Assignment 2

For STA-501
(Time Series Analysis and Forecasting)

Submitted by:

SHILPA SONI

2017IMSST008

Integrated M.Sc. Statistics
Sem-9
2017imsst008@curaj.ac.in

Submitted to:

Dr. Jitendra Kumar

Central University of Rajasthan,
Bandar Sindri, Dist.: Ajmer-305817, Rajasthan

Contents:

Problem Statement:

Choose a real dataset and fit an appropriate model to the dataset. Make a prediction based on the fit.

Dataset Chosen and Objective:

I have obtained the Monthly Milk Production dataset for Cows in the United States (Jan 1962 to Dec 1975) from https://www.stat.auckland.ac.nz/~ihaka/726/milk.txt and my objective is to fit an appropriate model to this dataset. Upon doing so, I hope to run a diagnostic check and see if the model is a suitable candidate for making forecasts. If so, I will forecast the milk production values for the year 1976 and plot the same.

Theory:

1. Introduction to Time Series Forecasting

A Time Series is defined as a series of data points recorded at different time intervals. The time order can be daily, monthly, or even yearly.

Time Series forecasting is the process of using a statistical model to predict future values of a time series based on past results.

Forecasting is the step where we want to predict the future values the series is going to take. Forecasting a time series is often of tremendous commercial value.

Forecasting a time series can be broadly divided into two types:

i. If we use only the previous values of the time series to predict its future values, it is called Univariate Time Series Forecasting.
ii. If we use predictors other than the series (like exogenous variables) to forecast it is called Multivariate Time Series Forecasting.

2. Introduction to ARIMA Models

ARIMA stands for Autoregressive Integrated Moving Average Model. It belongs to a class of models that explains a given time series based on its own past values i.e. its own lags and the lagged forecast errors. The equation can be used to forecast future values. Any 'non-seasonal' time series that exhibits patterns and is not a random white noise can be modeled with ARIMA models.

So, AutoRegressive Integrated Moving Average is a forecasting algorithm based on the idea that the information in the past values of the time series can alone be used to predict the future values.

ARIMA Models are specified by three order parameters: (p, d, q),

where,

- p is the order of the AR term
- q is the order of the MA term
- d is the number of differencing required to make the time series stationary

AR(p) Autoregression – a regression model that utilizes the dependent relationship between a current observation and observations over a previous period. An auto regressive (AR(p)) component refers to the use of past values in the regression equation for the time series.

I(d) Integration – uses differencing of observations (subtracting an observation from observation at the previous time step) in order to make the time series stationary. Differencing involves the subtraction of the current values of a series with its previous values d number of times.

MA(q) Moving Average – a model that uses the dependency between an observation and a residual error from a moving average model applied to lagged observations. A moving average component depicts the error of the model as a combination of previous error terms. The order q represents the number of terms to be included in the model.

3. The meaning of p, d and q in ARIMA model

3.1 The meaning of p

- p is the order of the Auto Regressive (AR) term. It refers to the number of lags of Y to be used as predictors.

3.2 The meaning of d

- The term Auto Regressive' in ARIMA means it is a linear regression model that uses its own lags as predictors. Linear regression models, as we know, work best when the predictors are not correlated and are independent of each other. So we need to make the time series stationary.

- The most common approach to make the series stationary is to difference it. That is, subtract the previous value from the current value. Sometimes, depending on the complexity of the series, more than one differencing may be needed.

- The value of d, therefore, is the minimum number of differencing needed to make the series stationary. If the time series is already stationary, then d = 0.

3.3 The meaning of q

- q is the order of the Moving Average (MA) term. It refers to the number of lagged forecast errors that should go into the ARIMA Model.

4. ARIMA model

An ARIMA model is one where the time series was differenced at least once to make it stationary, and we combine the AR and the MA terms. So the equation of an ARIMA model becomes:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + .. + \beta_p Y_{t-p} \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + .. + \phi_q \epsilon_{t-q}$$

ARIMA model in words:

Predicted Yt = Constant + Linear combination Lags of Y (upto p lags) + Linear Combination of Lagged forecast errors (upto q lags)

5. How to find the order of differencing (d) in ARIMA model

- As stated earlier, the purpose of differencing is to make the time series stationary. But we should be careful to not over-difference the series. An over differenced series may still be stationary, which in turn will affect the model parameters.

