Chapter 6: Box-Jenkins Methodology

- *The methodology consists several steps:*

Identification: *(see page 107).*

Estimation: *(see page 89).*

Diagnosis: *(see page 116).*

Forecasting:

1. *Stationarity analysis.*
2. *Invertibility analysis.*
3. *Residual analysis.*
    1. *Randomness*

       $H_0: Residuals\ are\ random$
       $H_1: Residuals\ are\ not\ random$ *Runs test around zero.*

    2. *Constant variance*
    3. *Follow the normal distribution*

       $H_0: Residuals\ follow\ normal\ distribution$
       $H_1: Residuals\ do\ not\ follow\ normal\ distribution$

       Shapiro test

    4. *Test if the residual of the fitted model up to lag k are uncorrelated*

       $H_0: \rho_1 = \rho_2 = \cdots = \rho_k = 0$
       $H_1: at\ least\ two\ \neq 0$ $the\ Ljung - Box\ test$

4. *Fitting the lower model.*
5. *Fitting the higher model.*

The theoretical forms of ACF and PACF for the models: $AR(p)$, $MA(q)$ *and* $ARMA(p,q)$

Model	$ACF\,(\rho_k)$	$PACF\ (\phi_{kk})$
$AR(1)$	*Approach zero exponentially or in a sinusoidal manner*	*Cut off completely after the 1st time lag*
$AR(2)$	*Approach zero exponentially or in a sinusoidal manner*	*Cut off completely after the 2nd time lag*
$AR(p)$	*Approach zero exponentially or in a sinusoidal manner*	*Cut off completely after time lag* p
$MA(1)$	*Cut off completely after the 1st time lag*	*Approach zero exponentially or in a sinusoidal manner*
$MA(2)$	*Cut off completely after the 2nd time gap*	*Approach zero exponentially or in a sinusoidal manner*
$MA(q)$	*Cut off completely after a time gap* q	*Approach zero exponentially or in a sinusoidal manner*
$ARMA(p,q)$	*Gradually approaching zero after (q-p) lags exponentially or in a sinusoidal manner*	*Gradually approaching zero after (p-q) lags exponentially or in a sinusoidal manner*

<u>Steps of Time series analysis:</u>

1. *Checking stationarity. (Make an appropriate transformations if need)*

Differencing can help stabilise the mean of a time series by removing changes in the level of a time series . Box-Cox can help make the variance constant.(R code)

```
(lambda <-BoxCox.lambda(x.D1))
x.B<-BoxCox(x.D1,lambda)
```

2. *Checking ACF and PACF.*
3. *Checking the coefficients.*
4. *Diagnose the Residuals.*
   a. *Random, PAC, L-Jung Box and normality graphs*
   b. *Randomness test*
   c. *Normality test*
5. *If we have more than model, we use AIC or BIC to compare.*
6. *Forecasting.*

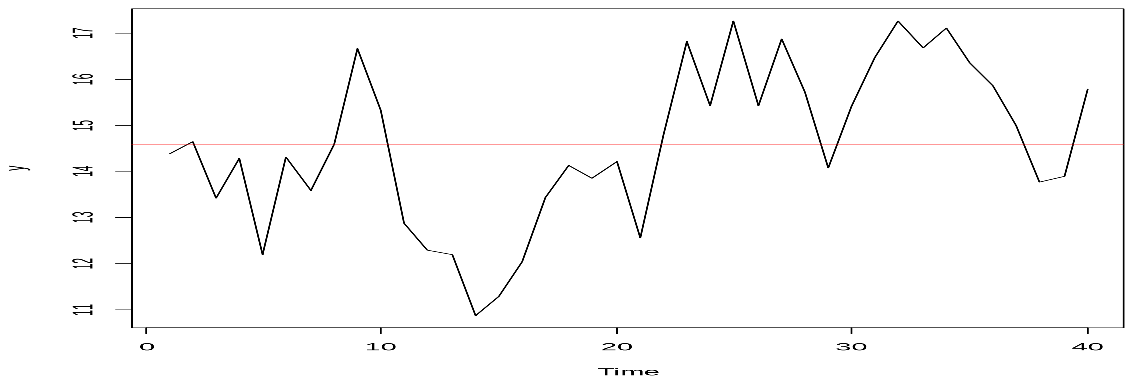
***Exercise 1* using R:**

the packages used in time series analysis

```
install.packages("forecast")
install.packages("tseries")
install.packages("randtests")
install.packages("astsa")
install.packages("lmtest")
library(forecast)
library(tseries)
library(randtests)
library(astsa)
library(lmtest)
```

1. Checking stationarity of the series:

```
> d=read.csv("ex.csv")
> d=c(14.383,14.649,13.416,14.288,12.201,14.307
,13.586,14.592,16.660,15.332,12.884
,12.296,12.201,10.873,11.290,12.049
,13.435,14.137,13.852,14.213,12.562
,14.801,16.812,15.427,17.268,15.427
,16.869,15.712,14.080,15.408,16.471
,17.268,16.679,17.116,16.357,15.863
,14.991,13.776,13.890,15.787)
> d=ts(d,frequency=1)
> plot(d)
> #plot.ts(d)
```



The data seems to be stationary in the mean.

```
> shapiro.test(d)
        Shapiro-Wilk normality test

data:  d
```

