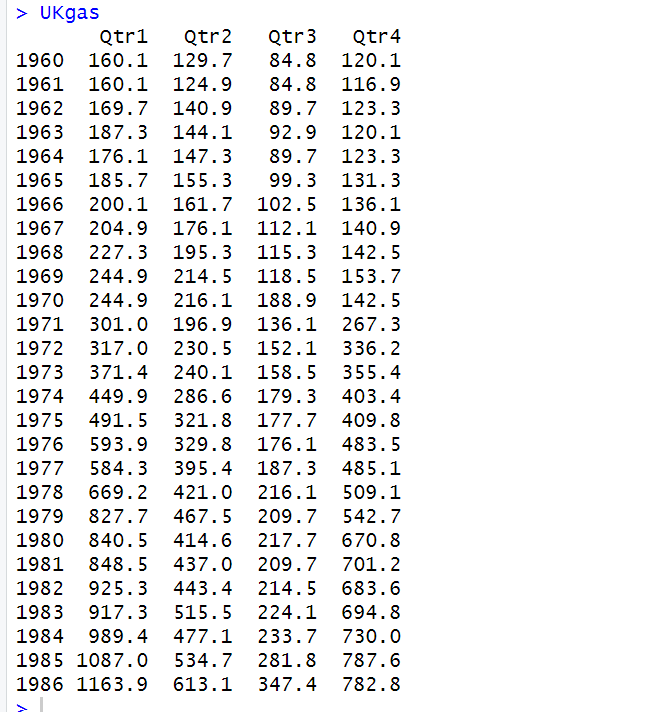
**Course Summative Assessment**

Conduct a thorough analysis of the dataset UKgas, available in the datasets package. Your analysis should be typed and should include a number of figures (at least 8, but more may be warranted) that you generate in R yourself. You should discuss an ARIMA or SARIMA fit, an ARIMA or SARIMA forecast, an exponential smoothing fit, and an exponential smoothing forecast. The (S)ARIMA fit will be the most important part of your analysis, as you should be most equipped to discuss this in detail based on what was covered in this course

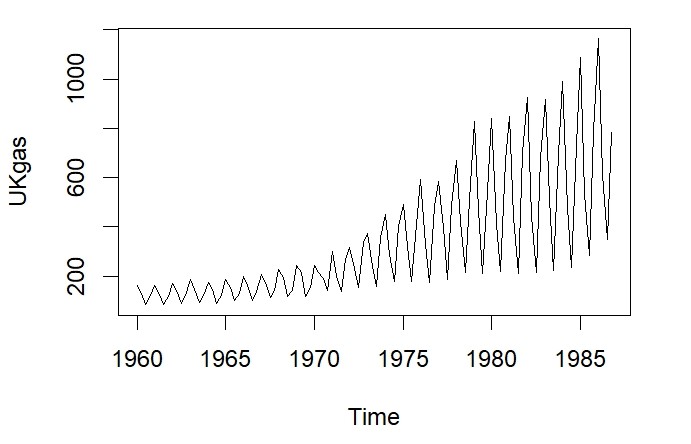
**Solution:**

1. **Exploring the dataset – UKgas**

UKgas dataset contains details of the consumption of gas in UK from 1960 to 1986 which is collected on quarterly basis. Below is the snapshot of the data in this time series.



When we plot this dataset, the graph looks like below:



1. **Making the timeseries stationary:**

As we can see from the UKgas plot, we can observe trend, seasonality and variability of data. We will have to difference the timeseries in order to make it stationary. First, we will have to take log differencing to remove variability, then we will use lag-4 differencing in order to remove seasonality. We will then difference the data once again in order to remove trend.

Below is the R code:

X.diff<-diff(diff(log(UKgas),lag=4))

The mean of this differenced time series is -5.935059e-05 , which is equal to zero. Hence we will be considering the model as mean zero model.

