Chapter 6: Box-Jenkins Methodology

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* 1^{st} *time lag*
$AR(2)$	*Approach zero exponentially or in a sinusoidal manner*	*Cut off completely after the* 2^{nd} *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* 1^{st} *time lag*	*Approach zero exponentially or in a sinusoidal manner*
$MA(2)$	*Cut off completely after the* 2^{nd} *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* $(q\text{-}p)$ *lags exponentially or in a sinusoidal manner*	*Gradually approaching zero after* $(p\text{-}q)$ *lags exponentially or in a sinusoidal manner*

Steps of Time series analysis:

1. *Checking stationarity. (Make an appropriate transformation 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 of Box-Cox transformation:

```
(lambda <-BoxCox.lambda( x ))
```

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

2. *Checking ACF and PACF and Finding the appropriate model.*
3. *Checking the coefficients.*

   Test for significance of the estimated parameters.

4. *Diagnose the Residuals.*
   a. *Random, PAC, L-Jung Box and normality graphs.*
   b. *Residuals are uncorrelated.*

   *Test if the residual of the fitted model up to lag k are uncorrelated. We examine the correlation up to lag **12, 24, 36 and 42.***

$$\begin{matrix} H_0: \rho_1 = \rho_2 = \cdots = \rho_k = 0 \\ H_1: at\ least\ two\ \neq 0 \end{matrix} \quad the\ Ljung - Box\ test$$

Also, autocorrelation function (ACF & PACF) must be free of any spikes (all the bars are within the blue band).

   c. *Randomness test by use Runs test.*

   The randomness of the residuals is tested by ***Runs test*** around zero.

$$\begin{matrix} H_0: Residuals\ are\ random \\ H_1: Residuals\ are\ not\ random \end{matrix} \quad \textit{(Runs test around zero).}$$

   d. *Normality test by use **Shapiro test.***

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

   e. *Mean of the residuals is zero.*

   Use t-test : $H_o: E(\varepsilon_t) = 0 \quad vs \quad H_A: E(\varepsilon_t) \neq 0$

5. *If we have more than model, we use AIC or BIC to compare.*

6- Forecasting.

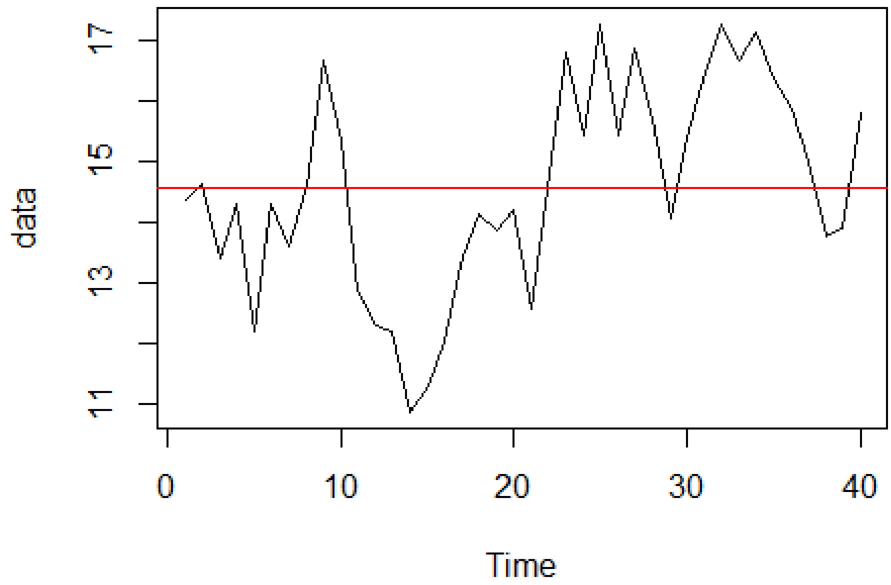
Exercise 1 using R:

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 stationary of the series:

```
d<- read.csv(file.choose(),header = T)
d=ts(d)  #time-series objects
plot(d) ; abline(h =mean(d),col="red")
```



The data seems to be stationary in the mean.

- **Normality test.**

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

data:  d
W = 0.9688, p-value = 0.3296
```

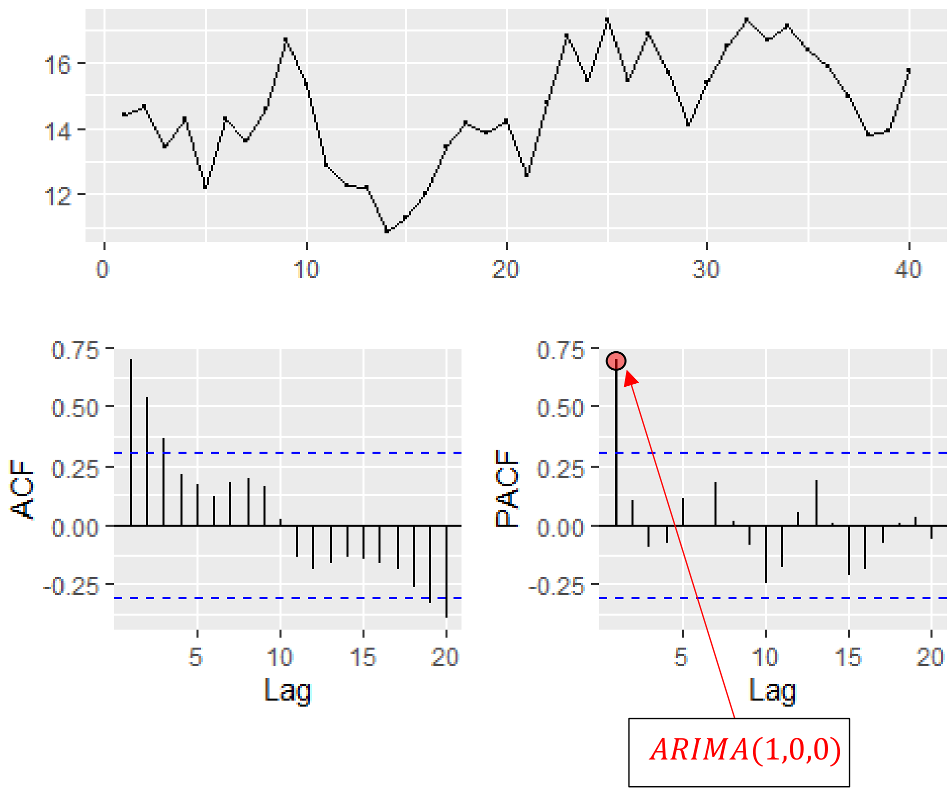
$H_0: data\ follow\ normal \quad v.s \quad H_1: data\ do\ not\ follow\ normal$

$p-value > 0.05, \quad we\ Acept\ H_0$

The data seems to be stationary in the variance.

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

