Time series with R

V. Lefieux

LIBRARIES

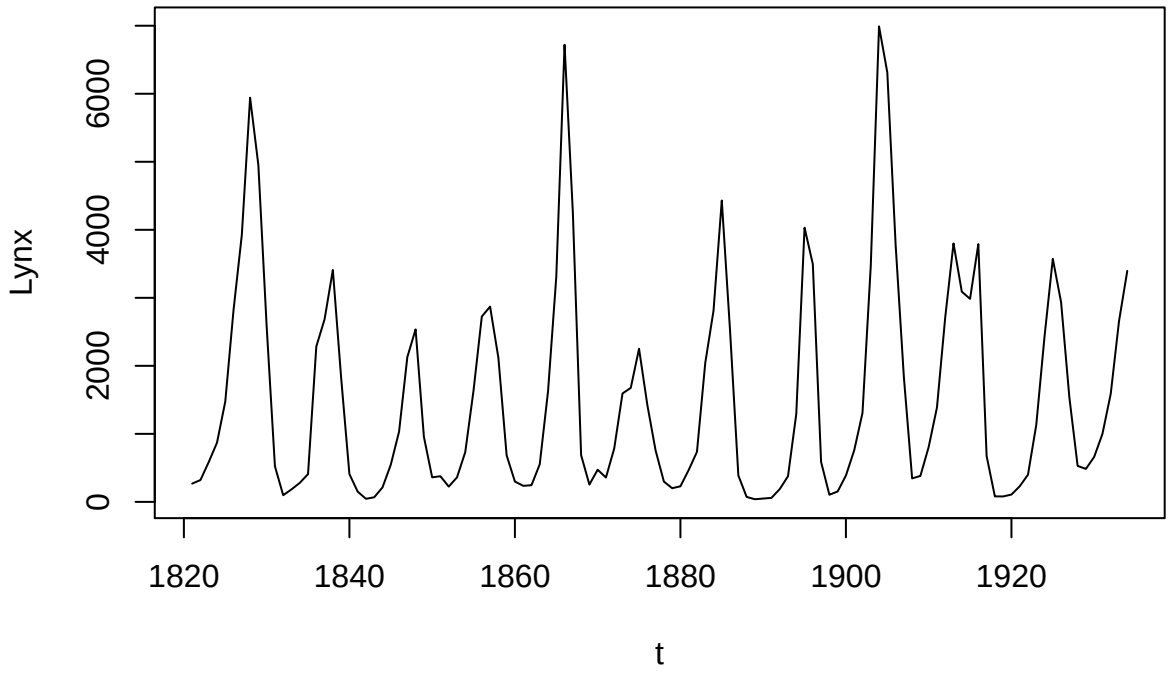
```
library(caschrono) # SARIMA modeling
library(TSA) # Periodogram
```

TIME SERIES EXAMPLES

In the same order than the course slides

Series *lynx*: numbers of annual lynx trappings in Canada from 1821 to 1934

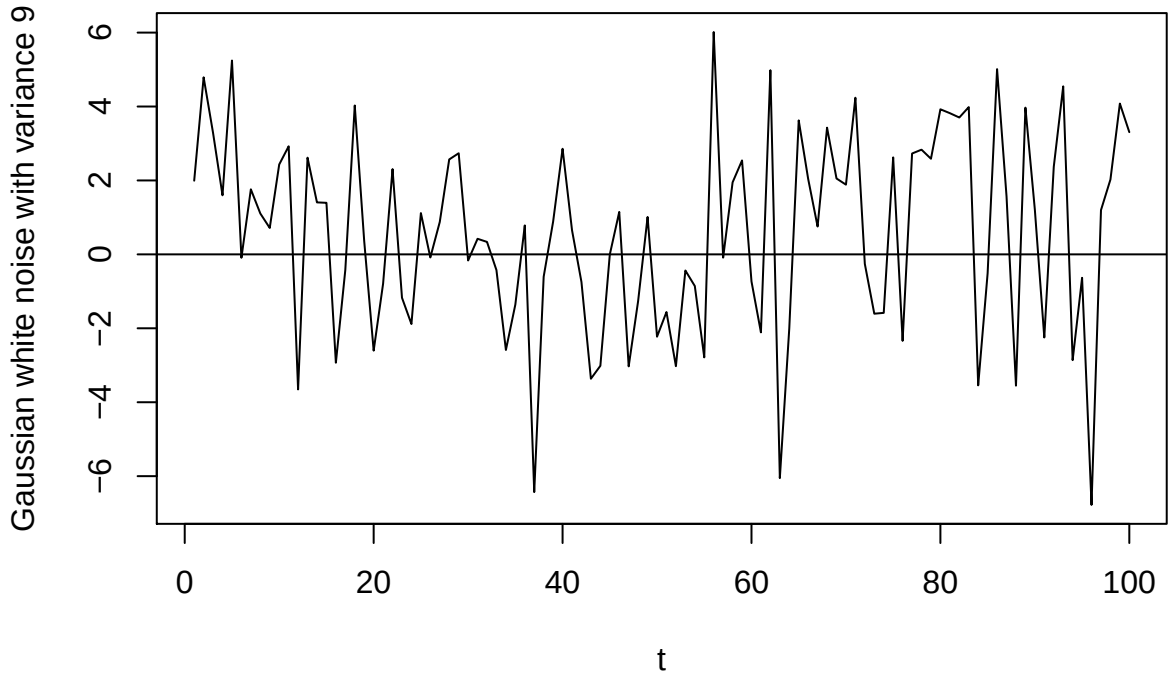
```
plot(lynx,xlab="t",ylab="Lynx")
```



Gaussian white noise with variance 9 ($n = 100$)

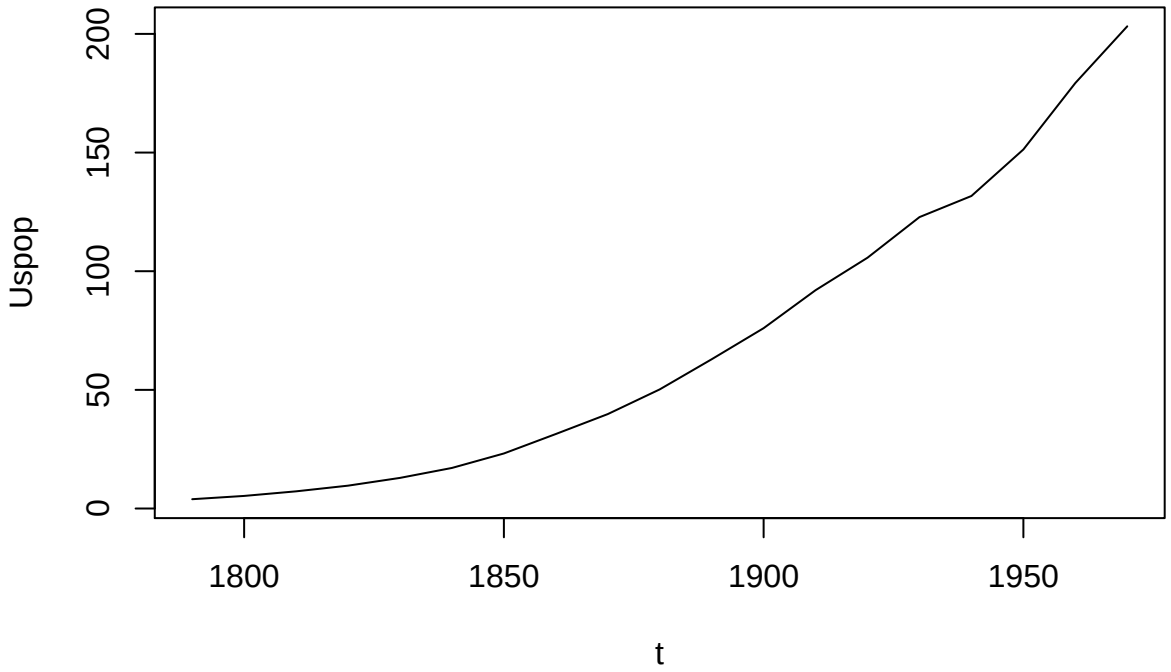
The seed is fixed (at 1789) to share results.

```
set.seed(1789)
plot(ts(rnorm(100,sd=3),start=1,end=100),xlab="t",ylab="Gaussian white noise with variance 9")
abline(h=0)
```