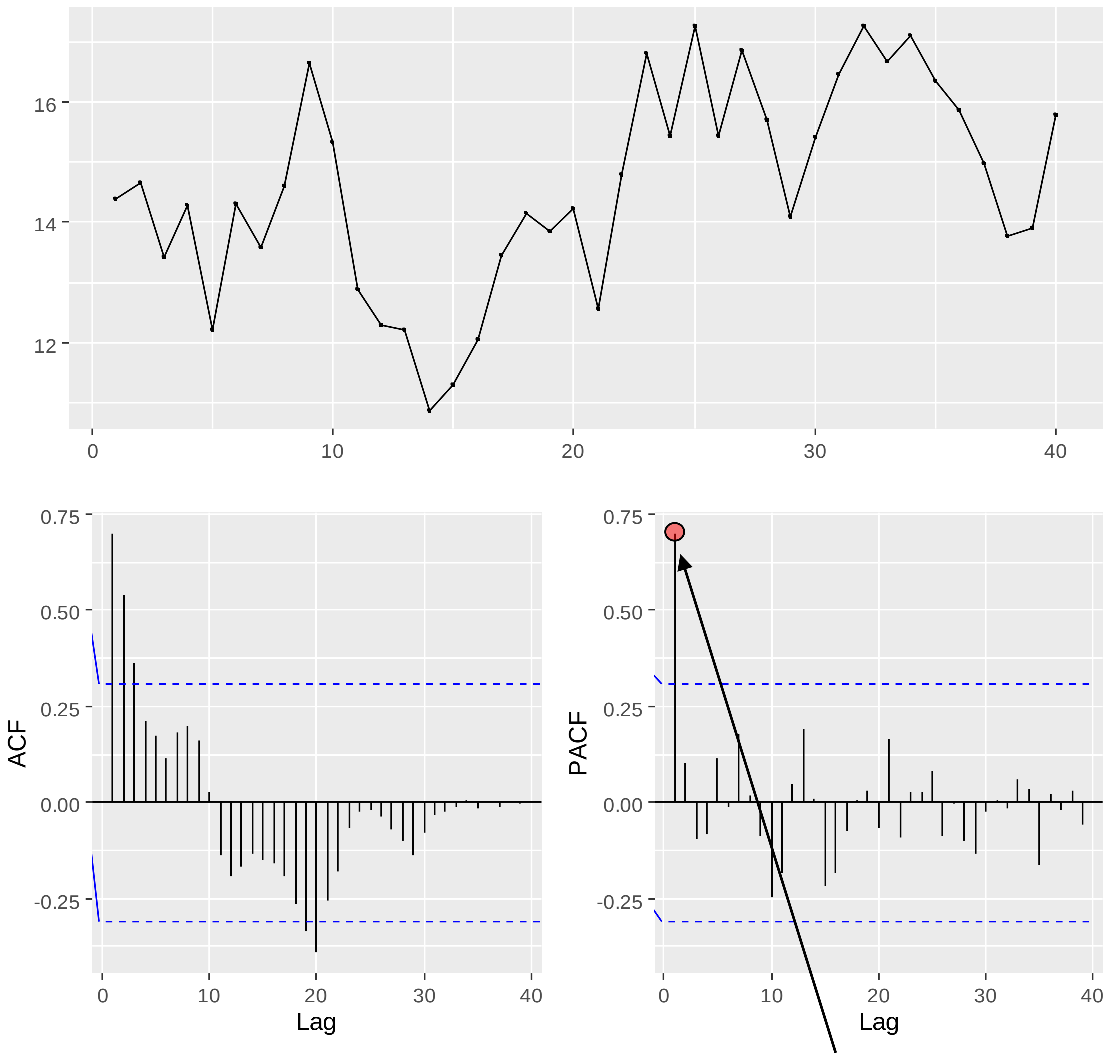
W = 0.9688, p-value = 0.3296

The data seems to be stationary in the variance.

2. Finding the appropriate model using ACF and PACF plot:

```
> ggtsdisplay(d,lag.max=20)
> #acf (d,lag.max=20)
> #pacf(d,lag.max=20)
```

Or



Approach zero exponentially or in a sinusoidal manner

Cut off completely after the 1st time lag

$ARIMA(1,0,0)$

- *Determine the model:*

$ARIMA(1,0,0)$:

```
> fit1=arima(d,order=c(1,0,0))
> fit1
Call:
arima(x = d, order = c(1, 0, 0))
Coefficients:
         ar1  intercept
      0.6909    14.6309
s.e.  0.1094     0.5840
sigma^2 estimated as 1.447:  log likelihood = -64.47,  aic = 134.94
```

3. Testing the coefficients for:

```
> coeftest(fit1)

z test of coefficients:
          Estimate  Std. Error  z value   Pr(>|z|)
ar1       0.69090   0.10945     6.3126    2.744e-10 ***
intercept 14.63095  0.58402     25.0523   < 2.2e-16 ***
```

$$H_0: \phi_1 = 0 \quad vs \quad H_1: \phi_1 \neq 0$$

p-value = 2.744e-10< 0.05, we reject H_0

$ARIMA(0,0,1)$:

```
> fit2=arima(d,order=c(0,0,1))
> coeftest(fit2)
z test of coefficients:
          Estimate  Std. Error  z value  Pr(>|z|)
ma1       0.55701   0.12509     4.453    8.467e-06 ***
intercept 14.58814  0.33366     43.721   < 2.2e-16 ***
```