```
ggtsdisplay(d,lag.max=20 )
```




```
# Or use
#* acf(d,lag.max=20)
#* pacf(d,lag.max=20)
```

The ACF Approach zero exponentially or in a sinusoidal manner. The PACF Cut off completely after the **1**st time lag, so we suggest the model ARIMA(1,0,0)

ARIMA(1,0,0) model

```
(model1=arima(d,order=c(1,0,0)))
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 ARIMA(1,0,0):

```
coeftest(model1)


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 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

$p - value \;=\; 2.744e^{-10} < \; 0.05, we\ reject\ H_0.$

The constant term and coefficient of AR1 is significantly different from zero ,thus must be kept in the model.

ARIMA(1,0,0) Model : $\hat{y} = 4.5224 + 0.6909\,\hat{y}_{t-1} + \varepsilon_t$

$c = \mu\left(1 - \emptyset_1 - \emptyset_2 - \cdots - \emptyset_p\right) = 14.6309(1 - 0.6909) = 4.5224$

- Not: ARIMA model in R

$$\begin{array}{ccccc} (1-\phi_1 B - \cdots - \phi_p B^p) & (1-B)^d y_t & = & c + (1+\theta_1 B + \cdots + \theta_q B^q)\varepsilon_t \\ \uparrow & \uparrow & & \uparrow \\ \text{AR}(p) & d \text{ differences} & & \text{MA}(q) \end{array} \qquad (8.2)$$

R uses a slightly different parameterisation:

$$(1-\phi_1 B - \cdots - \phi_p B^p)(y'_t - \mu) = (1+\theta_1 B + \cdots + \theta_q B^q)\varepsilon_t, \qquad (8.3)$$

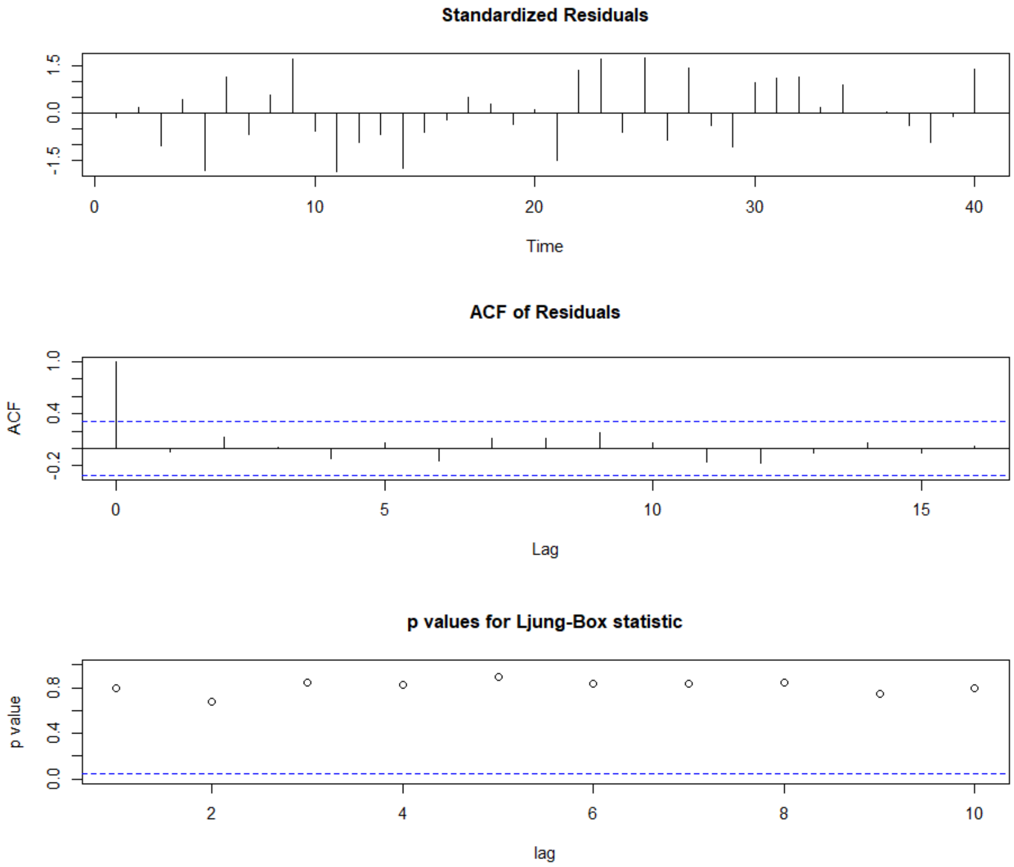
where $y'_t = (1-B)^d y_t$ and μ is the mean of y'_t. To convert to the form given by (8.2), set $c = \mu(1 - \phi_1 - \cdots - \phi_p)$.

4- Diagnosing the Residuals of model ARIMA(1,0,0)

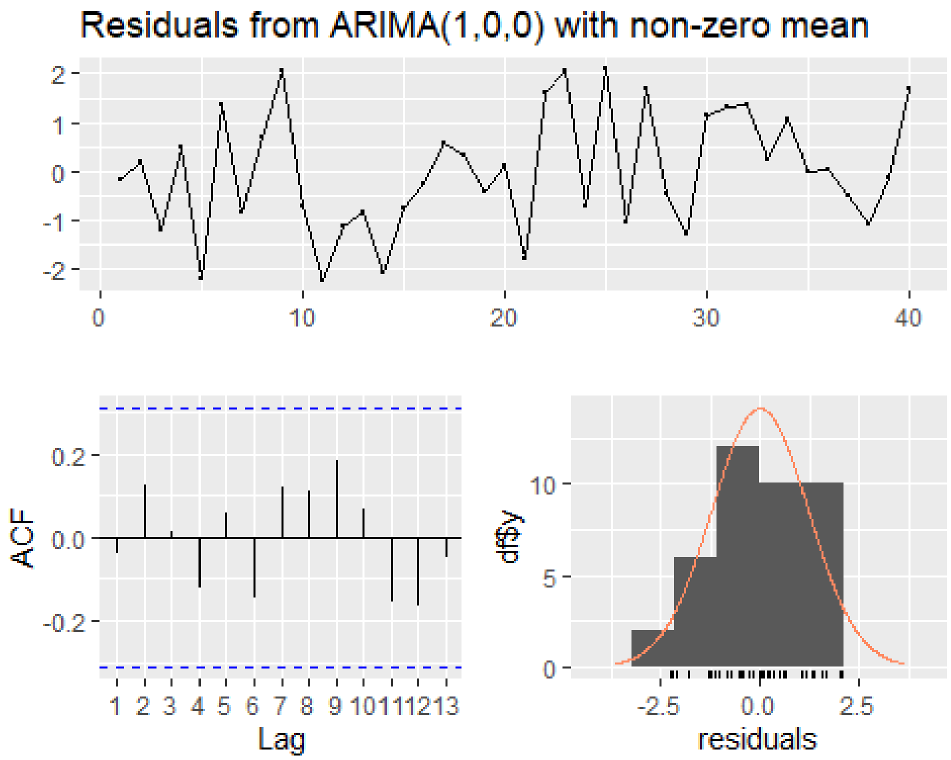
a. graphs.

b. Residuals are uncorrelated.