- So we should determine the right order of differencing. The right order of differencing is the minimum differencing required to get a near-stationary series which roams around a defined mean and the ACF plot reaches to zero fairly quick.

- If the autocorrelations are positive for many number of lags (10 or more), then the series needs further differencing. On the other hand, if the lag 1 autocorrelation itself is too negative, then the series is probably over-differenced.

- If we can't really decide between two orders of differencing, then we go with the order that gives the least standard deviation in the differenced series.

- Now, we will explain these concepts with the help of an example as follows:

- First, I will check if the series is stationary using the Augmented Dickey Fuller test (ADF Test), from the statsmodels package. The reason being is that we need differencing only if the series is non-stationary. Else, no differencing is needed, that is, d=0.

- The null hypothesis (Ho) of the ADF test is that the time series is non-stationary. So, if the p-value of the test is less than the significance level (0.05) then we reject the null hypothesis and infer that the time series is indeed stationary.

- So, in our case, if P Value > 0.05 we go ahead with finding the order of differencing.

Code:

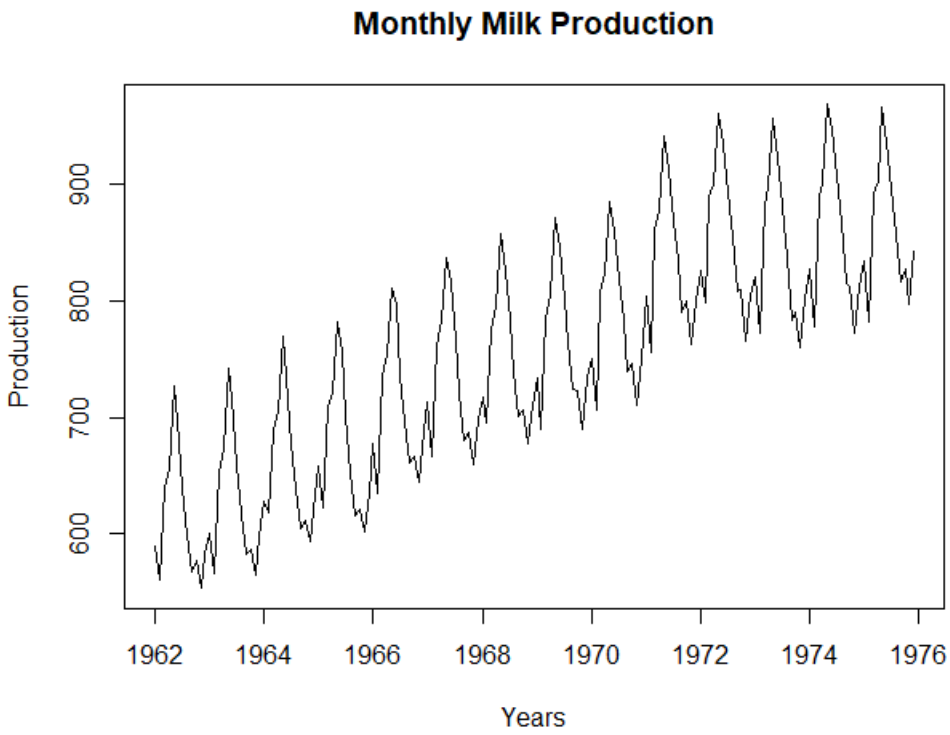
```
> #My data: Monthly Milk Production for Cows in the United States (Jan 1962 to
Dec 1975)
> data = read.csv("D:/Documents/Sem 9/STA 501 - Time Series
(JK)/cia2/milk.csv", header=TRUE)
> head(data)
```

```
    date     milk_prod
1  Jan-62       589
2  Feb-62       561
3  Mar-62       640
4  Apr-62       656
5  May-62       727
6  Jun-62       697
```

```
> summary(data)
```

```
      date            milk_prod
 Length:168         Min.   :553.0
 Class :character   1st Qu.:677.8
 Mode  :character   Median :761.0
                    Mean   :754.7
                    3rd Qu.:824.5
                    Max.   :969.0
```

```
> data = ts(data[,2], start=c(1962,1), frequency=12)
> #Initial Analysis of the type of data
> par(mfrow = c(1,1))
> plot(data, main = 'Monthly Milk Production', xlab='Years',
ylab='Production')
```