$$H_0: \theta_1 = 0 \quad vs \quad H_1: \theta_1 \neq 0$$

p-value = 8.467e-06 < 0.05, means, we reject H_0

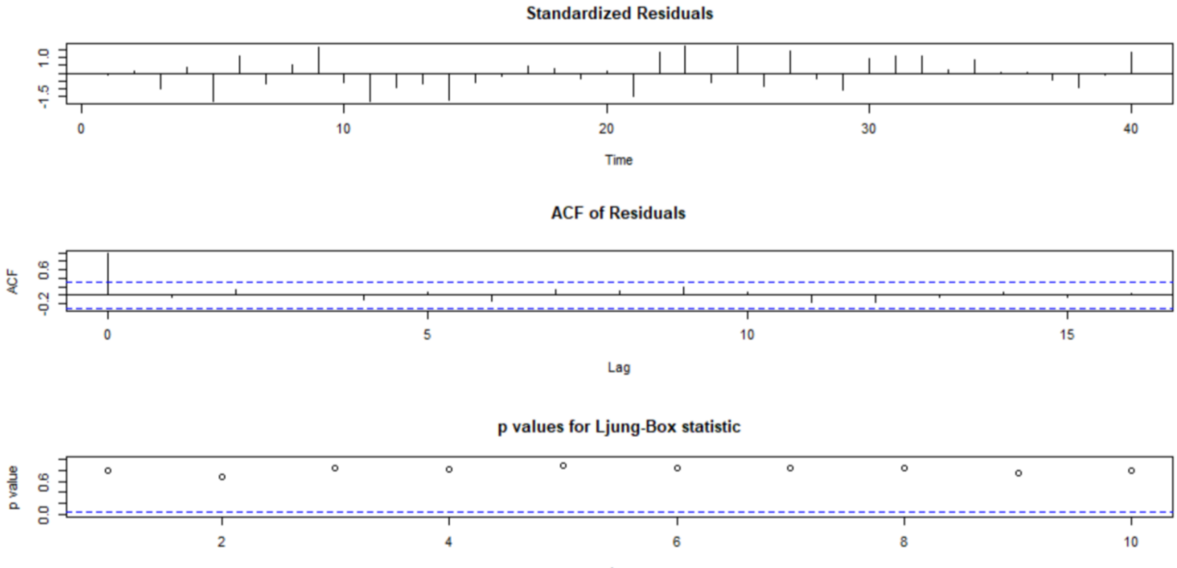
we have the models, $ARIMA(1,0,0)$ *and* $ARIMA(0,0,1)$

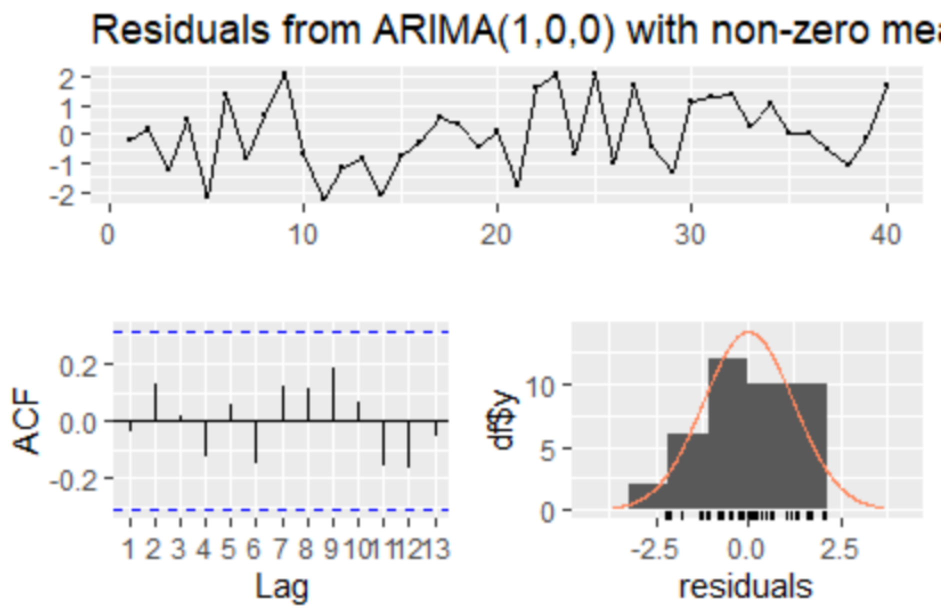
4. Diagnosing the Residuals.

$ARIMA(1,0,0)$

a) graphs.

```
> tsdiag(fit1)
> checkresiduals(fit1)
```





- *The residuals are random around the zero*
- *All p-values of the L-jung Box test > 0.05*
- *The ACF of the Residuals are zeros*
- *The residuals seem to be normal*

b) randomness test

$H_0: Residuals\ are\ random$
$H_1: Residuals\ are\ not\ random$

```
> fit1=arima(d,order=c(1,0,0))
> runs.test(fit1$r)
        Runs Test
data:  fit1$r
statistic = -1.6018, runs = 16, n1 = 20, n2 = 20, n = 40, p-value = 0.7487
alternative hypothesis: nonrandomness
```

p-value= 0.7487 > 0.05, means, we accept H_0 (the residuals are random)

$H_0: E(\varepsilon_t) = 0\ vs\ \ H_1: E(\varepsilon_t) \neq 0$

c) normality test:

$H_0: Residuals\ follow\ normal$
$H_1: Residuals\ do\ not\ follow\ normal$

```
> shapiro.test(fit1$r)

        Shapiro-Wilk normality test
data:  fit1$r
W = 0.96633, p-value = 0.2737
```