```
tsdiag(model1)
```




```
checkresiduals(model1, lag= 12)
```



```
Ljung-Box test
data:  Residuals from ARIMA(1,0,0) with non-zero mean
Q* = 9.2425, df = 11, p-value = 0.5995

Model df: 1.   Total lags used: 12

checkresiduals(model1, lag= 24,plot=FALSE)


    Ljung-Box test
data:  Residuals from ARIMA(1,0,0) with non-zero mean
Q* = 22.899, df = 23, p-value = 0.4667

Model df: 1.   Total lags used: 24

checkresiduals(model1, lag= 36,plot=FALSE)


    Ljung-Box test
data:  Residuals from ARIMA(1,0,0) with non-zero mean
Q* = 29.715, df = 35, p-value = 0.721

Model df: 1.   Total lags used: 36
```

- Plot of residuals with time: The residuals are random around the zero.
- All p-values of the Ljung-Box test > 0.05. The residuals are uncorrelated.
- The ACF of the Residuals are zeros.
- Histogram: The residuals seem to be normal .

c. Randomness test

```
runs.test(model1$r)


    Runs Test
data:  model1$r
statistic = 0.32036, runs = 22, n1 = 20, n2 = 20, n = 40, p-value =
0.7487
alternative hypothesis: nonrandomness
```

$$\boldsymbol{H_0: Residuals\ are\ random \quad v.s \quad H_1: Residuals\ are\ not\ random.}$$

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

d. Normality test

```
shapiro.test(model1$residuals)


    Shapiro-Wilk normality test
data:  model1$residuals
W = 0.96633, p-value = 0.2737
```

$$H_0: Residuals\ follow\ normal \quad v.s \quad H_1: Residuals\ do\ not\ follow\ normal.$$

p-value= 0.2737 > 0.05 we accept H_0 , which means that the residuals are Normally distributed.

e. Mean of the residuals is zero.

```
t.test(model1$r)


    One Sample t-test

data:  model1$r
t = 0.031149, df = 39, p-value = 0.9753
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.3835818  0.3955808
sample estimates:
  mean of x
0.005999503
```

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

p-value > 0.05 , which means the acceptance of the zero-mean hypothesis of the residuals.

If we suggest other model ARIMA(0,0,1)

```
(model2=arima(d,order=c(0,0,1)))
Call:
arima(x = d, order = c(0, 0, 1))

Coefficients:
         ma1  intercept
      0.5570    14.5881
s.e.  0.1251     0.3337

sigma^2 estimated as 1.87:  log likelihood = -69.46,  aic = 144.92

BIC(model2)

[1] 149.9828
```

3- Testing the coefficients for ARIMA(0,0,1):

```
coeftest(model2)
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 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$\boldsymbol{H_0: \theta_1 = 0 \quad vs \quad H_1: \theta_1 \neq 0}$$