Lets plot the differenced timeseries

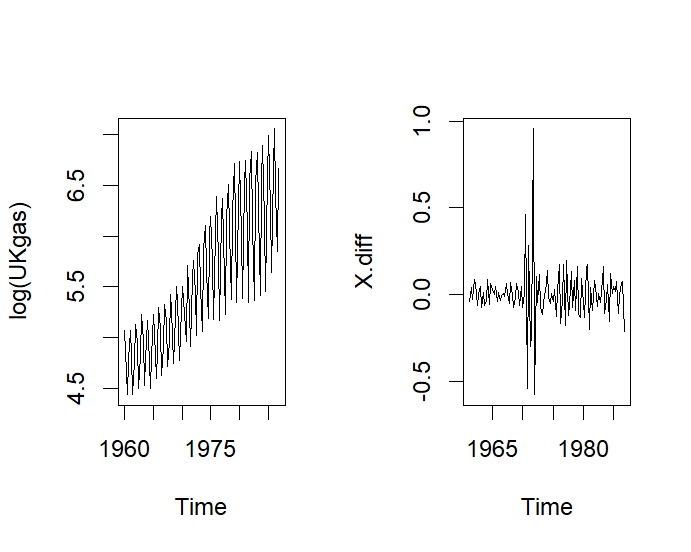
R code:

par(mfrow=c(1,2))

plot(log(UKgas))

plot(X.diff)

Output:



The differenced timeseries looks stationary with mean 0.

1. **Augmented Dickey Fuller Test Analysis:**

We run the adfTest on the differenced timeseries X.diff to check the p-values. Before running the test, we need to calculate the length of lag with the help of R

R code;

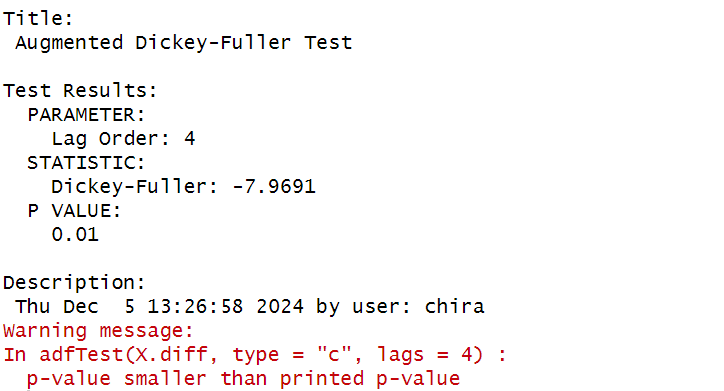
(length(X.diff)-1)^(1/3)

The value of length comes up as 4.641589. Lets consider lag=4 and lag=5 and run the adfTest. We will not be including constant here as we have considered mean is zero.

R code:

adfTest(X.diff,type='c',lags=4)

Output:



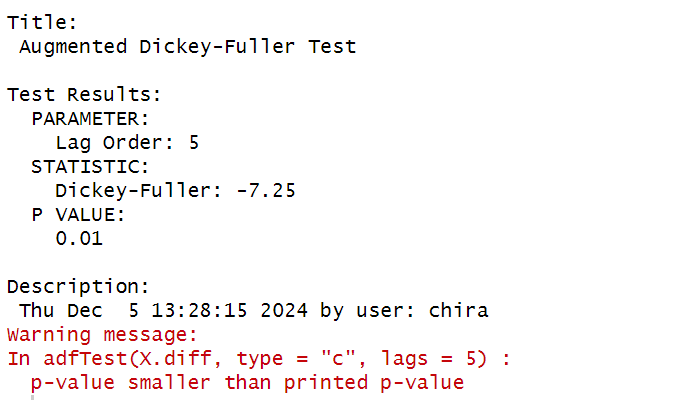
The p-value is less than 0.01, which means no additional differencing is required.

Lets check for lag=5.

R code:

adfTest(X.diff,type='c',lags=5)

Output:



For Lag=5, the p-value is less than 0.01, which means no additional differencing is necessary.

1. **ACF and PACF plots:**

We then plot the sample ACF and PACF plots of the differenced timeseries, X.diff for initial analysis.