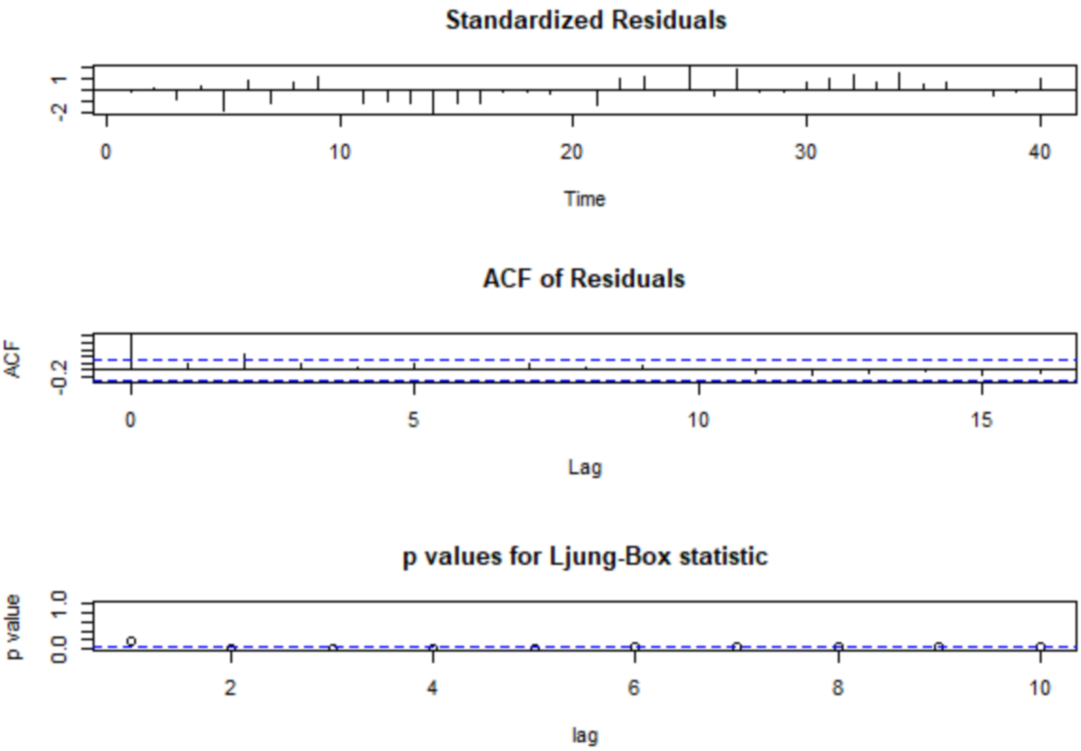
Accept H_0(Residuals follow normal)

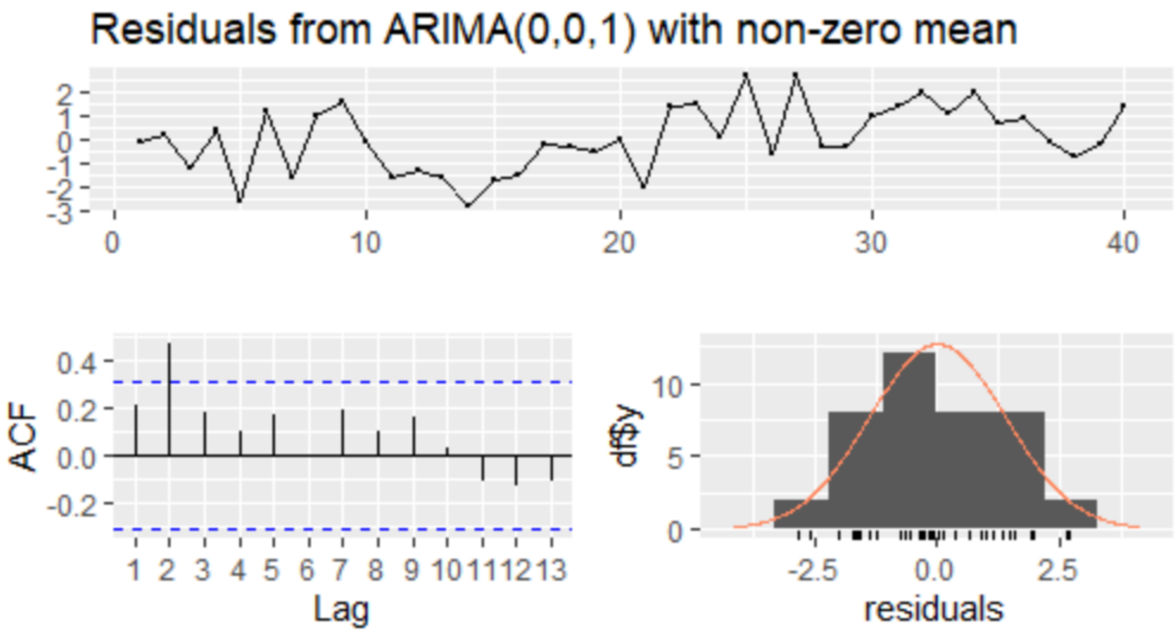
ARIMA(1,0,0) Model :

$$\hat{y} = 14.6309+0.6909\hat{y}_{t-1}+\varepsilon_t$$

$ARIMA(0,0,1)$

```
> fit2=arima(d,order=c(0,0,1))
> tsdiag(fit2)
> checkresiduals(fit2)
```





- *The residuals are random around the zero (Except for ρ_2, it could be a random error)*
- *Almost ll p-values of the L-jung Box test < 0.05*
- *The ACF of the Residuals are zeros*
- *The residuals seem to be normal*

The fitted model is not adequate

```
> fit2=arima(d,order=c(0,0,1))
> runs.test(fit2$r)
        Runs Test
data:  fit2$r
statistic = -0.96108, runs = 18, n1 = 20, n2 = 20, n = 40, p-value = 0.3365
alternative hypothesis: nonrandomness
```

p-value= 0.3365 > 0.05, means, we accept H_0 (the residuals are random)

$$H_0: E(\varepsilon_t) = 0 \ vs \ H_1: E(\varepsilon_t) \neq 0$$

Testing the normality of the residuals:

$$H_0: Residuals\ follow\ normal$$
$$H_1: Residuals\ do\ not\ follow\ normal$$

```
alternative hypothesis: two-sided

> shapiro.test(fit2$r)

        Shapiro-Wilk normality test
data:  fit2$r
W = 0.97718, p-value = 0.586
```

Accept H_0(Residuals follow normal)

ARIMA(0,0,1) Model :

$$\hat{y} = 14.58814 + \varepsilon_t - 0.55701\ \hat{\varepsilon}_{t-1}$$

More codes for checking normality and Randomness:

```
> hist(fit1$r)
> qqnorm(fit1$r)
> qqline(fit1$r)
> acf(fit1$r)
> pacf(fit1$r)
```