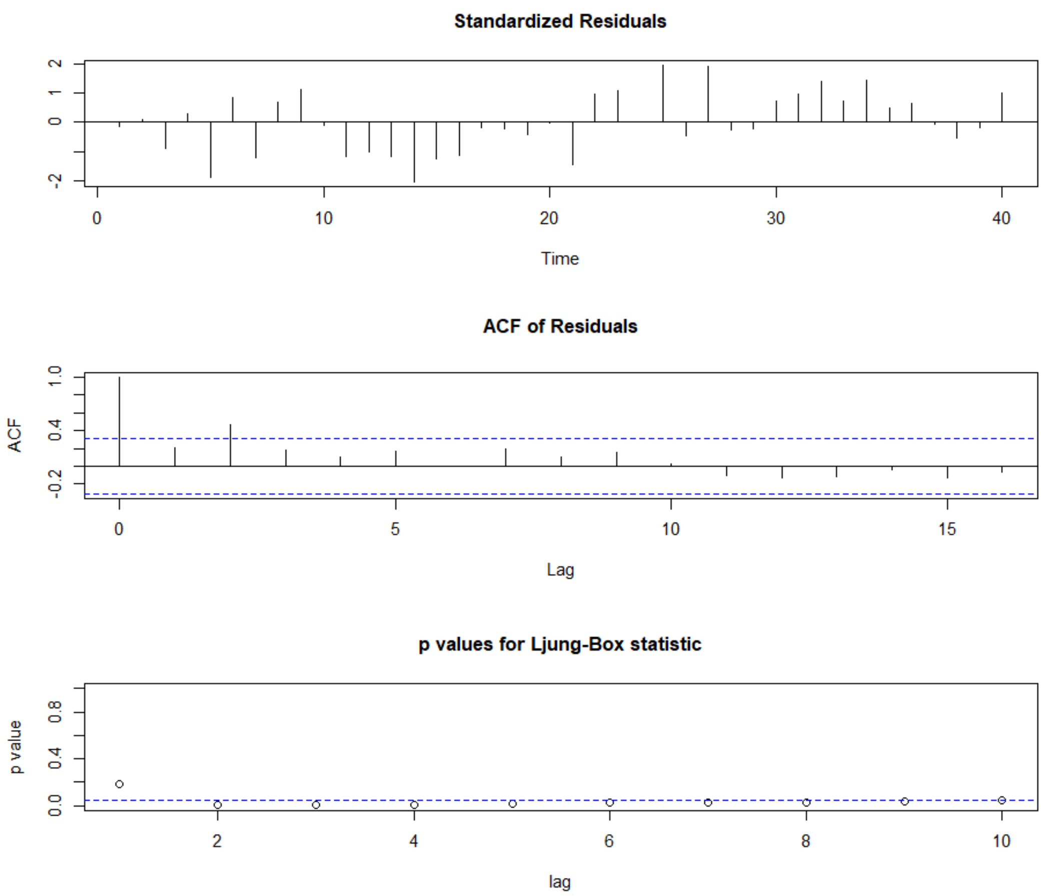
p-value = 8.467e-06 < 0.05, means, we reject H_0

The constant term and coefficient of MA1 is significantly different from zero ,thus must be kept in the model.

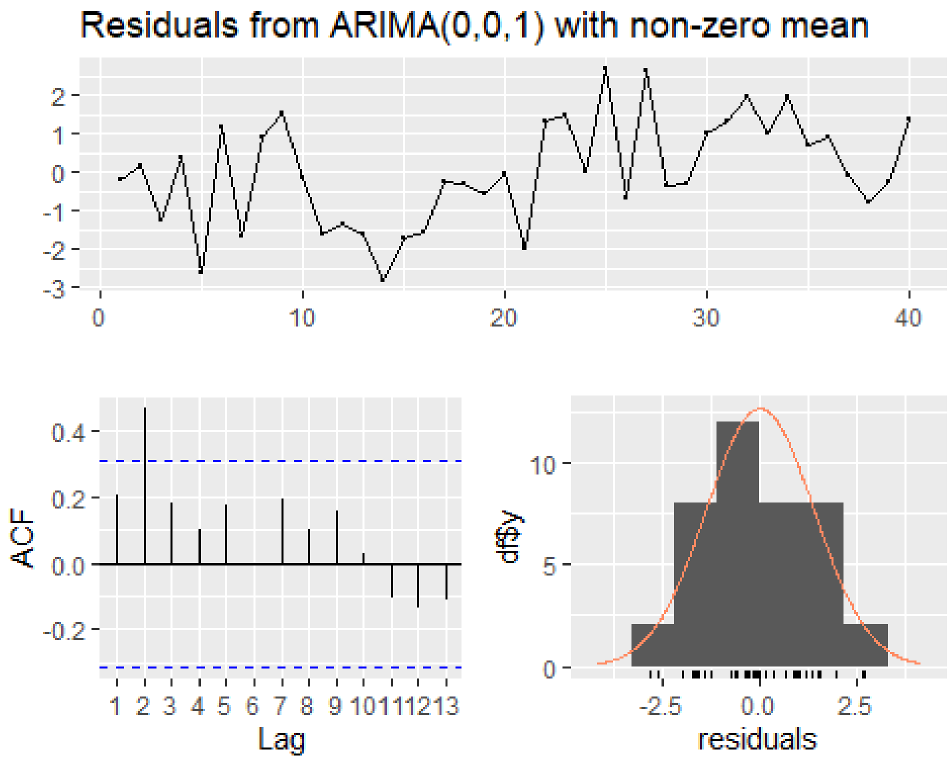
4- Diagnosing the Residuals of model ARIMA(0,0,1)

a.graphs. b.Residuals are uncorrelated.

```
tsdiag(model2)
```



```
checkresiduals(model1, lag= 12)
```




```
 Ljung-Box test
data:  Residuals from ARIMA(0,0,1) with non-zero mean
Q* = 20.109, df = 11, p-value = 0.04387
Model df: 1.   Total lags used: 12
```

```
checkresiduals(model2, lag= 24,plot=FALSE)
    Ljung-Box test
data:  Residuals from ARIMA(0,0,1) with non-zero mean
Q* = 43.852, df = 23, p-value = 0.005478
Model df: 1.   Total lags used: 24
```

```
checkresiduals(model2, lag= 36,plot=FALSE)
    Ljung-Box test
data:  Residuals from ARIMA(0,0,1) with non-zero mean
Q* = 49.797, df = 35, p-value = 0.05004
Model df: 1.   Total lags used: 36
```