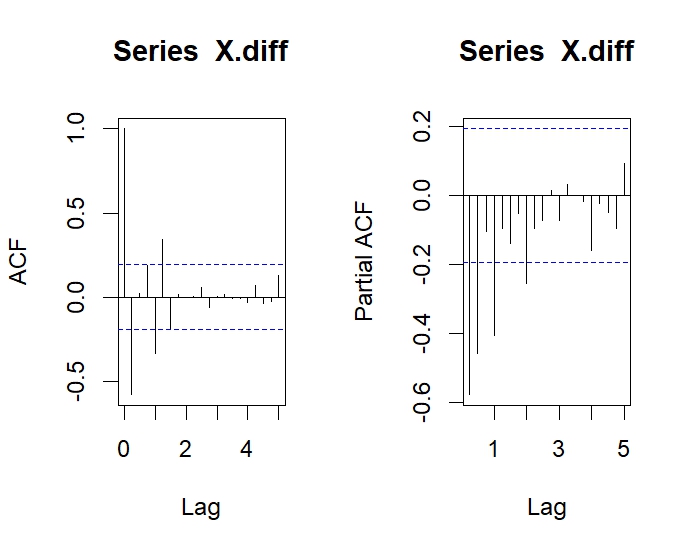
R code:

par(mfrow=c(1,2))

acf(X.diff)

pacf(X.diff)

Output:



ACF Plot:

In the ACF plot, we can see significant spikes at lags 1,4 and 5. Apart from that all other lags are within the dotted lines and decaying. This indicates the presence of seasonal MA component Q and MA component q. Let us guess the value of Q=1 and q =1 with differencing 0.

PACF Plot:

In the PACF plot, there are significant lags at 1,2,4,8. Apart from that, all other lags are within the dotted lines and showing a gradual decline. We can guess the value of AR p=2 and seasonal AR P=1 based on the plot. There is seasonality in the order of 4, so we will consider s=4 with differencing =0.

Our model guess= ARIMA(2,0,1)(1,0,1)[4]

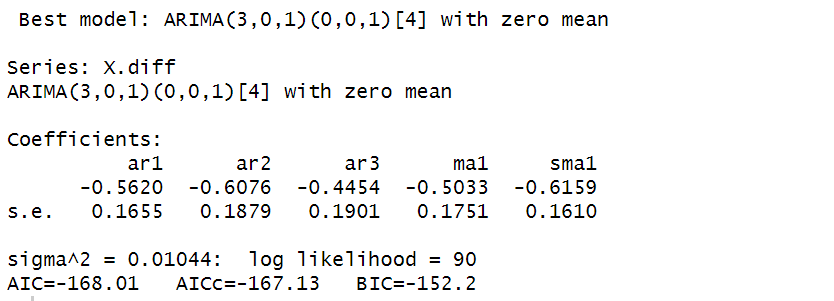
1. **Auto.arima model calculation:**

We will run auto.arima function to check the model selected by this function based on the value of AICc. We are not considering drift, hence allowdrift=FALSE

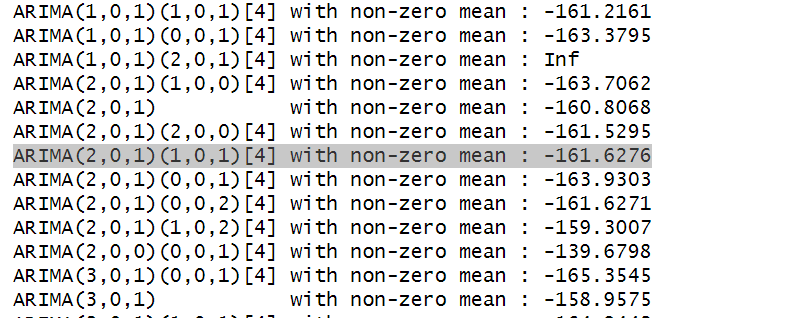
R code:

auto.arima(X.diff,ic = "aicc",trace = TRUE, allowdrift =FALSE, approximation = FALSE)

Output:



The best model selected is ARIMA(3,0,1)(0,0,1)[4] with mean zero. If we consider the model guessed on the basis of PACF and ACF ARIMA(2,0,1)(1,0,1)[4], the AICc= -161.627, which is little higher than the AICc of the selected model.