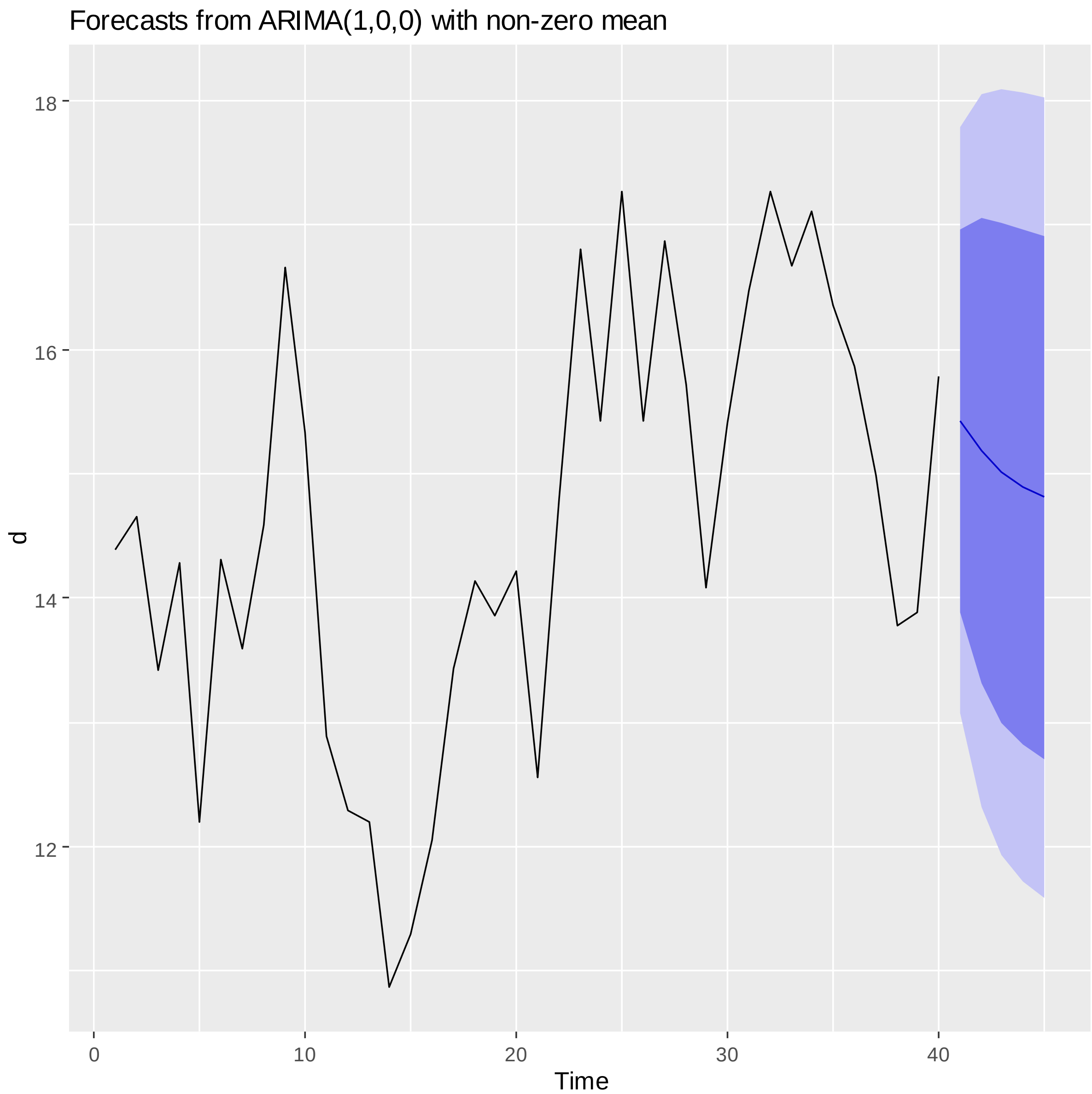
5. *Using AIC or BIC to choose between $ARIMA(1,0,0)$ and $ARIMA(0,0,1)$*

```
> fit1=arima(d,order=c(1,0,0))
> fit2=arima(d,order=c(0,0,1))

> fit1$aic
   [1] 134.9385
> fit2$aic
   [1] 144.9162
```

For forecasting we will use the model with lowest AIC. which is $ARIMA(1,0,0)$

6. Forecasting using $ARIMA(1,0,0)$*:*

```
> f=forecast(fit1, h=5)
> autoplot(f)
> f
Point    Forecast  Lo 80     Hi 80     Lo 95     Hi 95
41       15.42967  13.88817  16.97116  13.07215  17.78718
42       15.18278  13.30916  17.05641  12.31732  18.04825
43       15.01221  12.99927  17.02515  11.93369  18.09074
44       14.89436  12.81822  16.97051  11.71917  18.06956
45       14.81294  12.70729  16.91859  11.59263  18.03325
```



Exercise 2

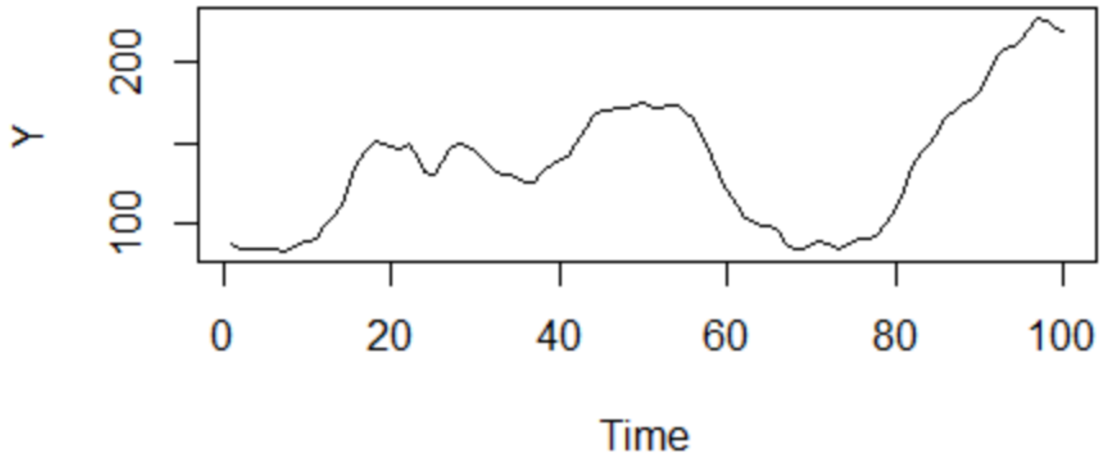
For WWWusage data, is a time series of the number of users on a server every minute for 100 minutes, do the following:

1- Plot the series and check its stationarity in mean and variance.
2- plot the ACF and PACF , suggest a preliminary model for the data.
3- Fit the suggested models and get acquainted with the R output.
4- Predict number of users for next 10 minutes.

