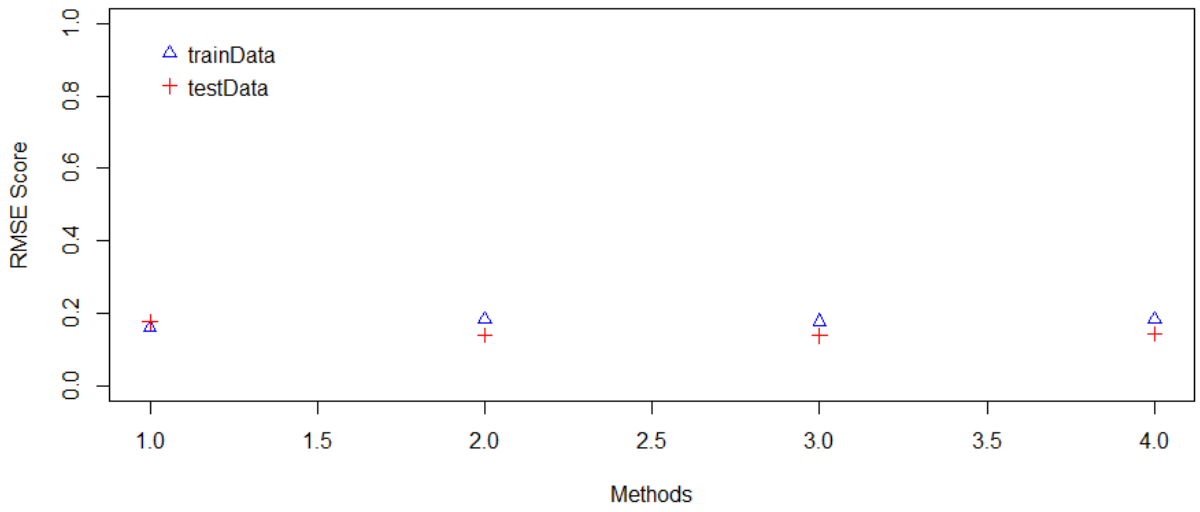
Estimation step - Best Model Selection

We can estimate the model's parameters using Maximum Likelihood method and least squares method in order to choose the best model for prediction.

The best fit model can be selected based on Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values. The idea is to choose a model with minimum AIC and BIC values. However, if we choose the best model for prediction based on AIC criterion, we would end with a model that best fits the data, since AIC criterion is based on maximum likelihood method. By using the least squared method, we minimize the sum of squared residuals of the model under consideration.

After trying and experimenting with different models, taking different AR and MA parts each time, we decided to keep the model with the lower RMSE in the training set. The model that scored the best was ARIMA with 2 parts of AR.

From the four below model we plotted, the ARIMA_200 gave us the best prediction regarding RMSE.