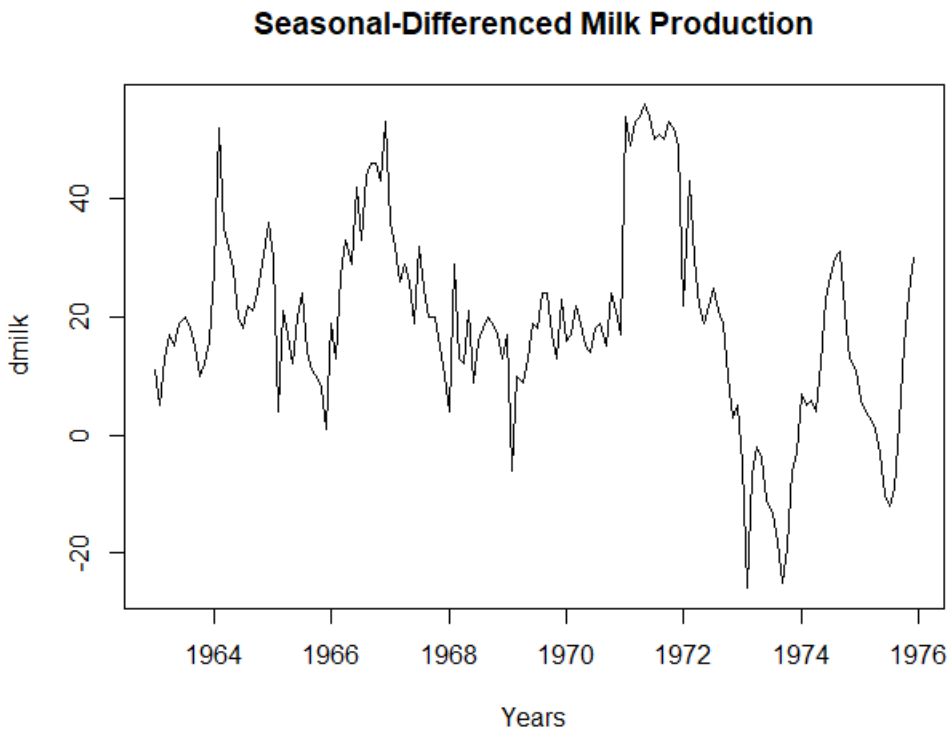


The plot shows a strong seasonal effect and an upward trend (which levels off towards the end of the series). There is no evidence that the variability of the series is increasing with its mean level, so no variance stabilizing transformation is required.

Stage 1 of Analysis:

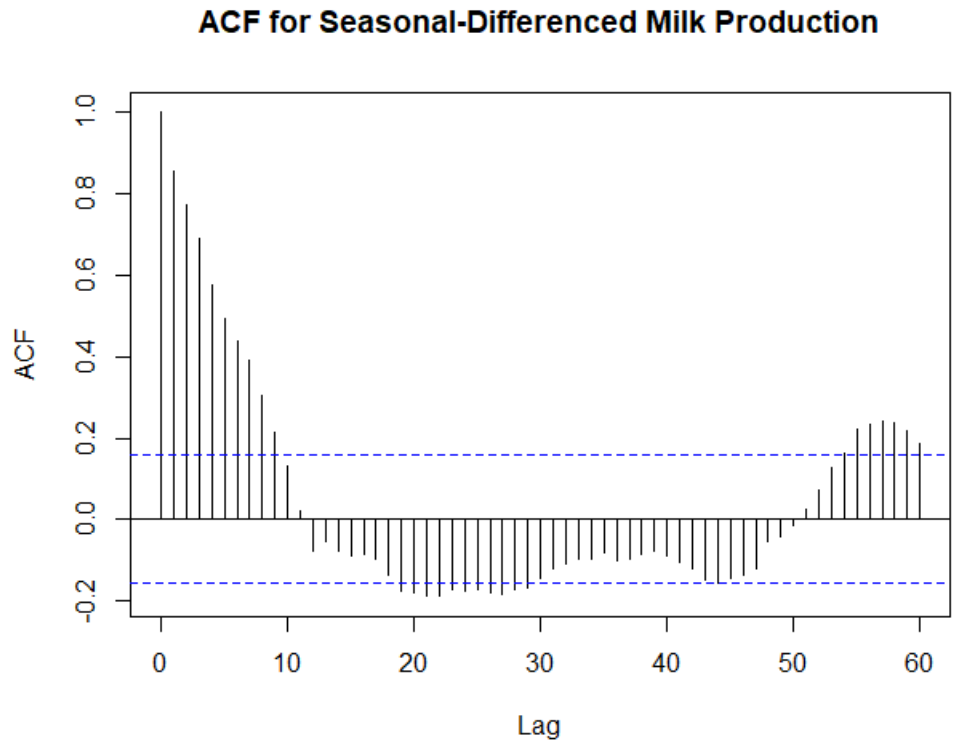
To find the amount of differencing required to make the series stationary. The most visible feature of the series is the seasonal pattern, so the appropriate first step is to apply seasonal differencing as follows:

```
> #Applying Seasonal Differencing
> par(mfrow = c(1,1))
> seas_diff = diff(data, lag=12)
> plot(seas_diff, ylab='Seasonal-Differenced Milk Production')
```



The plot shows long-term variation which could represent non-stationarity. This can be checked by looking for slow decay of the ACF(autocorrelation function) of the differenced series as follows:

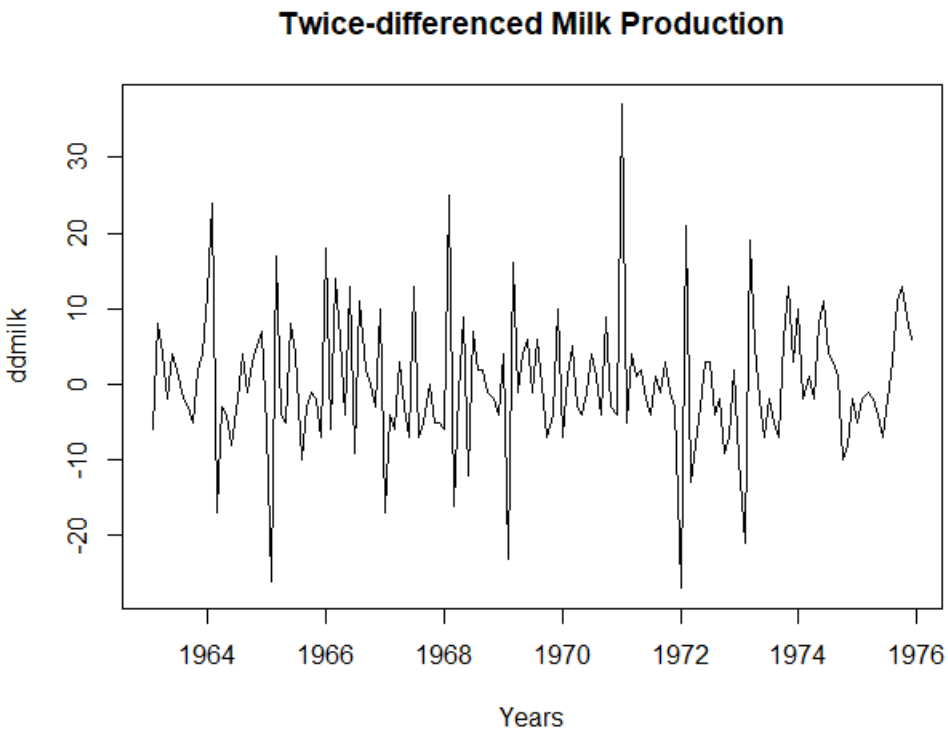
```
> #Checking the slow decay of the ACF of the Seasonal-Differenced TS
> acf(ts(seas_diff), main='ACF for Seasonal-Differenced Milk Production',
lag.max=60)
> pacf(ts(diff_2),main='PACF for Twice-differenced Milk Production',
lag.max=60)
```