Solution:

1. Checking stationarity of the series:

```
> data1 = read.csv("C:/ WWWusage.txt", sep="",header=TRUE)
> Y=ts(data1$Y,frequency=1)
> plot(Y)
```



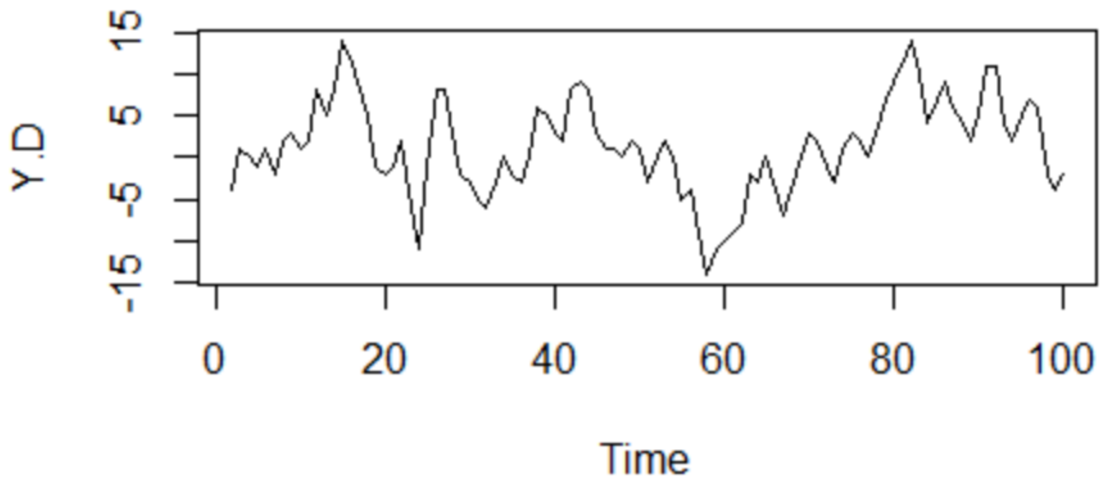
The data seems to be not stationary in the mean.

```
> shapiro.test(Y)

        Shapiro-Wilk normality test

data:  Y
W = 0.9373, p-value = 0.0001325
```

The data is not stationary in the variance.

- *First starting by taking the first difference:*

```
> Y.D<-diff(Y,difference=1)
> plot(Y.D)
```



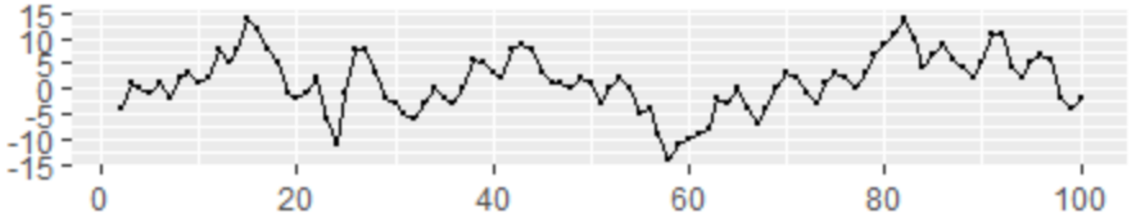
The data now is seems to be stationary in the mean

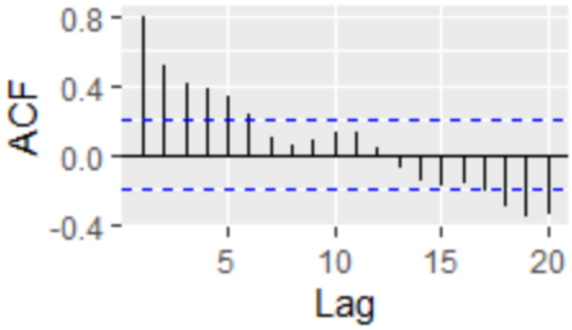
```
Shapiro-Wilk normality test


data:  Y.D
W = 0.9891, p-value = 0.5997
```

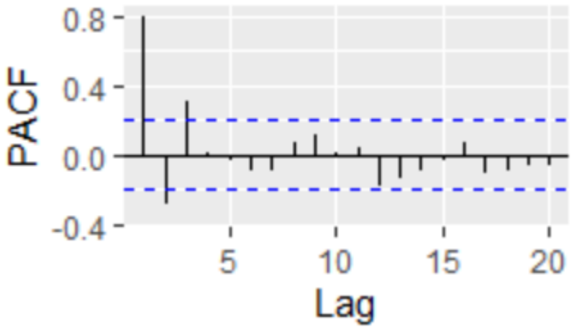
The data now is stationary in the variance.