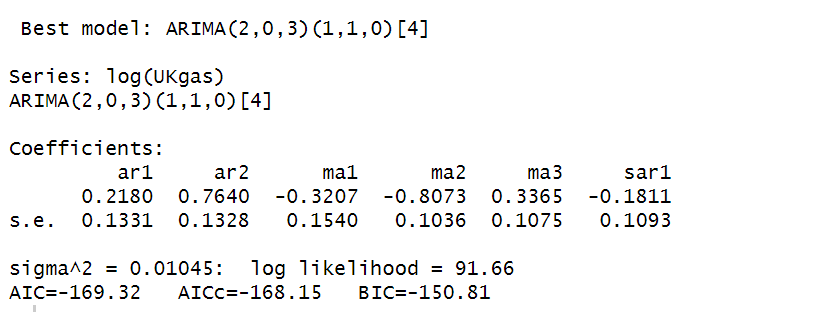


If we run auto.arima() on log differenced UKgas dataset, it predicts model as ARIMA(2,0,3)(1,1,0)[4]. This model and the model prediction for differenced dataset are quite different.

R code:

auto.arima(log(UKgas),trace = TRUE, allowdrift =FALSE, approximation = FALSE)

Output:



We will consider the model selected by auto.arima() and define the model equation and model summary.

1. **Model Equation.**

Considering the model selected by auto.arima() ARIMA(3,0,1)(0,0,1)[4] with mean zero, below are the expected values of the coefficients of the model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| AR1 | AR2 | AR3 | MA1 | SMA1 |
| -0.5620 | -0.6076 | -0.4454 | -0.5033 | -0.6159 |

Model Equation is as below:

equation

Where equation where Xt=log-transformed UKgas.

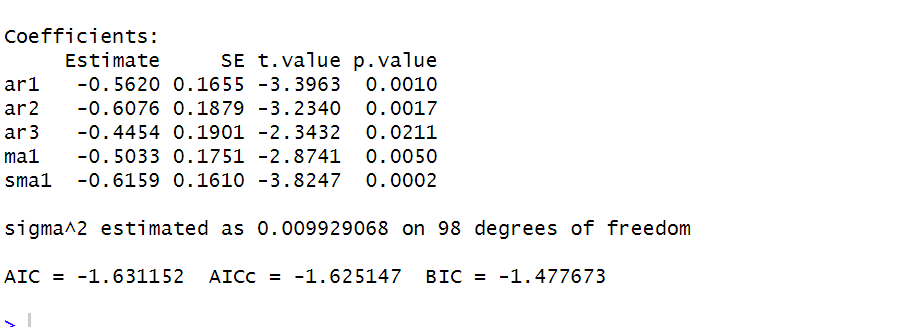
1. **Model Summary using sarima():**

Below is the model diagnostics by sarima() in R.

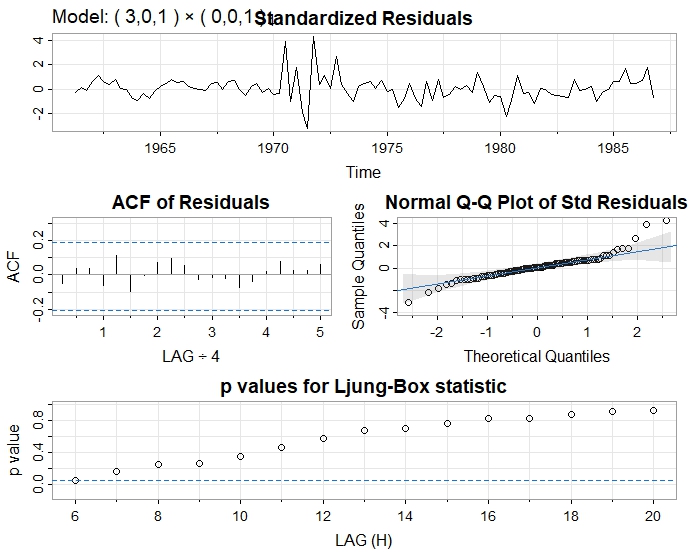
R code:

sarima(X.diff,3,0,1,0,0,1,4,no.constant = TRUE)

Output:



The estimated value of coefficients are same as predicted by auto.arima().Even the standard deviation value estimated by both the functions are same and the value is very less.