The above plot shows that the ACF does decay slowly and that additional differencing is required. The ACF does not show a seasonal pattern, so it is appropriate to use simple differencing as follows:

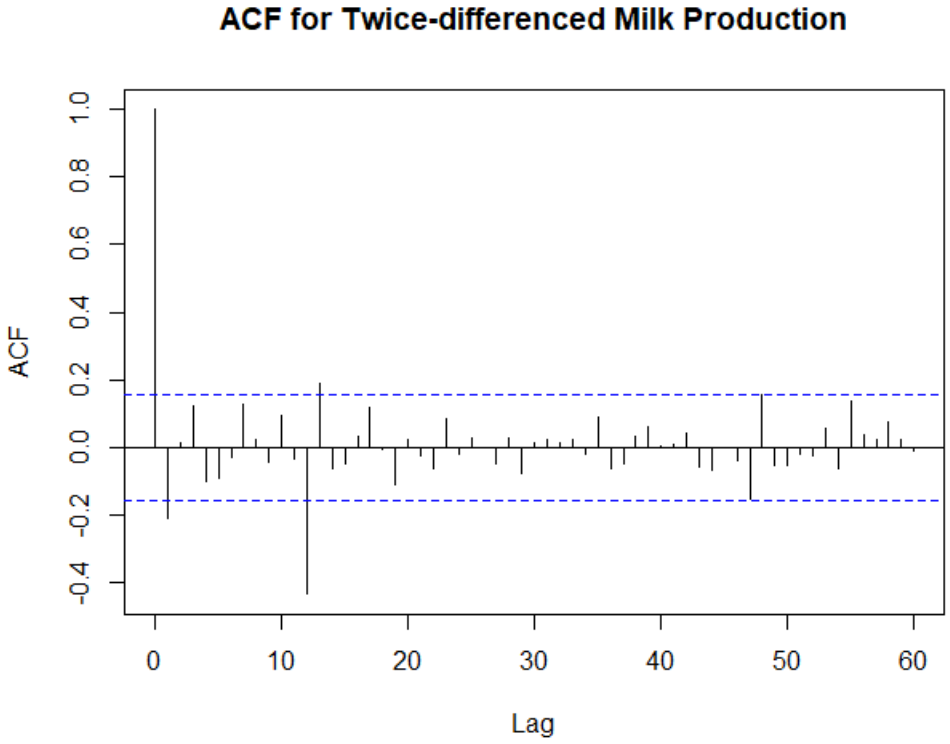
```
> #The twice-differenced milk production series
> diff_2 = diff(seas_diff)
> plot(diff_2, main='Twice-differenced Milk Production', ylab='ddmilk',
xlab='Years')
```