2. Finding the appropriate model using ACF and PACF plot:

```
> ggtsdisplay(Y.D,lag.max=20)
```







Approach zero exponentially or in a sinusoidal manner

Cut off completely after the 3rd time lag

$\boxed{ARIMA(3,1,0)}$

- *Determine the model:*

$ARIMA(3,1,0)$:

```
> fit1=arima(Y,order=c(3,1,0))

> fit1

Call:

arima(x = Y, order = c(3, 1, 0))

Coefficients:

         ar1      ar2      ar3

      1.1513  -0.6612  0.3407

s.e.  0.0950   0.1353  0.0941

sigma^2 estimated as 9.363:  log likelihood = -252,  aic = 511.99
```

3. Testing the coefficients for:

```
> coeftest(fit1)

z test of coefficients:

    Estimate Std. Error z value  Pr(>|z|)
ar1  1.151340   0.094984 12.1214 < 2.2e-16 ***
```

```
ar2 -0.661227   0.135263 -4.8885 1.016e-06 ***
ar3  0.340713   0.094146  3.6190 0.0002957 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

1) For ϕ_1:

$$H_0: \phi_1 = 0 \quad vs \quad H_1: \phi_1 \neq 0$$

p-value = 2.2e-16 < *0.05, we reject* H_0

2) For ϕ_2:

$$H_0: \phi_2 = 0 \quad vs \quad H_1: \phi_2 \neq 0$$

p-value 1.016e-06< *0.05, we reject* H_0

3) For ϕ_3:

$$H_0: \phi_3 = 0 \quad vs \quad H_1: \phi_3 \neq 0$$

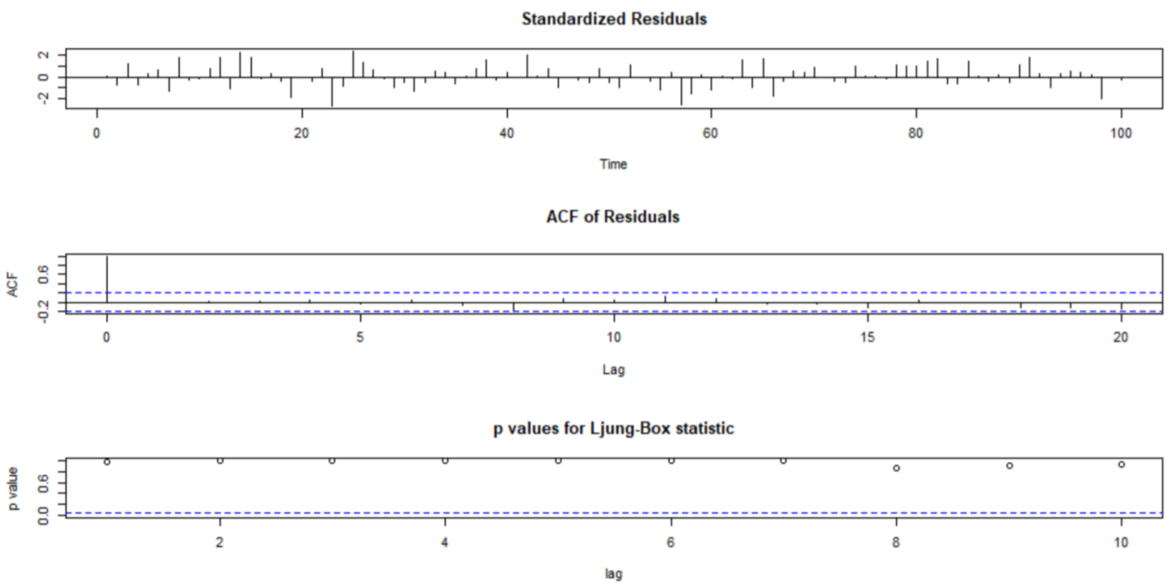
p-value 0.0002957< *0.05, we reject* H_0

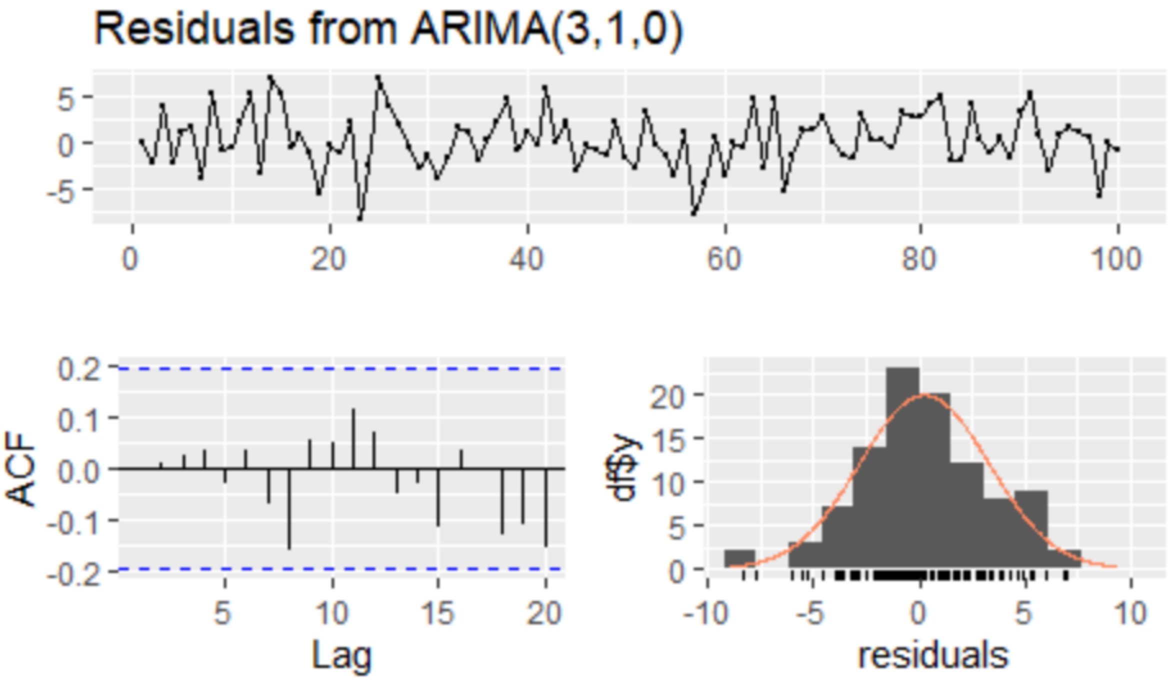
4. Diagnosing the Residuals.

$ARIMA(1,0,0)$

a) graphs.