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

***The fitted model is not adequate*.**

c. Randomness test

```
runs.test(model2$r)
    Runs Test
data:  model2$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)

d. Normality test

```
shapiro.test(model2$residuals)
    Shapiro-Wilk normality test
data:  model2$residuals
W = 0.97718, p-value = 0.586
```

p-value= 0.58 > 0.05 , *Accept H_0(Residuals follow normal)*

e. Mean of the residuals is zero.

```
t.test(model2$r)
    One Sample t-test
data:  model2$r
t = 0.033222, df = 39, p-value = 0.9737
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.4355887  0.4501366
sample estimates:
  mean of x
0.007273971
```

p-value > 0.05 , which means the acceptance of the zero-mean hypothesis of the residuals.

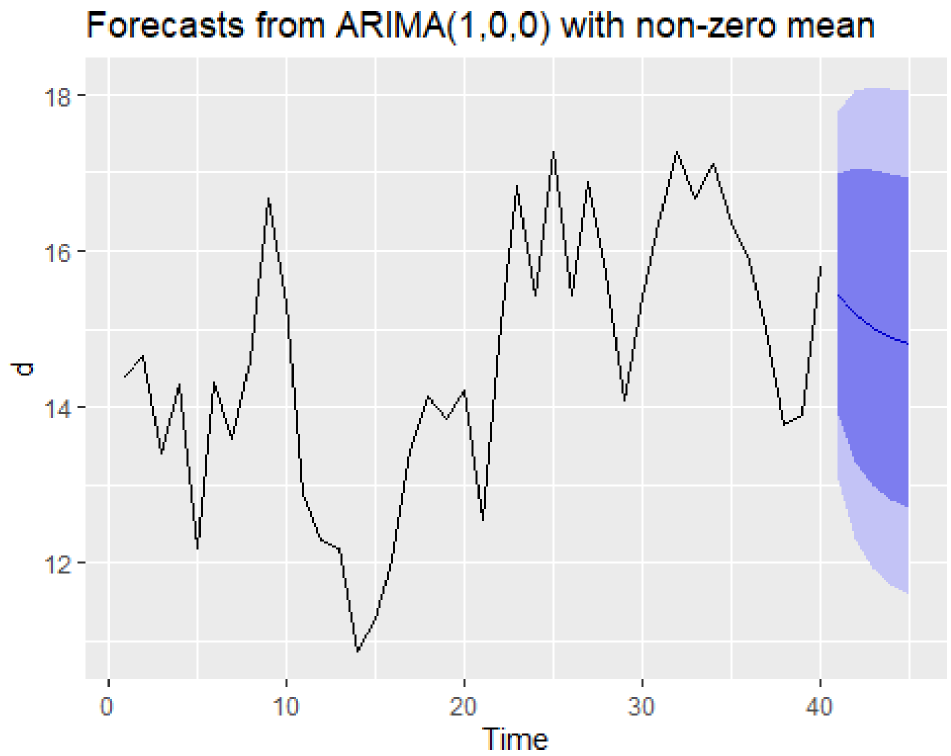
5- Using AIC or BIC to choose between ARIMA(1,0,0) and ARIMA(0,0,1)

```
model1$aic
[1] 134.9385
model2$aic
[1] 144.9162
BIC(model1)
[1] 140.0051
BIC(model2)
[1] 149.9828
```

The best model with lowest AIC and BIC . Which is ARIMA(1,0,0,0)

6- Forecasting using ARIMA(1,0,0):

```
(f=forecast(model1, h=5))