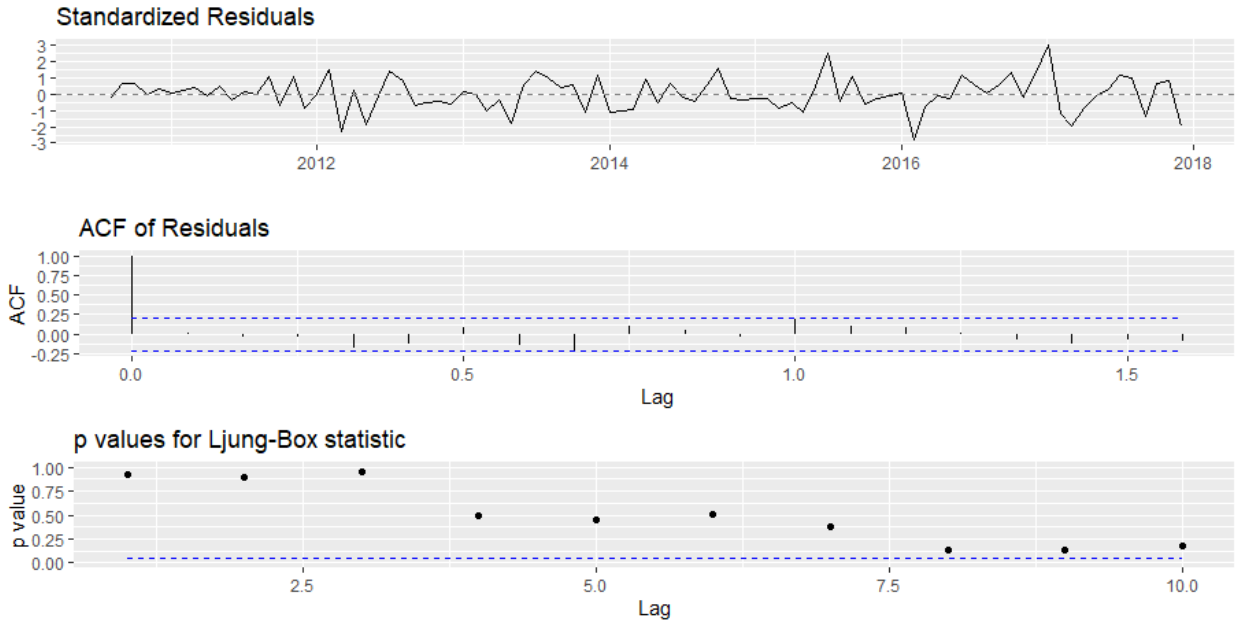
Diagnostic checking step

In this step, we examine if the chosen (estimated) model fits the data reasonably well. We have to test if the residuals of the estimated model are uncorrelated, homoscedastic and normal, i.e. white noise.

Test of Autocorrelation of Residuals and Zero mean

The residuals should be uncorrelated. If there are correlations between residuals, then there is information left in the residuals which should be used in computing forecasts. Different tests can be used to test the autocorrelation assumption of residuals. We will plot the ACF and PACF of residuals and conduct a Box-Pierce and Ljung-Box test.

Moreover, the residuals should have zero mean. If the residuals have a mean other than zero, then the forecasts are biased.



We can see that lag 0.7 is just about touching the significance interval. To test whether there is significant evidence for non-zero correlations we can carry out a Ljung-Box test. In R we can use the Box-test() function.

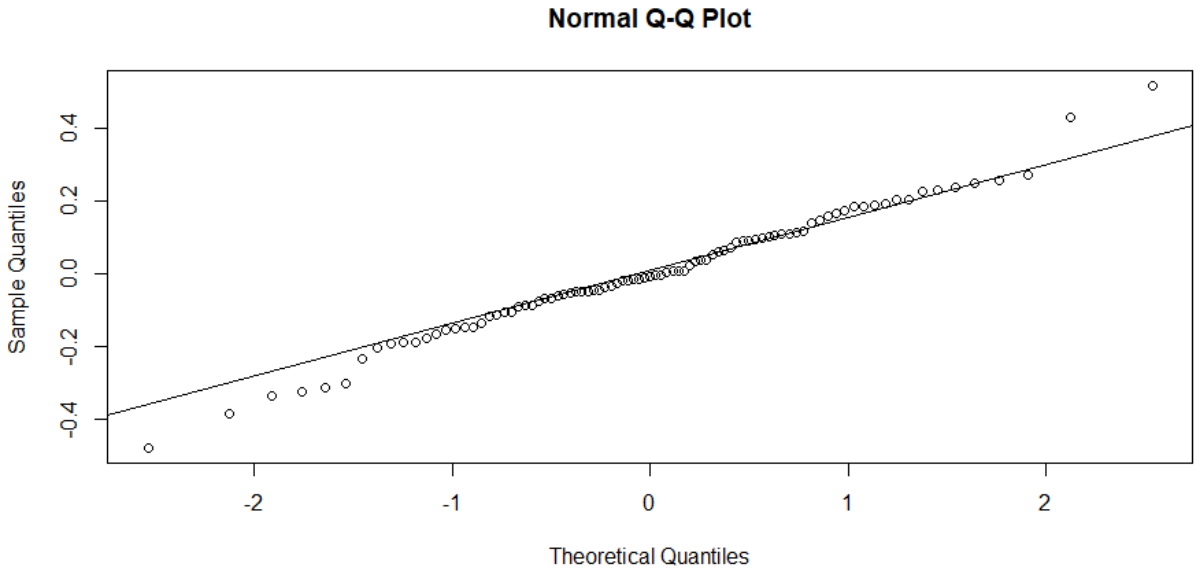
```
> Box.test(mod_200$residuals, lag = 20, type='Ljung-Box')

        Box-Ljung test

data:  mod_200$residuals
X-squared = 27.148, df = 20, p-value = 0.1312

>
```

Here we see that the p-value is 0.1312. If the p value is greater than 0.05 then the residuals are independent which we want for the model to be correct. Ljung-Box test shows no lag with p-value less than 0.05, which suggests the absence of autocorrelation.

Normality Test of residuals

It is useful to check if the residuals are normally distributed. Qqnorm is a generic function the default method of which produces a normal QQ plot of the values in y. Qqline adds a line to a "theoretical", by default normal, quantile-quantile plot which passes through the probs quantiles, by default the first and third quartiles. From the below plot we can assume that the residuals are normally distributed.