```
> tsdiag(fit1)
> checkresiduals(fit1)
```





- *The residuals are random around the zero*
- *All p-values of the L-jung Box test > 0.05*
- *The ACF of the Residuals are zeros*
- *The residuals seem to be normal*

b) randomness test

H_0: *Residuals are random*
H_1: *Residuals are not random*

```
> runs.test(fit1$r)
Runs Test

data:  fit1$r
statistic = 0.20102, runs = 52, n1 = 50, n2 =
50, n = 100, p-value = 0.8407
alternative hypothesis: nonrandomness
```

p-value= 0.8407 > 0.05, means, we accept H_0 (the residuals are random)

$$H_0: E(\varepsilon_t) = 0 \; vs \; H_1: E(\varepsilon_t) \neq 0$$

c) normality test:

$H_0: Residuals\ follow\ normal$
$H_1: Residuals\ do\ not\ follow\ normal$

```
> shapiro.test(fit1$r)

	Shapiro-Wilk normality test

data:  fit1$r
W = 0.98913, p-value = 0.5951
```

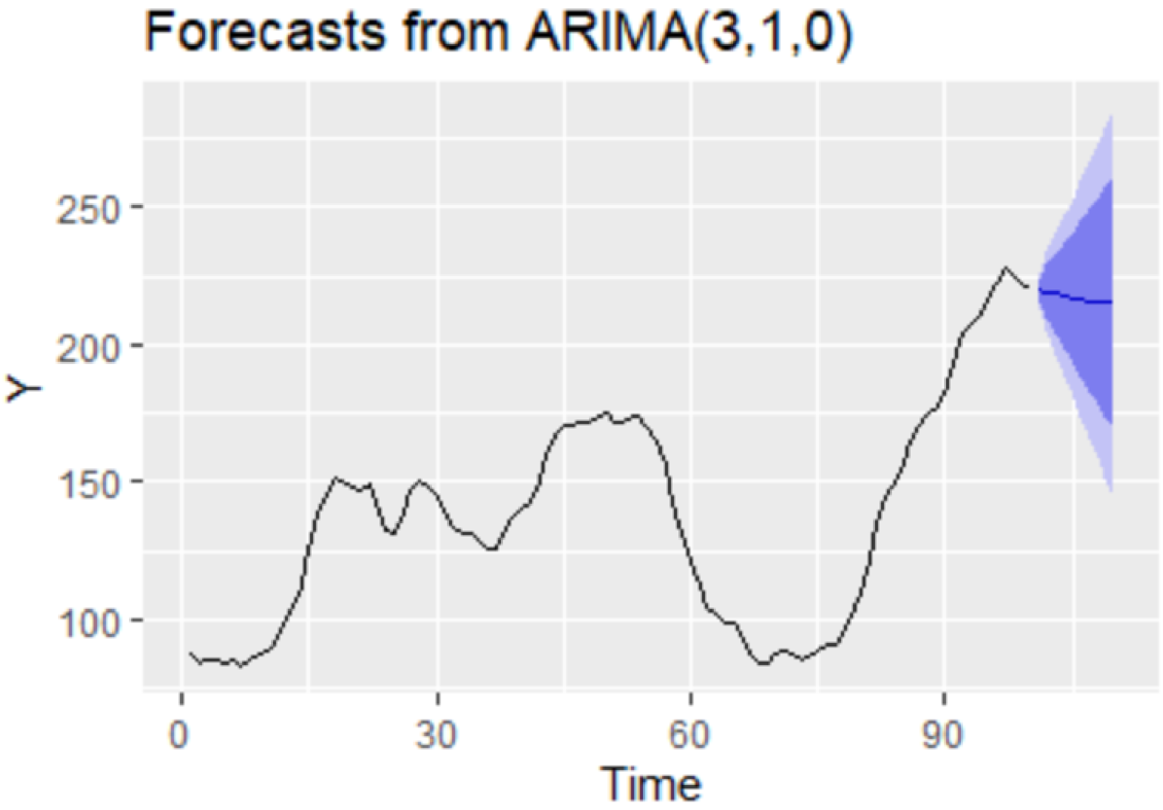
Accept H_0(Residuals follow normal)

ARIMA(3,1,0) Model :

$$(1 - \phi_1 B - \phi_2 B^2 - \phi_3 B^3)(1 - B)\, y_t = \epsilon_t \gg$$
$$\gg (1 - 1.1513B + 0.6612B^2 - 0.3407B^3)(1 - B)\, y_t = \epsilon_t$$

6. Forecasting:

```
> f=forecast(fit1, h=10)
> autoplot(f)
> f
Point Forecast    Lo 80    Hi 80    Lo 95
101       219.6608 215.7393 223.5823 213.6634
102       219.2299 209.9265 228.5332 205.0016
103       218.2766 203.8380 232.7151 196.1947
104       217.3484 198.3212 236.3756 188.2489
105       216.7633 193.2807 240.2458 180.8498
106       216.3785 188.3324 244.4246 173.4858
107       216.0062 183.3651 248.6473 166.0860
108       215.6326 178.5027 252.7624 158.8474
109       215.3175 173.8431 256.7919 151.8879
110       215.0749 169.3780 260.7719 145.1874
       Hi 95
101 225.6582
102 233.4581
103 240.3585
104 246.4479
105 252.6768
106 259.2713
107 265.9264
108 272.4178
109 278.7471
110 284.9625
```


Forecasts from ARIMA(3,1,0)
250
200
150
100
Y
0
30
60
90
Time