Series *uspop*: 10-years USA population from 1790 to 1990 (in millions) from 1790 to 1990

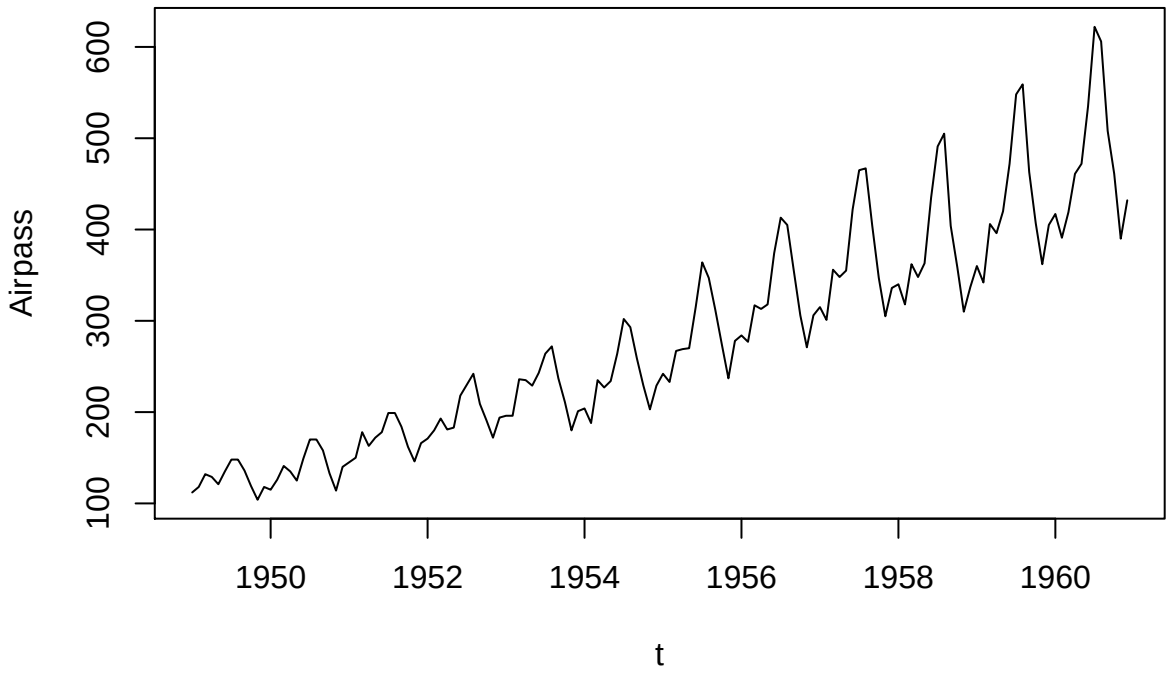
```
plot(uspop,xlab="t",ylab="Uspop")
```



Series *airpass*: monthly number international airline passengers (in millions) from January 1949 to December 1960

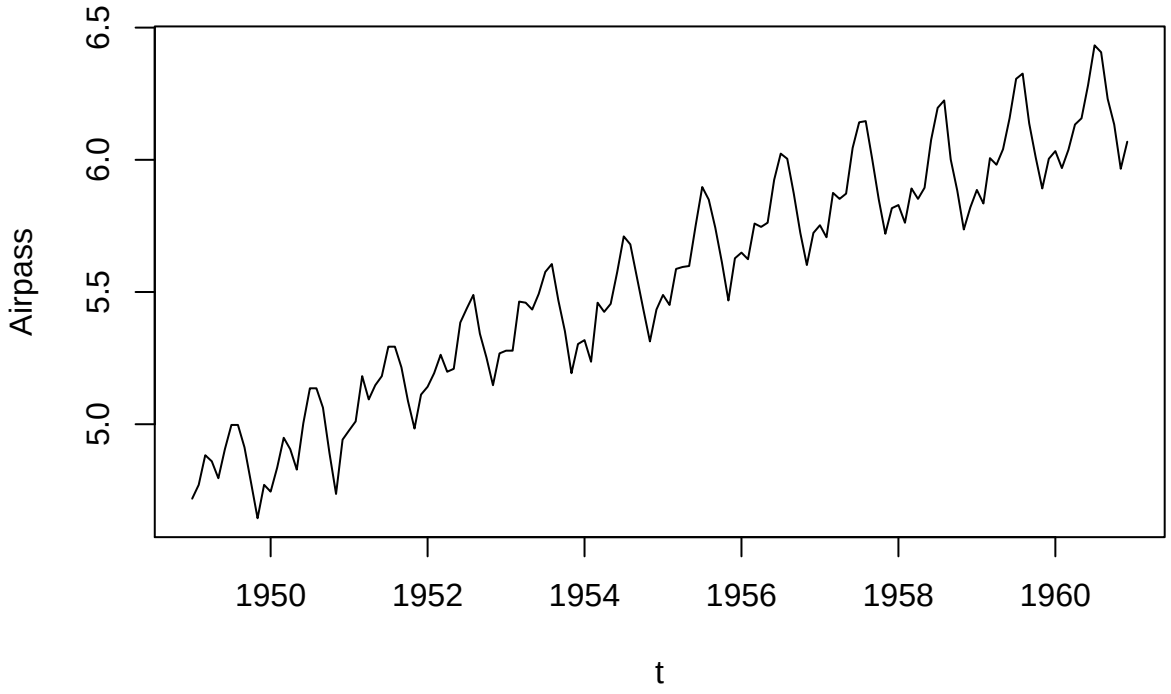
Time series

```
plot(AirPassengers,xlab="t",ylab="Airpass")
```



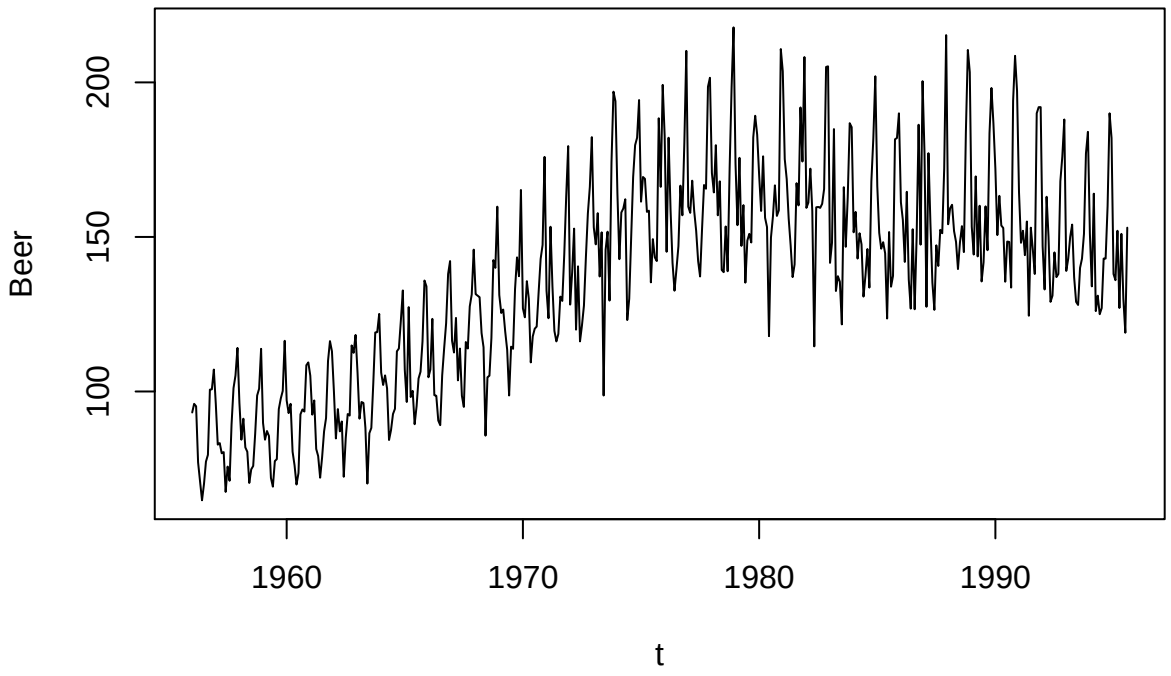
Logarithm of the time series

```
plot(log(AirPassengers),xlab="t",ylab="Airpass")
```