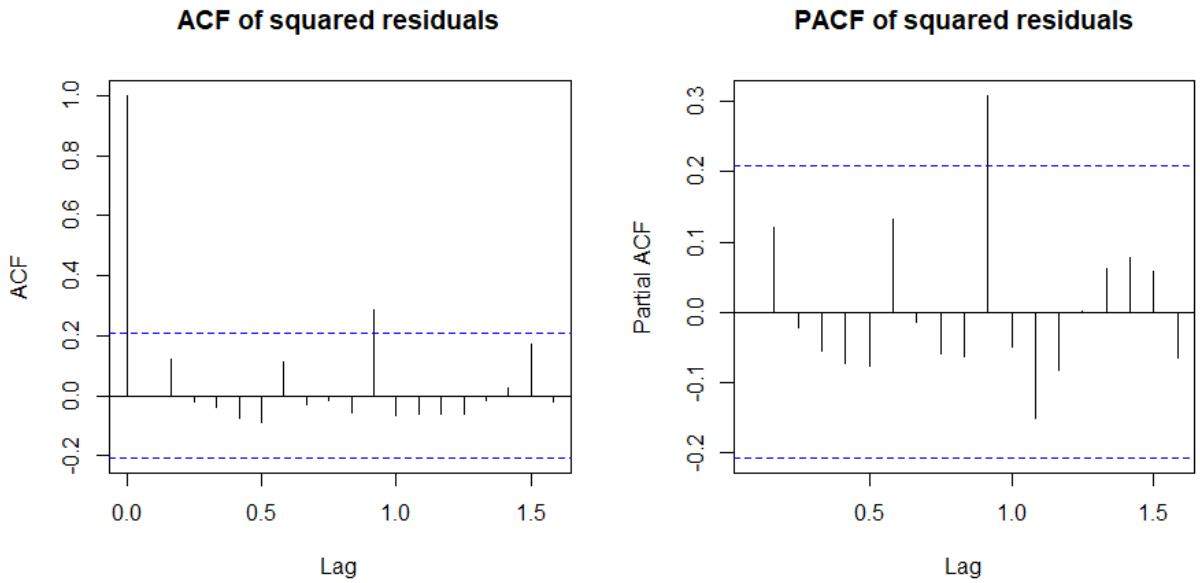
Moreover, the p-value (0.6035) in excess of 0.05 for the Shapiro-Wilk normality test supports the null hypothesis of normal distribution of residuals. We come to the same conclusion by noticing that the qq-plot displays a straight line. No pattern is apparent on the plot of residuals against the predicted values, or the residuals over time. Therefore, we consider the model adequate.



Test of Heteroskedasticity of Residuals

In addition to above essential properties, it is useful for the residuals to also have constant variance. Different tests (directly or indirectly) can be used to test the heteroscedasticity assumption of residuals:

- Autocorrelation and partial autocorrelation plots of squared residuals



The PACF and ACF of squared residuals have one significant lag at 0.9. Let's test if this is significant.
We can conduct an autocorrelation test of squared residuals. P value is less than the significant level of 0.05 and therefore we can conclude that the residuals of our ARIMA prediction model is stationary.

```
> adf.test(residuals^2, alternative ="stationary")

        Augmented Dickey-Fuller Test

data:  residuals^2
Dickey-Fuller = -4.8474, Lag order = 4, p-value = 0.01
alternative hypothesis: stationary
```

All the above graphs and tests show that the method produces forecasts that appear to account for all available information. The mean of the residuals is close to zero and there is no significant correlation in the residuals series. The time plot of the residuals shows that the variation of the residuals stays much the same across the historical data, and therefore the residual variance can be treated as constant. The residual plots appear to be centered around 0 as noise, with no pattern. Therefore, the Arima200 model is a fairly good fit.

Forecasting step

Having chosen the model ARIMA200, we will now predict the monthly price of electricity in the first half of 2019. Serveral methods of forecasting were implemented to measure which model would be the best for forecasting a horizon of 6 months. The accuracy measure was RMSE. When creating a predictive model, the goal is to create a model where it captures the signal rather than the noise of the data. It is to be noted that the RMSE score will return a result for both the "training set" and the "test set". RMSE score on the training set is a metric which measures how much signal and noise is explained by the model.

Generally speaking, the score will be lower for the "test set". The training loss is calculated over the entire training dataset. Likewise, the validation loss is calculated over the entire validation dataset. The training

set is typically at least 4 times as large as the validation (80%-20%). Given that the error is calculated over all samples, we could expect up to approximately 4X the loss measure of the validation set.

To sum up, the prediction for the electricity prices for the first 6 months of 2019 would be:

- **JANUARY: 65.40938,**
- **FEB: 65.33704,**
- **MARCH:65.65618,**
- **APRIL: 66.01185,**
- **MAY: 66.25212,**
- **JUNE: 66.50791**