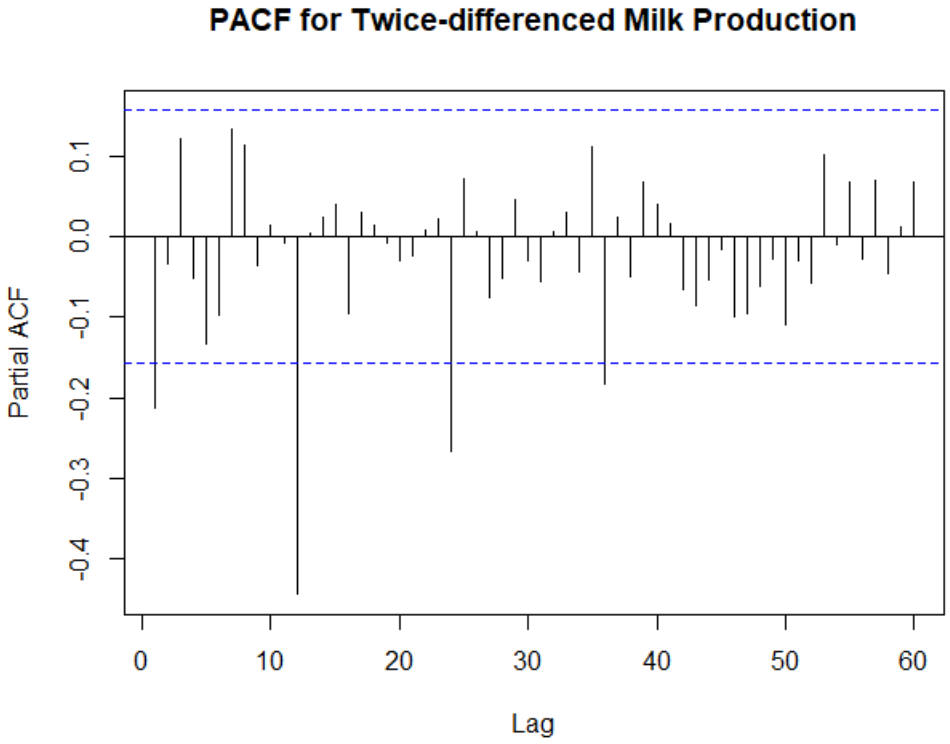
The result of applying the two kinds of differencing is shown by the ACF plot above. The series now looks stationary. (This could be confirmed by looking at the ACF plot above).



The result of applying the two kinds of differencing is shown by the plot above. The series now looks stationary. This can be confirmed by looking at the ACF plot below:

```
> #Checking the ACF and PACF of the Twice-differenced Milk Production TS
> acf(ts(diff_2), main='ACF for Twice-differenced Milk Production',
lag.max=60)
```


ACF for Twice-differenced Milk Production
ACF
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
0
10
20
30
40
50
60
Lag


PACF for Twice-differenced Milk Production
Partial ACF
0.1
0.0
-0.1
-0.2
-0.3
-0.4
0
10
20
30
40
50
60
Lag

An initial model for forecasting can now be determined by examining the ACF and PACF above plots for the twice-differenced series.

- The ACF plot shows sharp cut-off; after lag 1 for a nonseasonal component and after lag 12 for the nonseasonal part. The large value at lag 13 is a natural consequence of using a product seasonal model.
- The PACF shows exponential decay at multiples of the seasonal period and what could be seen as co-sinusoidal behaviour at non-seasonal lags. This suggests using a product moving average model.

It can be concluded that the appropriate model would be

$$\text{ARIMA}(0,1,1) \times (0,1,1)_{12}.$$

$$\nabla\nabla_{12}12\,\mathrm{Y_t} \;=\; (1 + \theta_1 \mathrm{L})(1 + \Theta_1 \mathrm{L}^{12})\varepsilon_t$$

Fitting this model produces the following results.

```
> #Fitting the model
> library(forecast)
> fit = auto.arima(diff_2, approximation=FALSE,trace=FALSE)
> summary(fit)
Series: diff_2
ARIMA(1,0,0)(0,0,1)[12] with zero mean

Coefficients:
ar1     sma1
-0.2253  -0.6190
s.e.   0.0783   0.0626

sigma^2 = 53.38:  log likelihood = -530.11
AIC=1066.21   AICc=1066.37   BIC=1075.34

Training set error measures:
ME     RMSE      MAE MPE MAPE
Training set 0.05431126 7.258962 5.589855 NaN  Inf
MASE         ACF1
Training set 0.5117473 -0.008779199
```