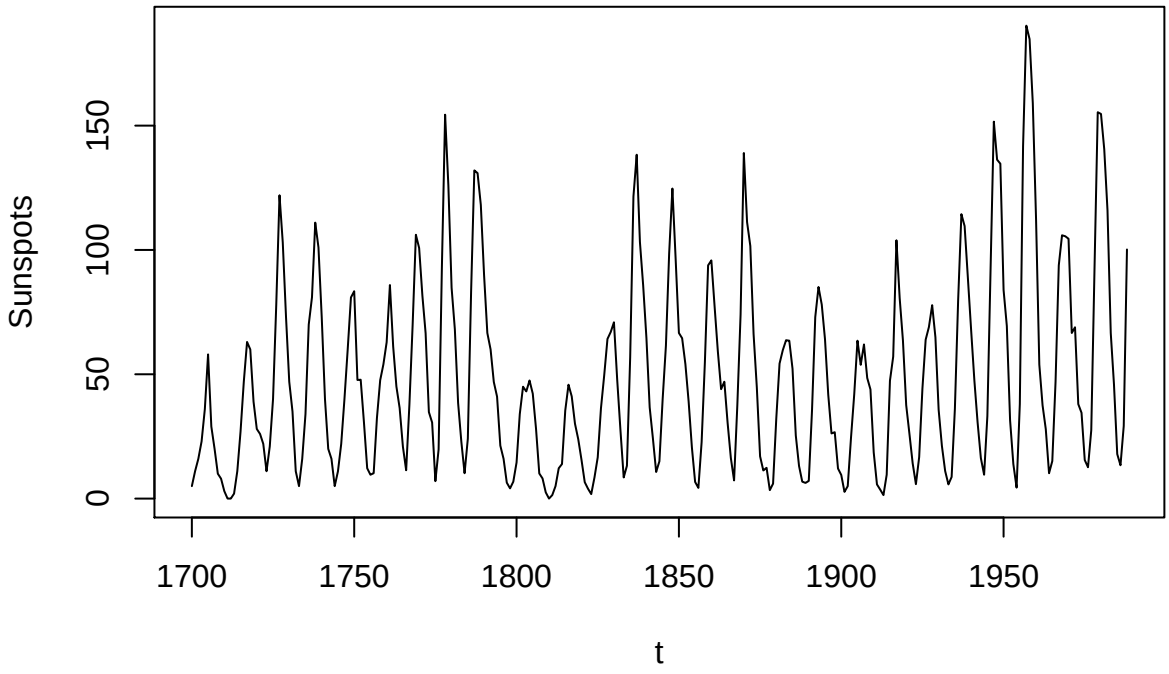
Series *beer*: monthly Australian beer production (in megaliters) from January 1956 to February 1991

```
beer=read.csv("../Data/beer.csv",header=F,dec=".",sep=",")
beer=ts(beer[,2],start=1956,freq=12)
plot(beer,xlab="t",ylab="Beer")
```



Series *sunspot*: yearly number of sunspots from 1790 to 1970

```
plot(sunspot.year,xlab="t",ylab="Sunspots")
```



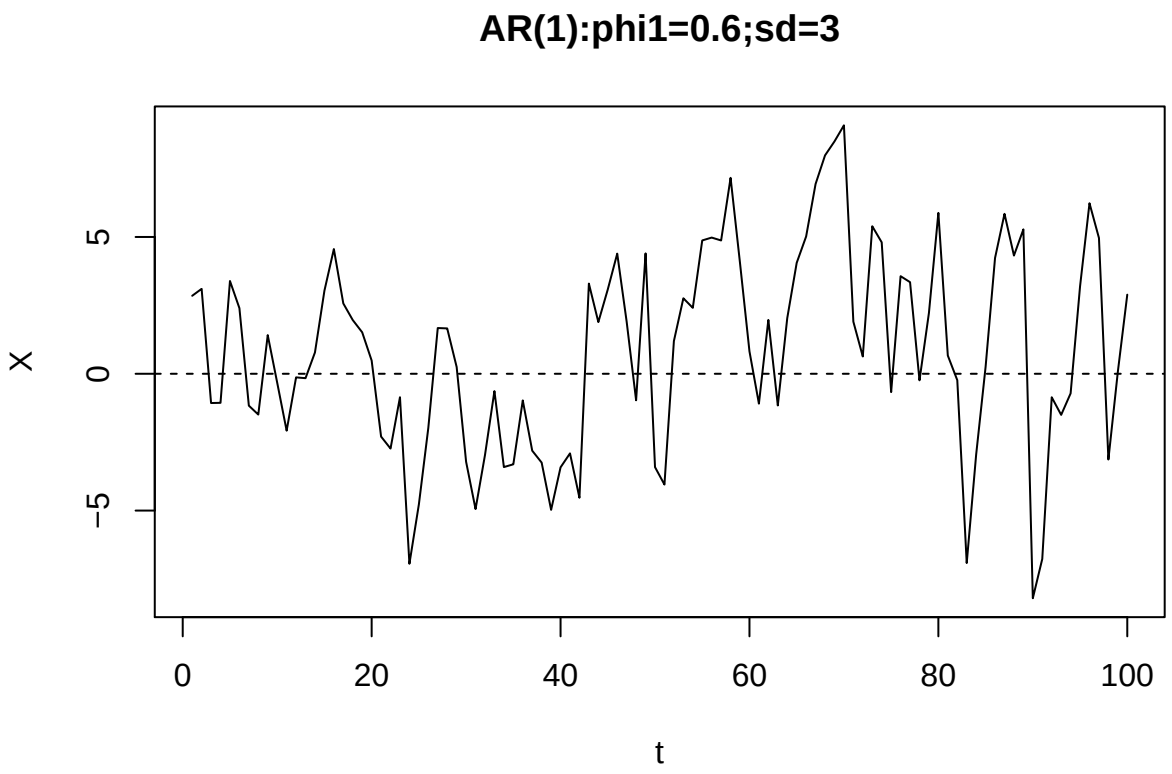
AR, MA & ARMA SIMULATIONS

We consider time series with $n = 100$ points.