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

autoplot(f)
```



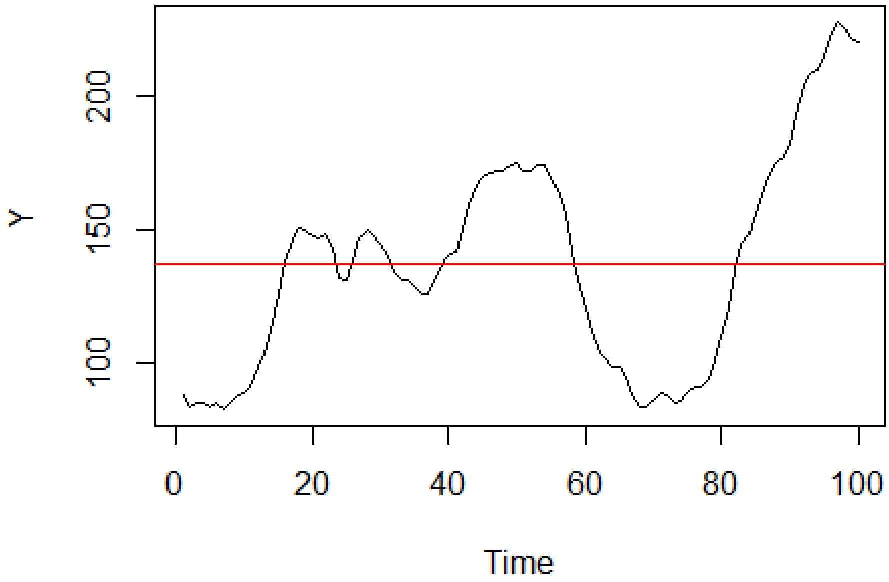
Exercise 2:

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.

Exercise 2 using R: WWWusage data.

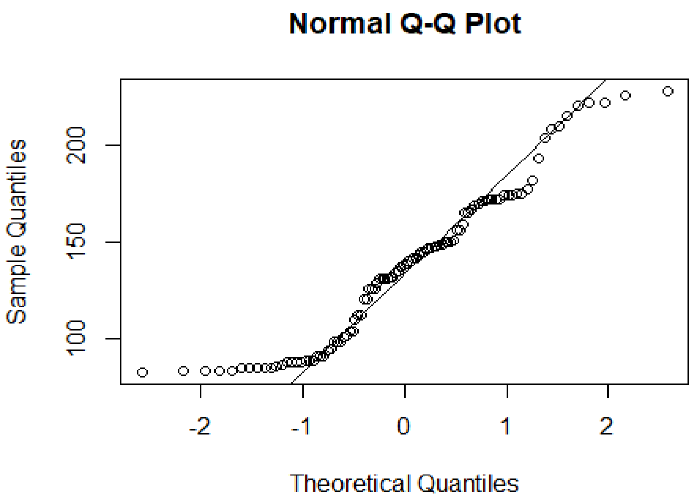
```
rm(list=ls())
 data <- read.csv(file.choose(),header = T)
 Y=ts(data)
 plot(Y) ; abline(h =mean(Y),col="red")
```



The data seems to be not stationary in the mean and variance.

- **Normality test:**

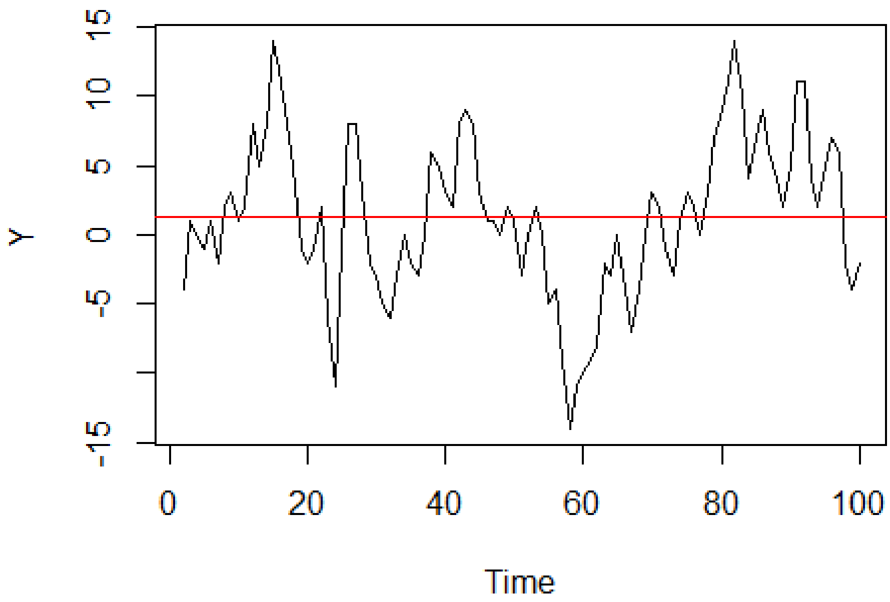
```
shapiro.test(Y)
    Shapiro-Wilk normality test
data:  Y
W = 0.9373, p-value = 0.0001325

qqnorm(Y); qqline(Y)
```



The data is not stationary in the variance. p-value = 0.00013 < 0.05 , we reject H_0, which indicates to instability in the variance. Also, qq-plot doesn't look normally distributed.

- **First starting by taking the first difference:**

```
Y.D<-diff(Y,difference=1)
plot(Y.D) ; abline(h =mean(Y.D),col="red")
```



The data now seems to be stationary in the mean.

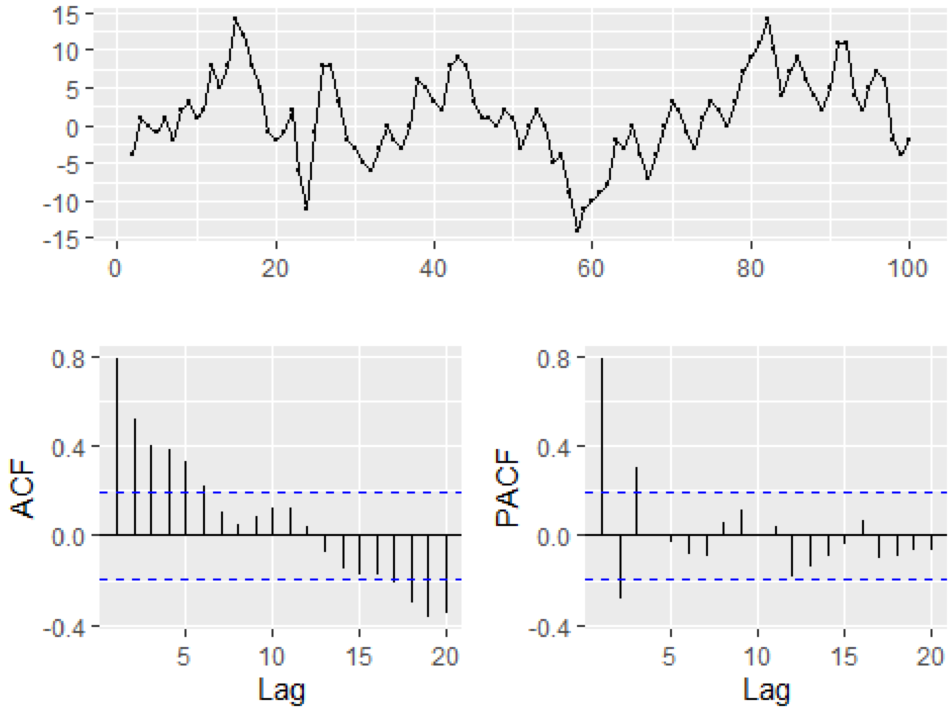
- **Normality test:**