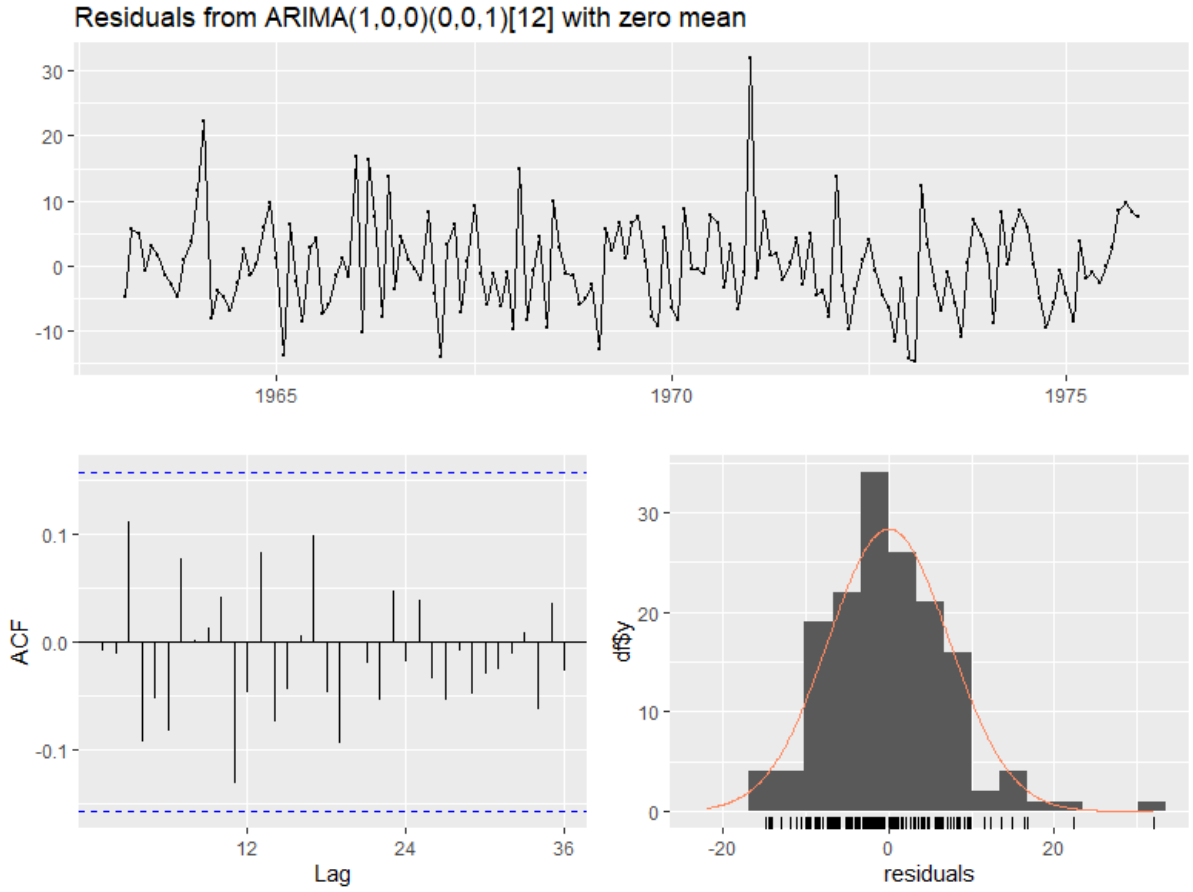
Both coefficients are significant, and when additional MA coefficients (simple or seasonal) are added to the model they are not significant. This model is a suitable candidate for making forecasts.

Now, we are going to make forecasts for milk production for the year 1976 as follows:

Before making forecasts, we will carry out a check of the model residuals. Diagnostic plots are shown below:

```
> #Checking Residuals
> checkresiduals(fit)

        Ljung-Box test

data:  Residuals from ARIMA(1,0,0)(0,0,1)[12] with zero mean
Q* = 16.896, df = 22, p-value = 0.7691

Model df: 2.   Total lags used: 24
```

```
> Box.test(diff_2)

        Box-Pierce test

data:  diff_2
X-squared = 6.9818, df = 1, p-value = 0.008234
```