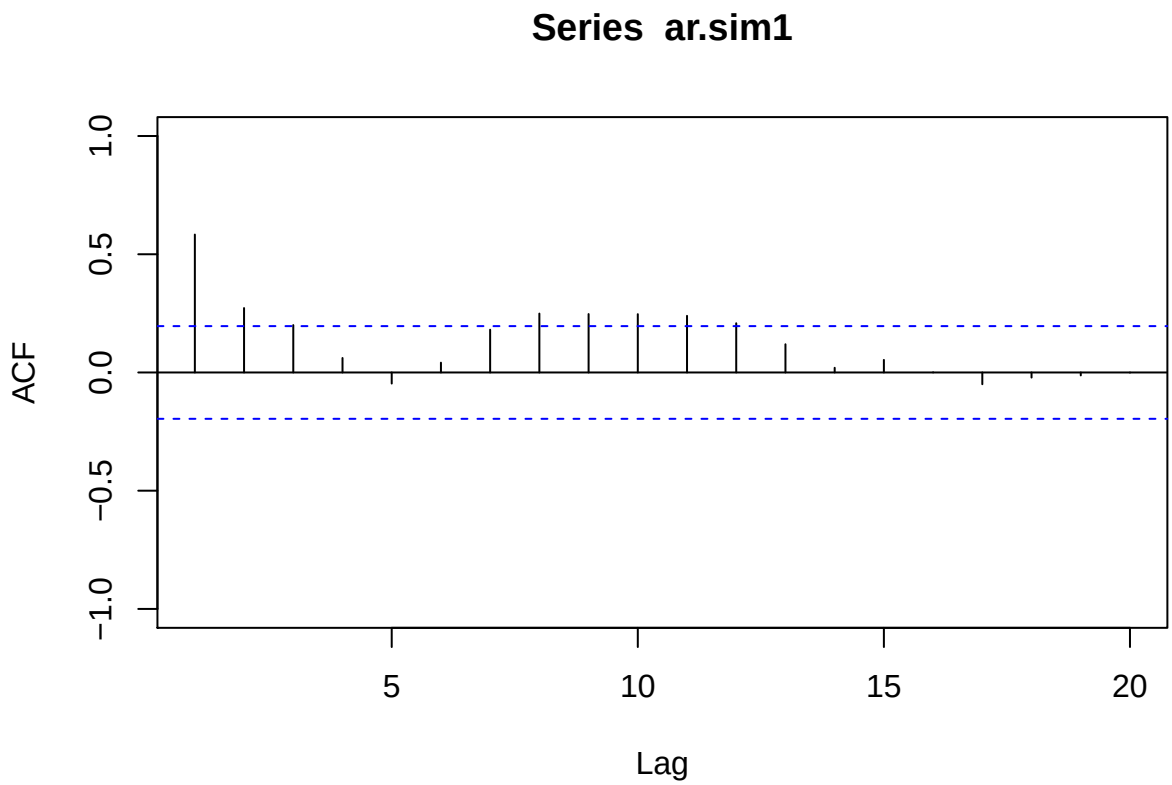
AR case

With an $AR(1)$ process such that $X_t = 0.6X_{t-1} + \varepsilon_t$ and $\text{Var}(X_t) = 3^2$:

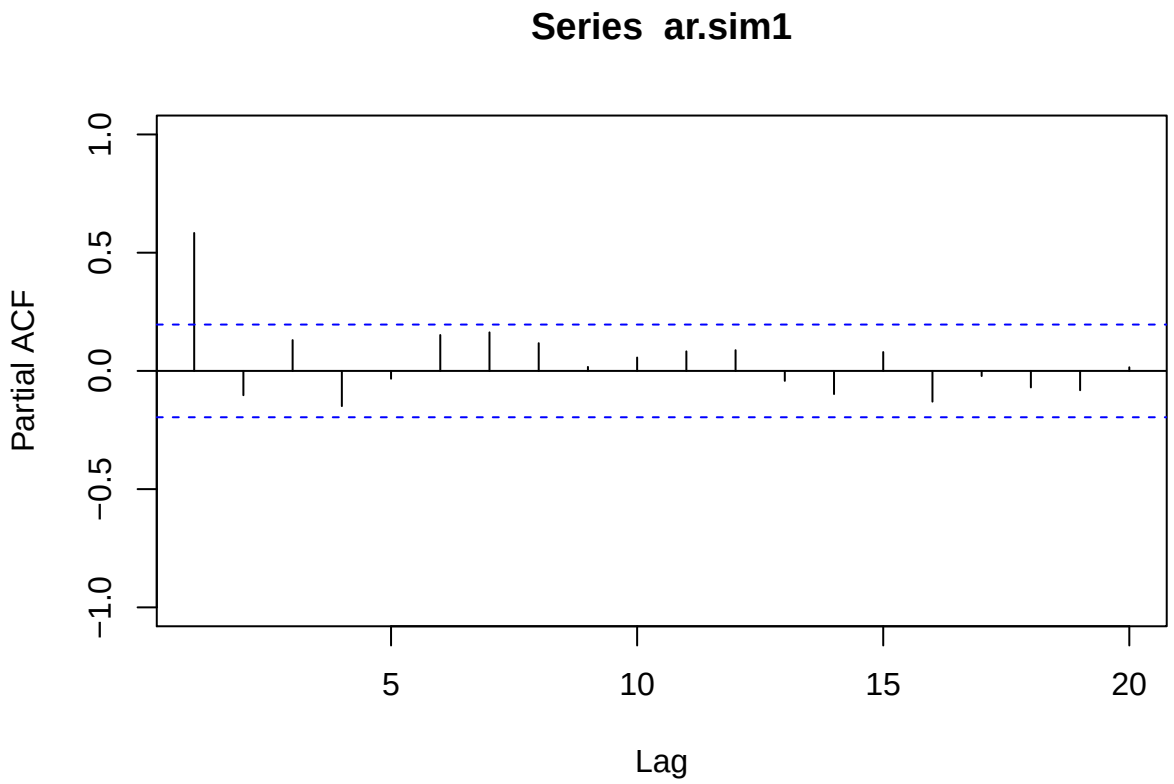
```
set.seed(1789)
ar.sim1=arima.sim(n=100,list(ar=0.6),sd=3)
plot(ar.sim1,xlab="t",ylab="X",main="AR(1):phi1=0.6;sd=3")
abline(h=0,lty=2)
```




```
plot(acf(ar.sim1,lag.max=20,plot=FALSE),ylim=c(-1,1))
```

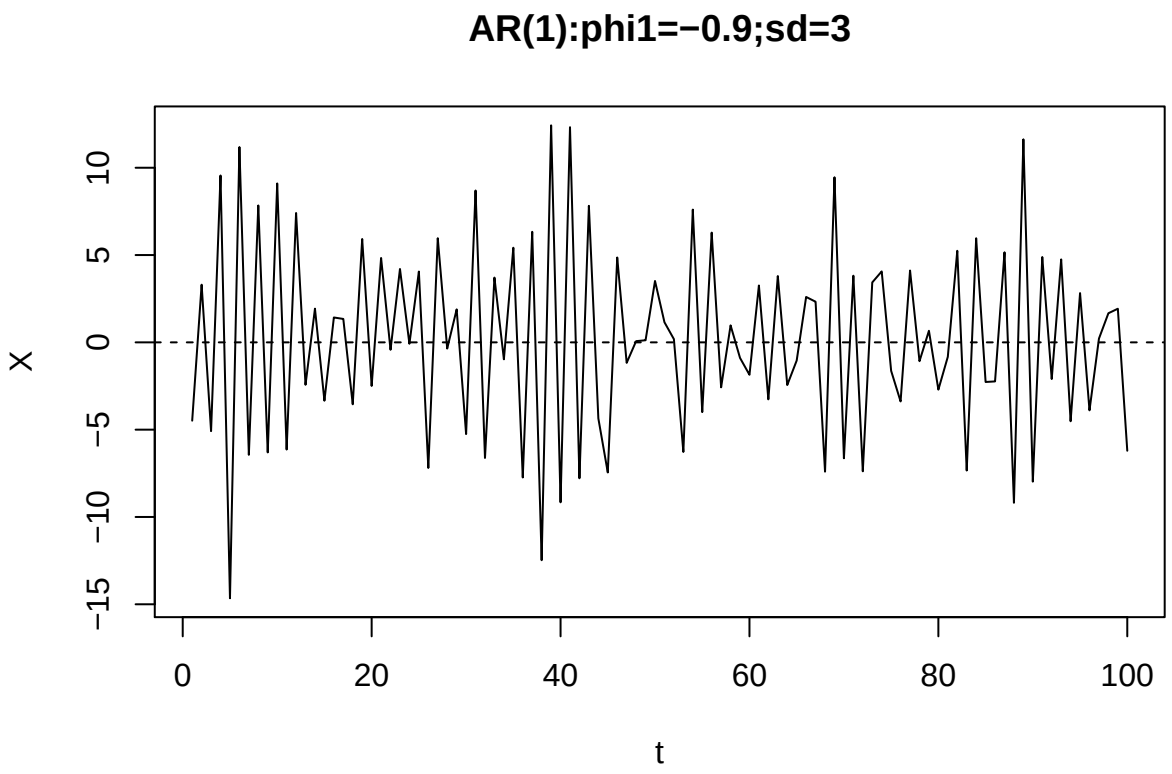


```r
plot(pacf(ar.sim1,lag.max=20,plot=FALSE),ylim=c(-1,1))
```