Standardized Residuals – The residual graph appears to be stationary and revolving around 0 mean. Nothing unusual.

ACF of Residuals – All the lag spikes are within the dotted lines, which makes it consistent with white noise model.

Normal Q-Q Plot – Most of the residuals lie along the diagonal line, which means the residuals normally distributed. There are some deviations along both the edges. Some residuals lie quite above the line. Overall, the plot looks good.

Ljung-Box Statistics – Except for the p-value at lag 6, all other p-values lie above the threshold line. P-value of lag 6 lie on the threshold line. After lag 6, all other p-values shows a gradual increase. Overall the p-values are well above the threshold line which makes us reject the null hypothesis of no correlation among the residuals.

1. **Model Forecasting using forecast function**

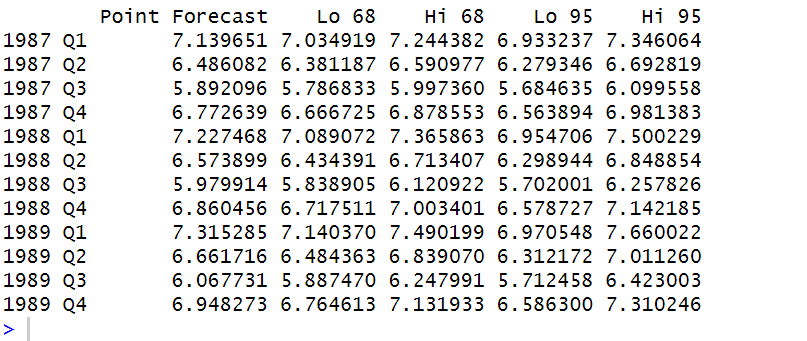
We run the forecast function on the log(UKgas) and forecast values for next 3 years ie 1987,1988 and 1989. While predicting values, we are considering 2 confidence interval of 68% and 95%.

R code:

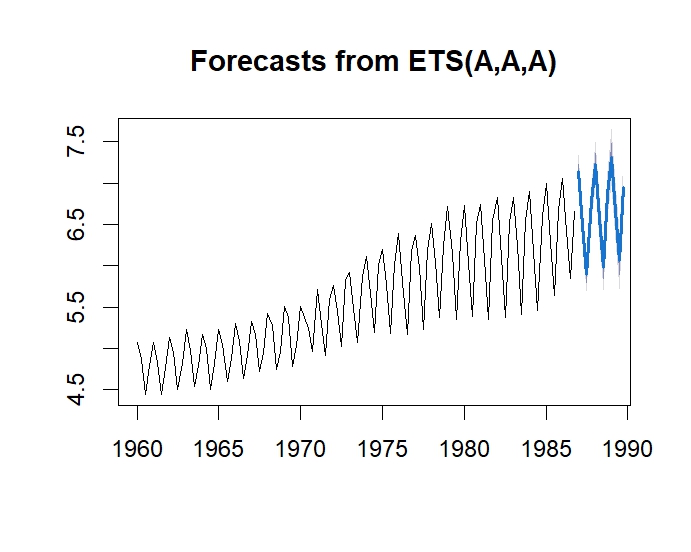
X.forecast2 <- forecast(log(UKgas), h = 12, level = c(.68, .95))

Output:

Below are the forecasted quarterly values for year 1987,1988 and 1989.



Plot:



As per the plot, the forecasted values are in sync with previous or past values. The forecasted values are showing an upward trend and seasonality.