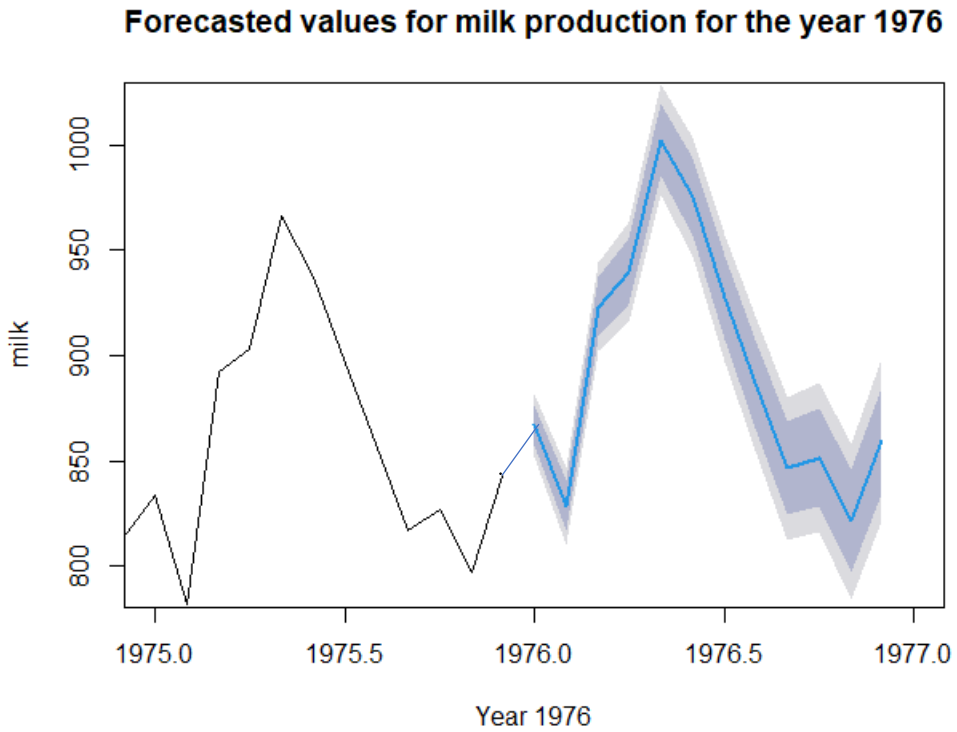


They show no problems with the residuals and that we can be confident that the forecasts and their standard errors are reasonable.

Table and Plot for forecasts for milk production for the year 1976 are shown below:

```
> #Forecasted values for milk production for the year 1976
> p = forecast(data,h = 12);p
         Point Forecast    Lo 80     Hi 80    Lo 95     Hi 95
Jan 1976       866.9927 857.3511  876.6344 852.2471  881.7384
Feb 1976       828.6343 816.7179  840.5508 810.4097  846.8590
Mar 1976       923.1375 909.3152  936.9599 901.9981  944.2770
Apr 1976       939.7290 924.2330  955.2250 916.0300  963.4280
May 1976      1002.2440 985.2379 1019.2501 976.2354 1028.2525
Jun 1976       975.4559 957.0630  993.8489 947.3263 1003.5856
Jul 1976       927.2905 907.6078  946.9733 897.1883  957.3927
Aug 1976       886.9869 866.0935  907.8803 855.0333  918.9406
Sep 1976       846.4018 824.3638  868.4397 812.6977  880.1058
Oct 1976       851.6340 828.5078  874.7601 816.2656  887.0024
Nov 1976       821.4789 797.3131  845.6446 784.5206  858.4371
Dec 1976       859.7003 834.5377  884.8629 821.2174  898.1832
```

```
> plot(p,xlim=c(1975,1977),ylim=c(790,1020), main='Forecasted values for milk
production for the year 1976', xlab='Year 1976', ylab='milk')
```



Interpretation of the forecast:

The forecasts show a predicted increase of roughly 3% to 4% in milk production. Because the lower confidence limit for the predictions is close to the values for 1975, we can be fairly sure that there will be small increase in production for 1976.