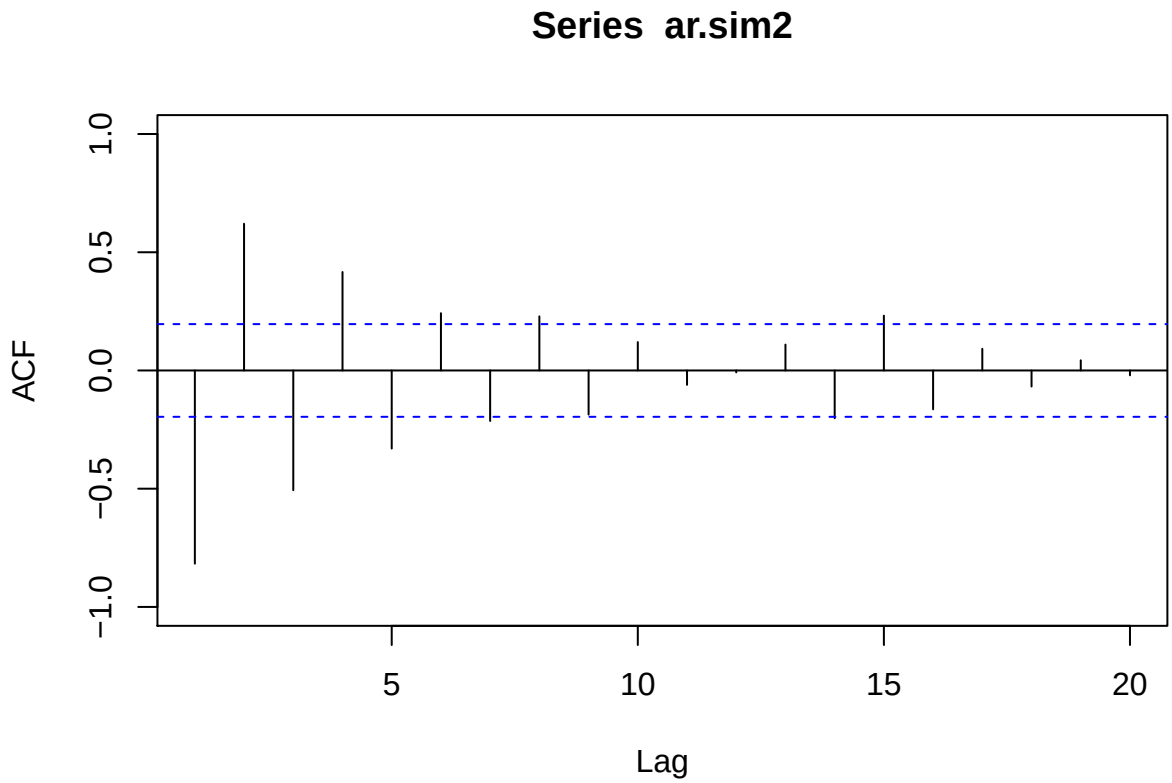


Consider the $AR(1)$ process such that $X_t = -0.8t - 1 + \varepsilon_t$ and $\text{Var}\,(X_t) = 3^2$:

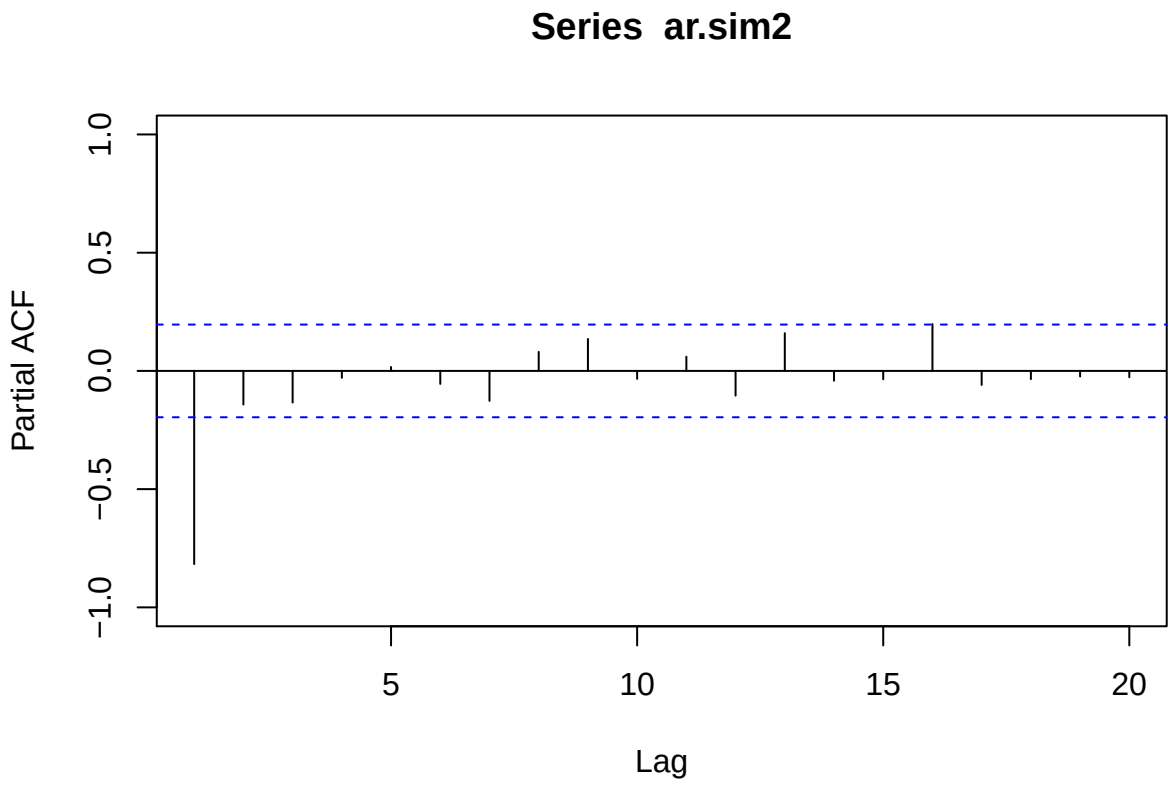
```r
set.seed(1789)
ar.sim2=arima.sim(n=100,list(ar=-0.9),sd=3)
plot(ar.sim2,xlab="t",ylab="X",main="AR(1):phi1=-0.9;sd=3")
abline(h=0,lty=2)
```




```
plot(acf(ar.sim2,lag.max=20,plot=FALSE),ylim=c(-1,1))
```