```
shapiro.test(Y.D)


    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)
```



The ACF Approach zero exponentially or in a sinusoidal manner. The PACF Cut off completely after the 3rd time lag, so we suggest the model ARIMA(3,1,0)

ARIMA(3,1,0) model:

```
(model1=arima(Y,order=c(3,1,0)))
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 ARIMA(3,1,0):

```
coeftest(model1)
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

Notice here the coefficient of AR1 ,AR2 and AR3 are significantly different from zero and hence must be retained in the model.

ARIMA(3,1,0) Model :

$$(\mathbf{1} - \boldsymbol{\phi_1 B} - \boldsymbol{\phi_2 B^2} - \boldsymbol{\phi_3 B^3})(\mathbf{1} - \boldsymbol{B})\, \boldsymbol{y_t} = \boldsymbol{\epsilon_t} \quad \gg$$

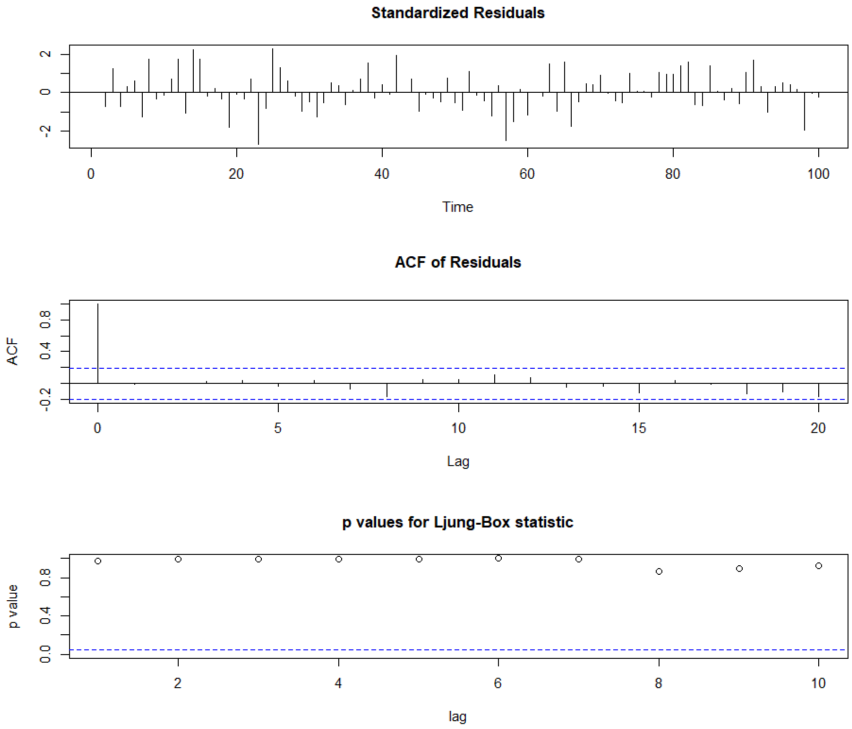
$$\gg (\mathbf{1} - \mathbf{1.1513}\boldsymbol{B} + \mathbf{0.6612}\boldsymbol{B^2} - \mathbf{0.3407}\boldsymbol{B^3})(\mathbf{1} - \boldsymbol{B})\, \boldsymbol{y_t} = \boldsymbol{\epsilon_t}$$

4- Diagnosing the Residuals of model ARIMA(3,1,0):

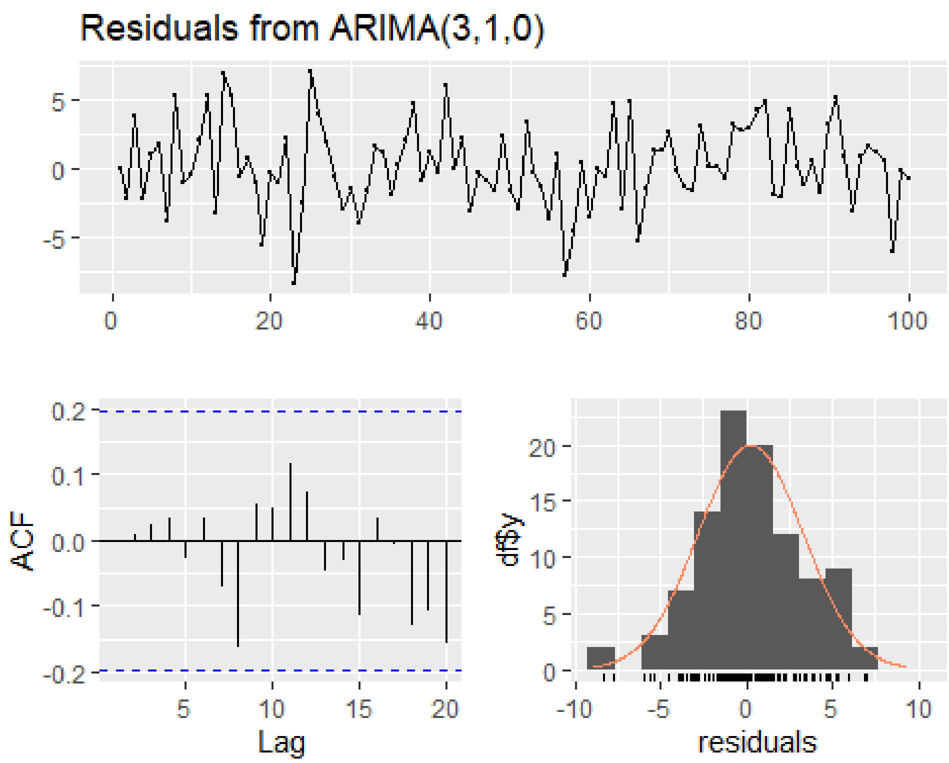
a. graphs.

b. Residuals are uncorrelated.