1. **Exponential smoothing Fit and Forecast**

We use HoltWinters() function for fitting one-step ahead forecasting on the log(UKgas) data.

R code:

X.HW <- HoltWinters(log(UKgas))

Output:



The above output shows the value predicted by Holt-Winter model. These values are then used in 3 smoothing equations - the level(l1), trend (b1) and seasonal (s1).

The value of a and b are the final values of level and trend. The values of s1 to s4 are the final values of seasonal parameter.

Example 1: Suppose we want to forecast value for Q2 of year 1988.

We will consider below equation for point forecast and substitute the values in the below equation



X=a+3\*b+s2

X=6.35670491+(6\* 0.02234170) + 0.08604617 = 6.57680128

This is the same value predicted by the forecast function.



Example 2: Suppose we want to forecast value for Q3 of year 1987.

We will consider below equation for point forecast and substitute the values in the below equation



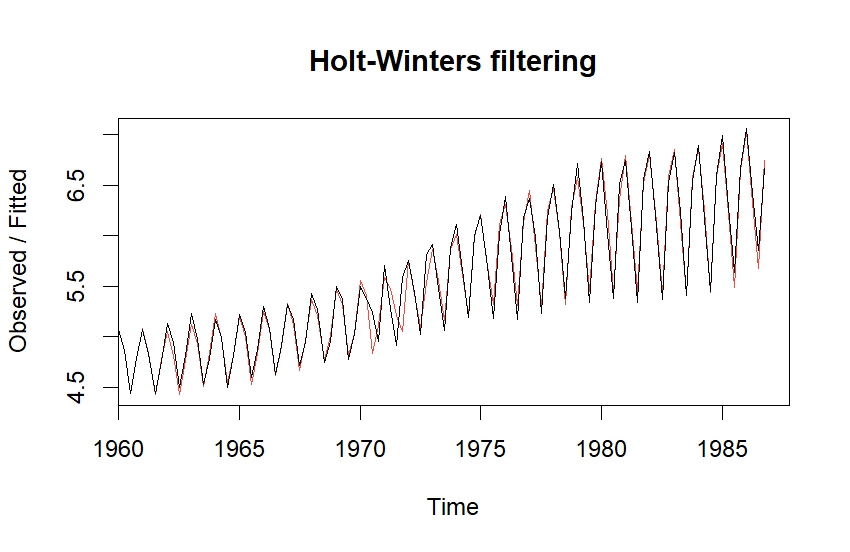
X=a+6\*b+s2

X=6.35670491+(3\* 0.02234170) - 0.53107110 = 5.89265891

This is the same value predicted by the forecast function.



Plot:



The above plot represents log(UKgas) dataset with Holt-Winter one step ahead forecasting of seasonal exponential smoothing fitted model.

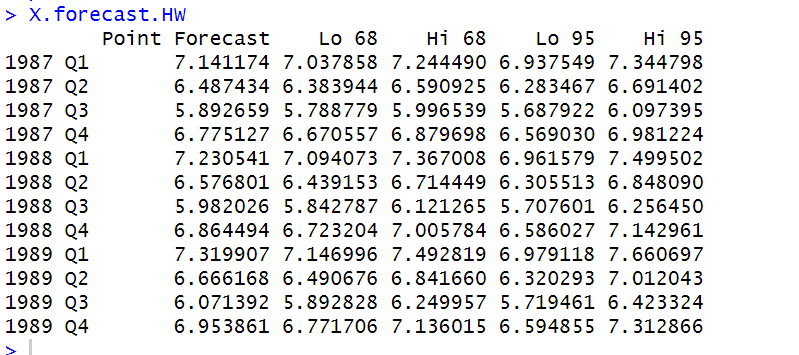
**Holt-Winter Forecasting:**

R code:

X.forecast.HW <- forecast(X.HW,h = 12,level = c(.68, .95))