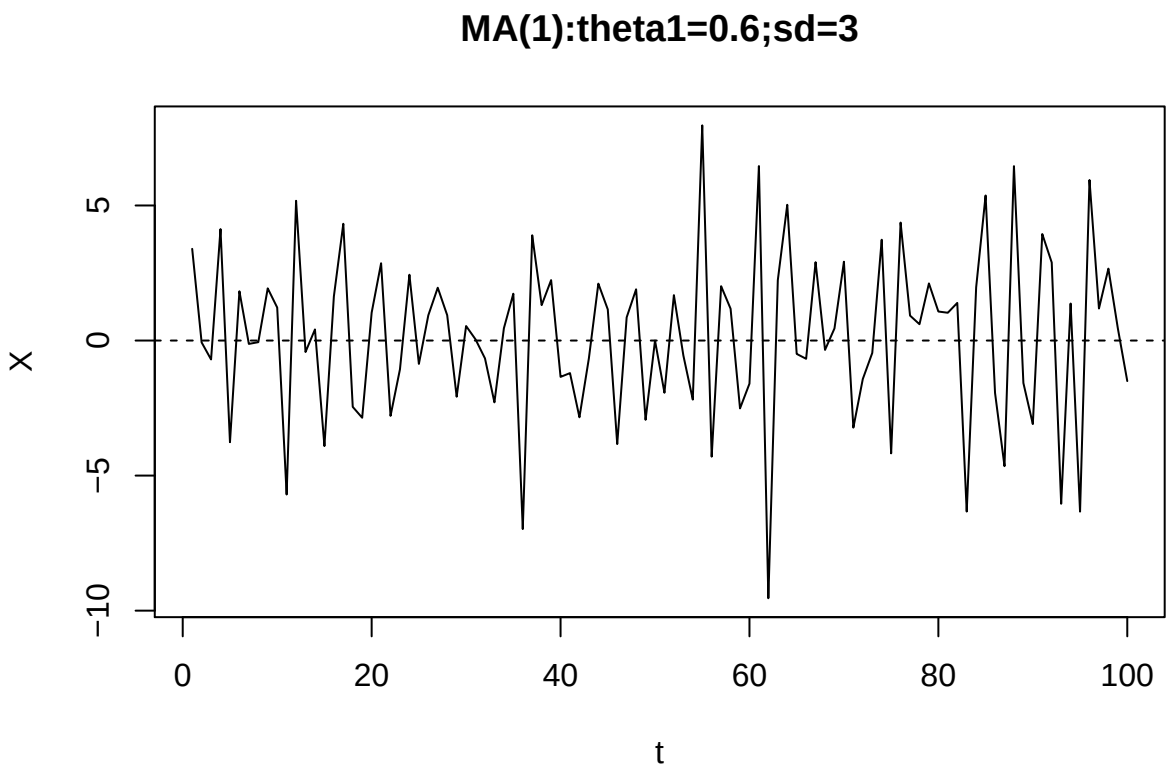
```r
plot(pacf(ar.sim2,lag.max=20,plot=FALSE),ylim=c(-1,1))
```



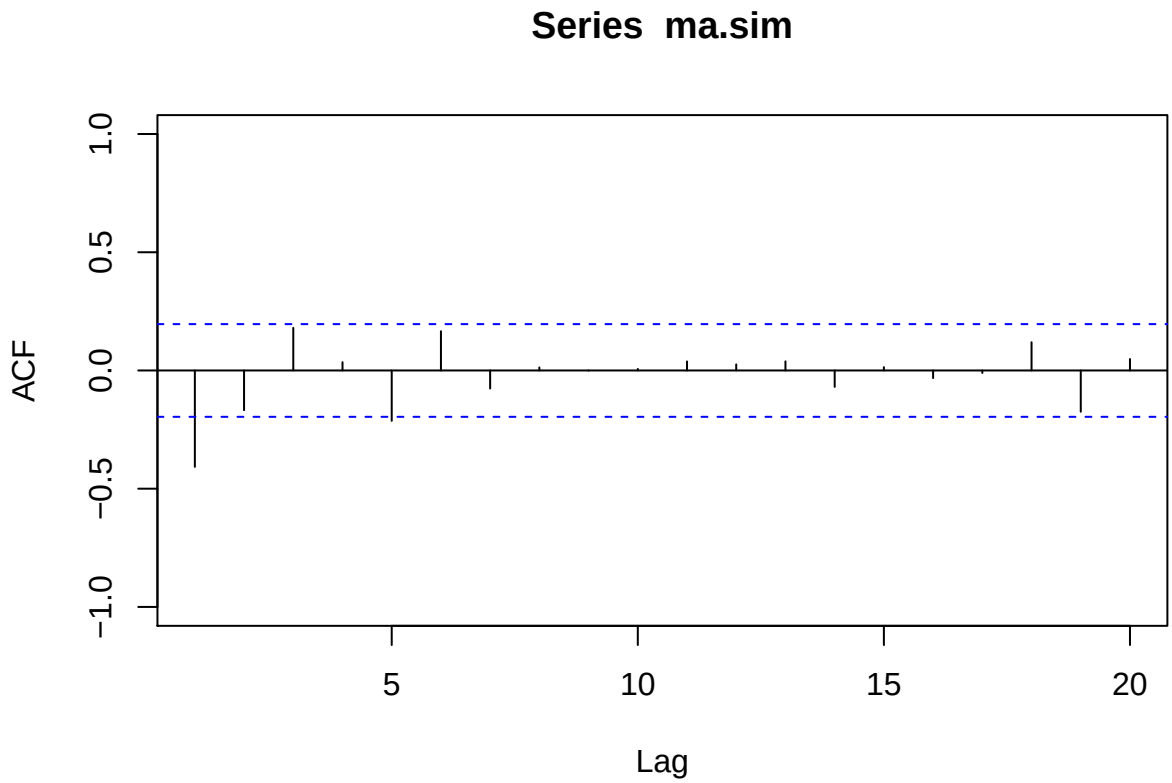
MA case

Consider the $MA(1)$ process such that $X_t = \varepsilon_t - 0.7\varepsilon_{t-1}$ and $\text{Var}\,(X_t) = 3^2$:

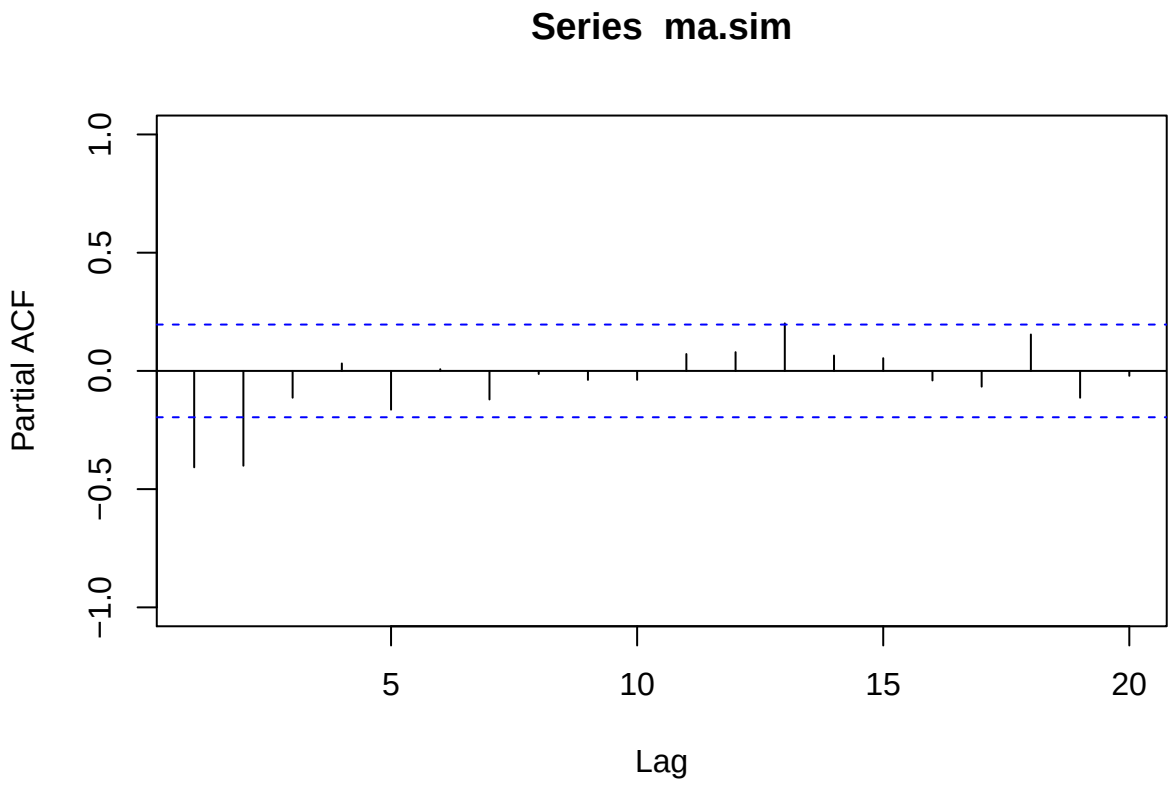
```r
set.seed(1789)
ma.sim=arima.sim(n=100,list(ma=-0.7),sd=3)
plot(ma.sim,xlab="t",ylab="X",main="MA(1):theta1=0.6;sd=3")
abline(h=0,lty=2)
```




```
plot(acf(ma.sim,lag.max=20,plot=FALSE),ylim=c(-1,1))
```