```
tsdiag(model1)
```




```
checkresiduals(model1, lag= 12)
```




```
   Ljung-Box test
data:  Residuals from ARIMA(3,1,0)
Q* = 6.6597, df = 9, p-value = 0.6725

Model df: 3.   Total lags used: 12
```

```
checkresiduals(model1, lag= 24,plot=FALSE)


    Ljung-Box test
data:  Residuals from ARIMA(3,1,0)
Q* = 20.393, df = 21, p-value = 0.4965

Model df: 3.   Total lags used: 24

checkresiduals(model1, lag= 36,plot=FALSE)


    Ljung-Box test
data:  Residuals from ARIMA(3,1,0)
Q* = 31.19, df = 33, p-value = 0.5574

Model df: 3.   Total lags used: 36

checkresiduals(model1, lag= 42,plot=FALSE)


    Ljung-Box test
data:  Residuals from ARIMA(3,1,0)
Q* = 38.516, df = 39, p-value = 0.4918

Model df: 3.   Total lags used: 42
```

- Plot of residuals with time: The residuals are random around the zero.
- All p-values of the Ljung-Box test > 0.05. The residuals are uncorrelated.
- The ACF of the Residuals are zeros.
- Histogram: The residuals seem to be normal .

c. Randomness test

```
runs.test(model1$r)


    Runs Test
data:  model1$r
statistic = 0.20102, runs = 52, n1 = 50, n2 = 50, n = 100,
 p-value =0.8407
alternative hypothesis: nonrandomness
```

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

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

d. Normality test

```
shapiro.test(model1$residuals)


    Shapiro-Wilk normality test
data:  model1$residuals
W = 0.98913, p-value = 0.5951
```

p-value= 0.595 > 0.05, Accept H_0 , which means that the Residuals follow normal.

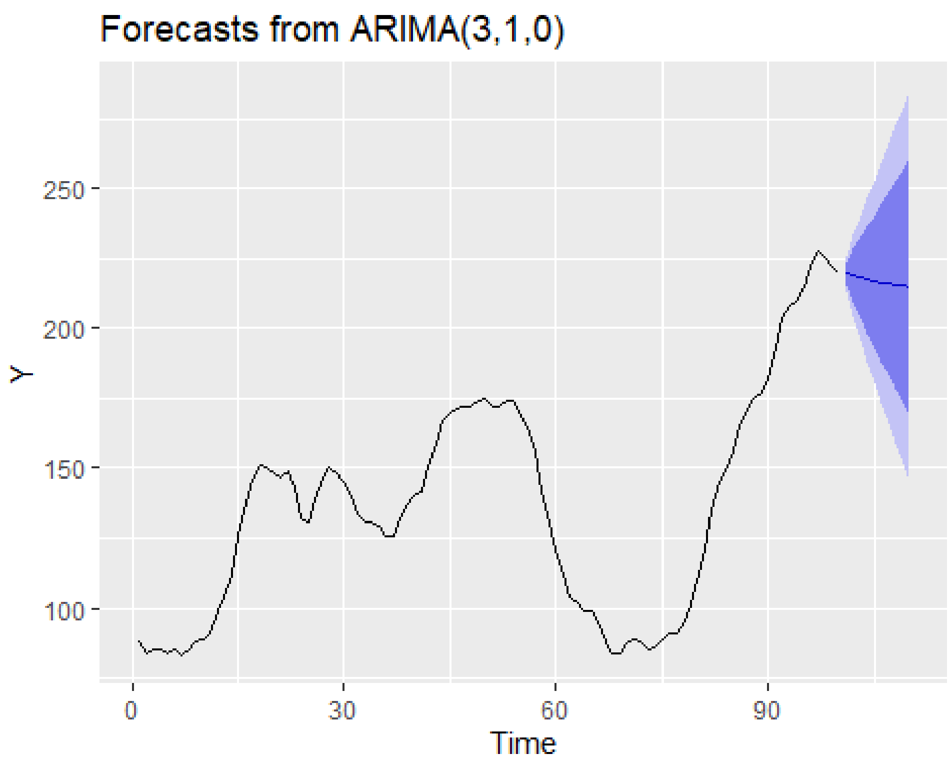
e. Mean of the residuals is zero.

```
t.test(model1$r)


    One Sample t-test
data:  model1$r
t = 0.75573, df = 99, p-value = 0.4516
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.3748326  0.8360087
sample estimates:
mean of x
 0.230588
```

p-value =0.4516 > 0.05 , which means the acceptance of the zero-mean hypothesis of the residuals.

6- Forecasting :

```
(f=forecast(model1, h=10))

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

autoplot(f)
```



- **plot the original time series as a black line, with the forecast values as a pink line:**

```
fits<-fitted(model1)
plot(Y,col = "black",lwd=2)
points(fits, col = "deeppink",type = "l",lwd=2,lty = 2)
points(f$mean,col = "blue",type = "l",lwd=3)
```