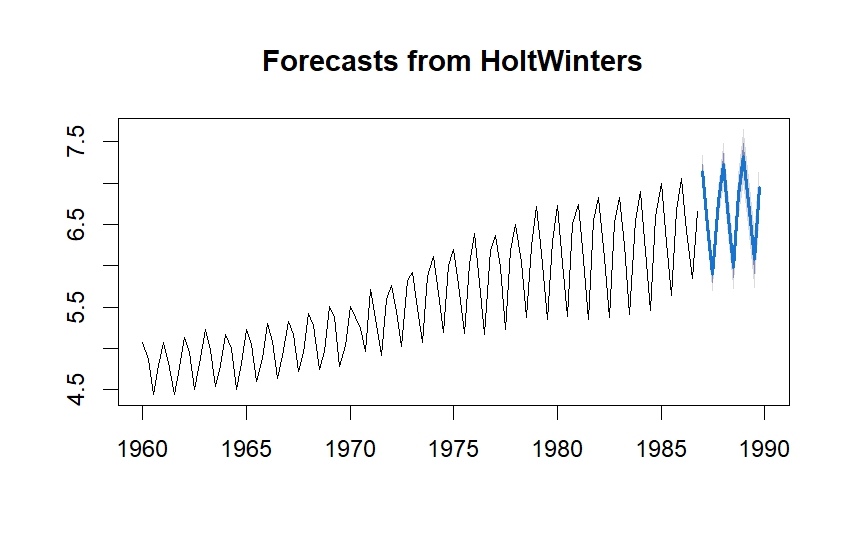
plot(X.forecast.HW, xlim = c(1960, 1990))

Output:



The above values are quarterly point forecasted values for next 3 years (1987-1989) using the Holt-Winter fit.

Plot:



The above plot shows forecasted values using Holt-Winter exponential smoothing fit.

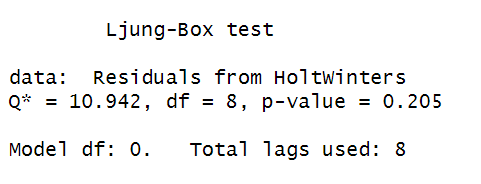
As we can see from the plot, the forecasted values are showing an upward trend and seasonality which makes the values acceptable.

**Model Diagnostics on Holt-Winter forecasted values**

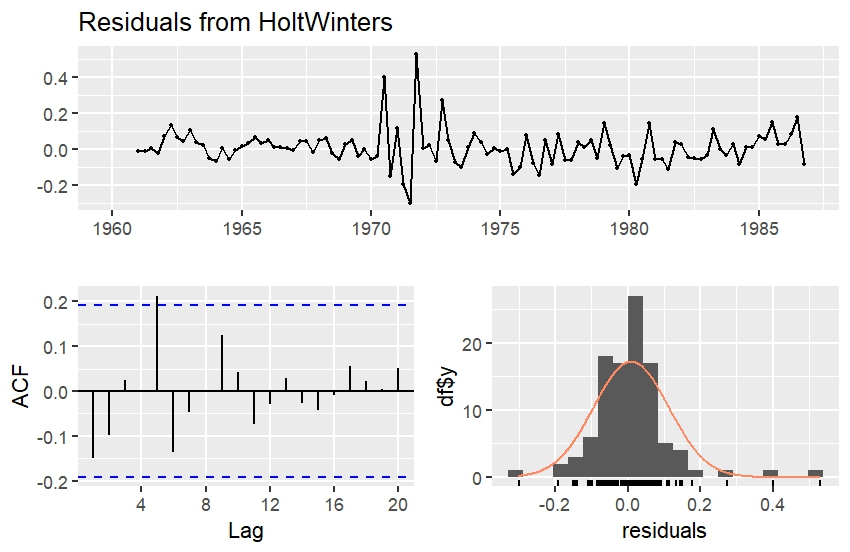
R code:

checkresiduals(X.forecast.HW)

Output:



Plot



The Residual graph seems stationary and revolve around the mean 0. In the ACF plot, the lag values are within the threshold lines except for the one at lag 5. This seems to be an acceptable graph. In the histogram plot, the residuals are normally distributed, except for the one at the end on both the sides.