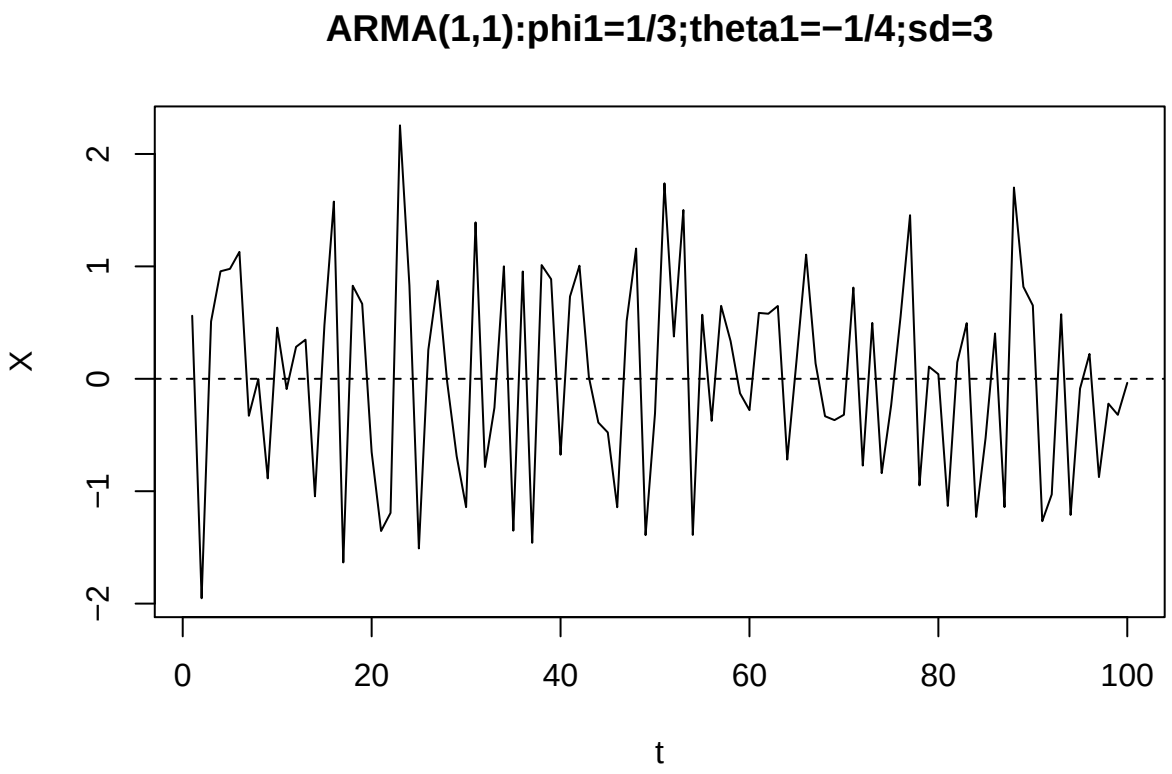
```r
plot(pacf(ma.sim,lag.max=20,plot=FALSE),ylim=c(-1,1))
```



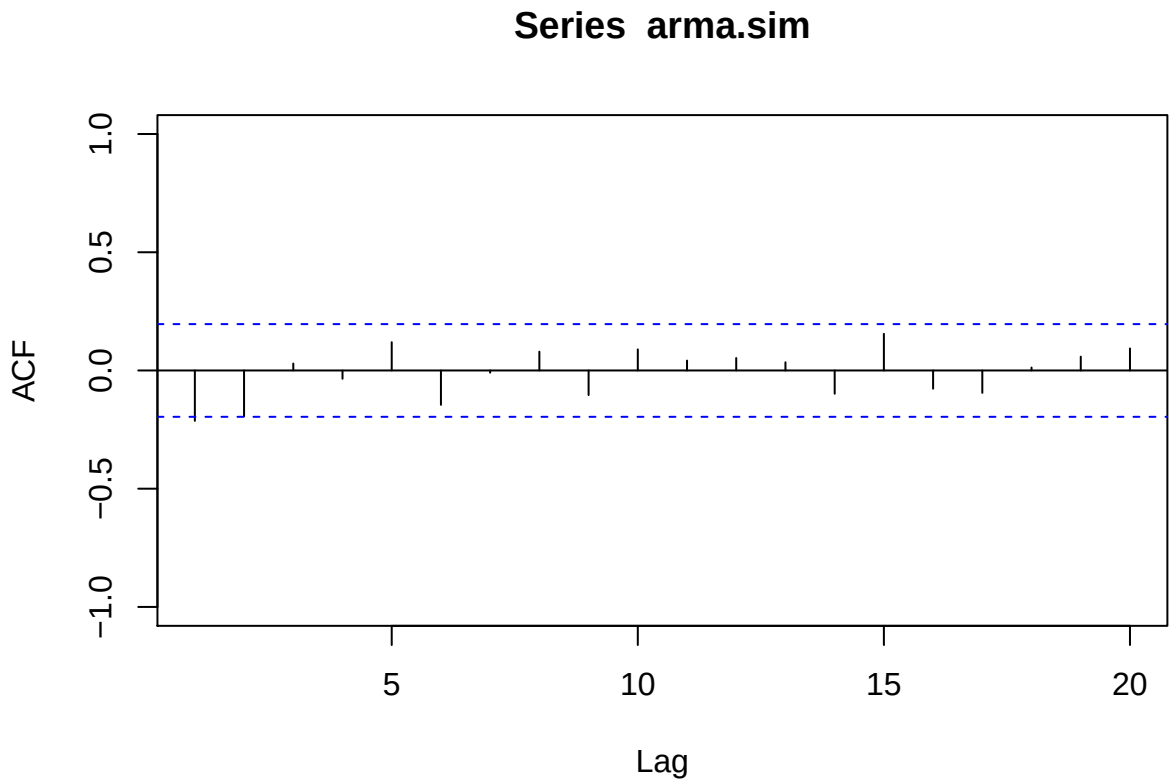
ARMA case

Consider the $ARMA(1,1)$ process such that $X_t = \frac{1}{3}X_{t-1} + \varepsilon_t - \frac{1}{4}\varepsilon_{t-1}$ and $\text{Var}\,(X_t) = 3^2$:

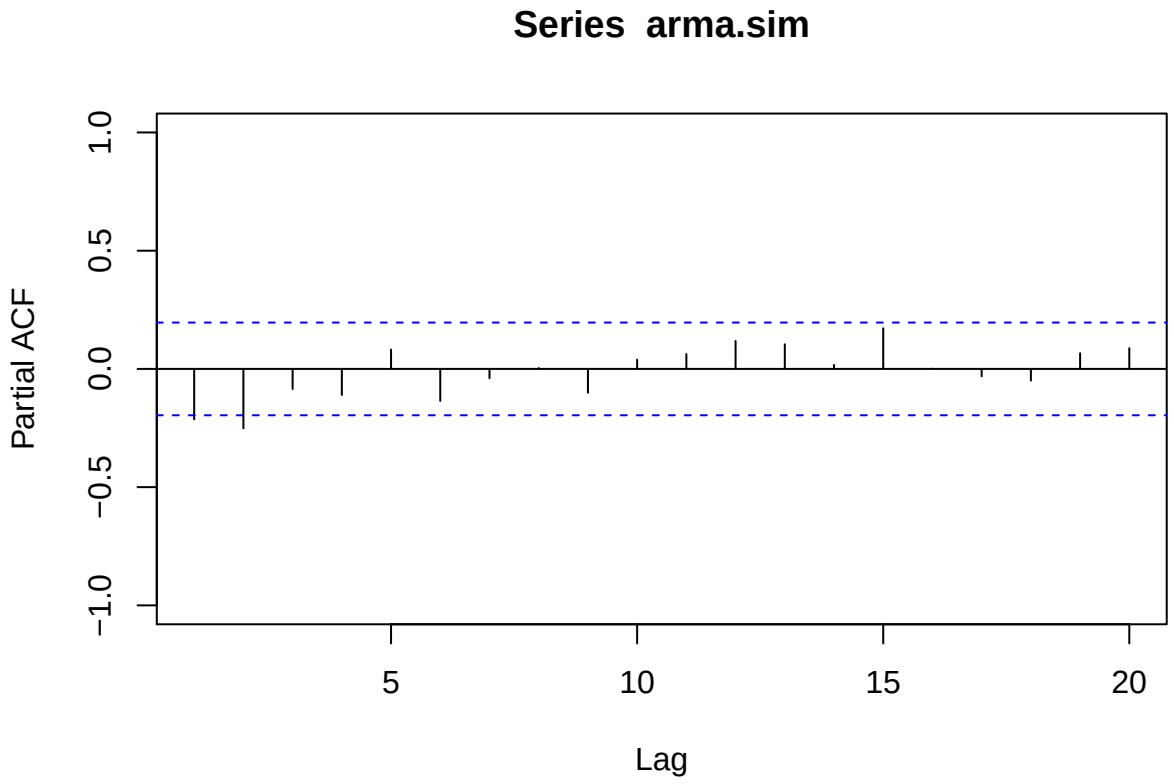
```r
arma.sim=arma.sim=arima.sim(n=100,list(ar=1/3,ma=-1/4),sd=1)
plot(arma.sim,xlab="t",ylab="X",main="ARMA(1,1):phi1=1/3;theta1=-1/4;sd=3")
abline(h=0,lty=2)
```




```
plot(acf(arma.sim,lag.max=20,plot=FALSE),ylim=c(-1,1))
```



```
plot(pacf(arma.sim,lag.max=20,plot=FALSE),ylim=c(-1,1))
```



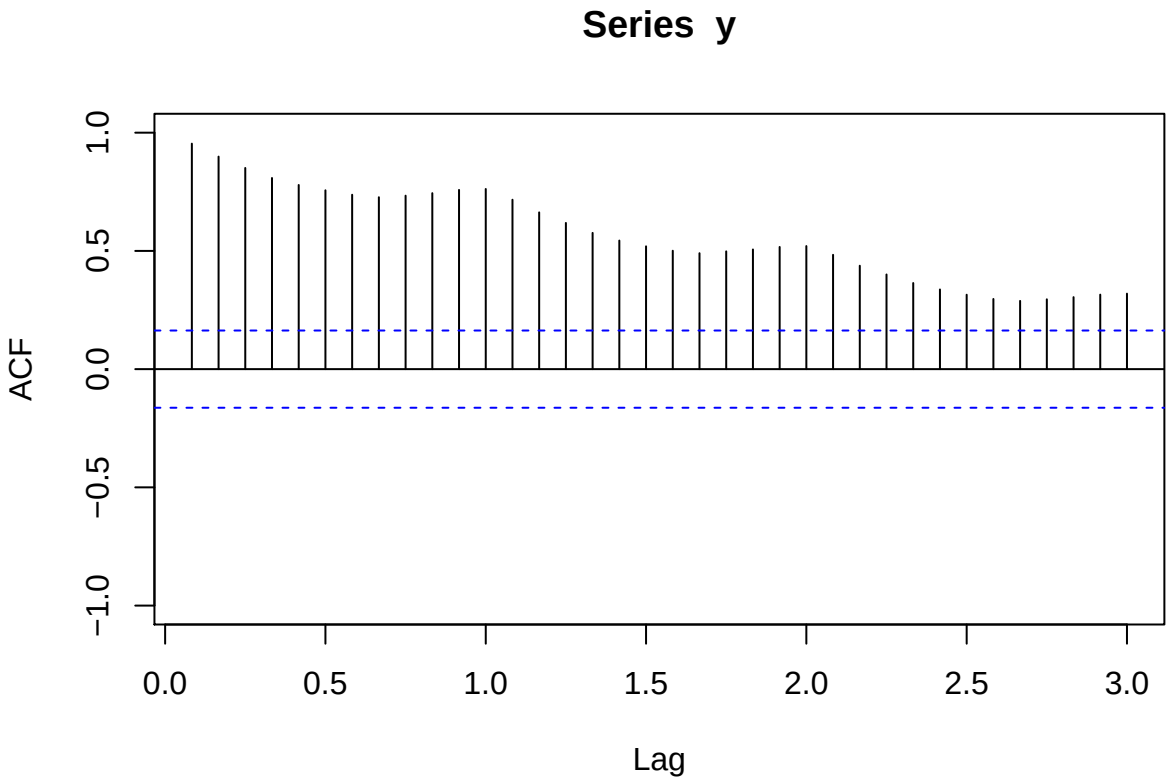
SARIMA MODELING EXAMPLE

We consider here the *airpass* series. X_t refers to the *airpass* series and we consider $Y_t = \log(X_t)$.

```
x=AirPassengers
y=log(x)
```

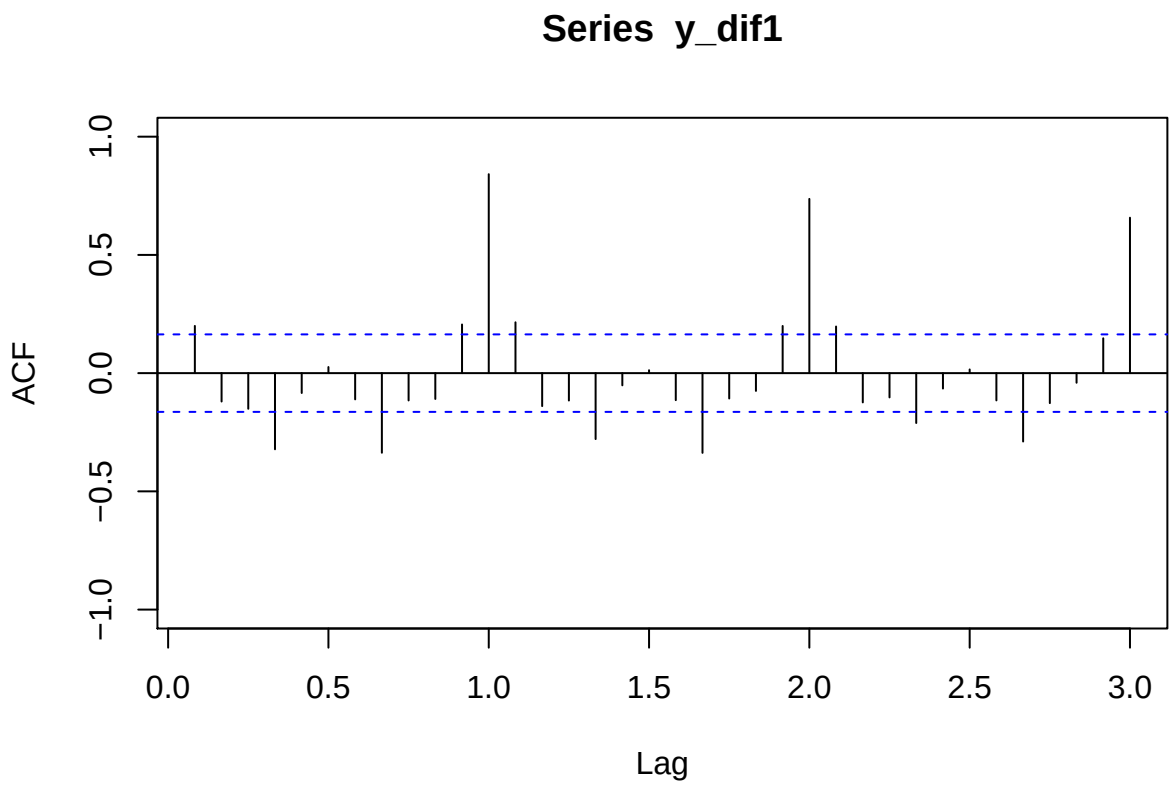
Stationarity

```
plot(acf(y,lag.max=36,plot=FALSE),ylim=c(-1,1))
```



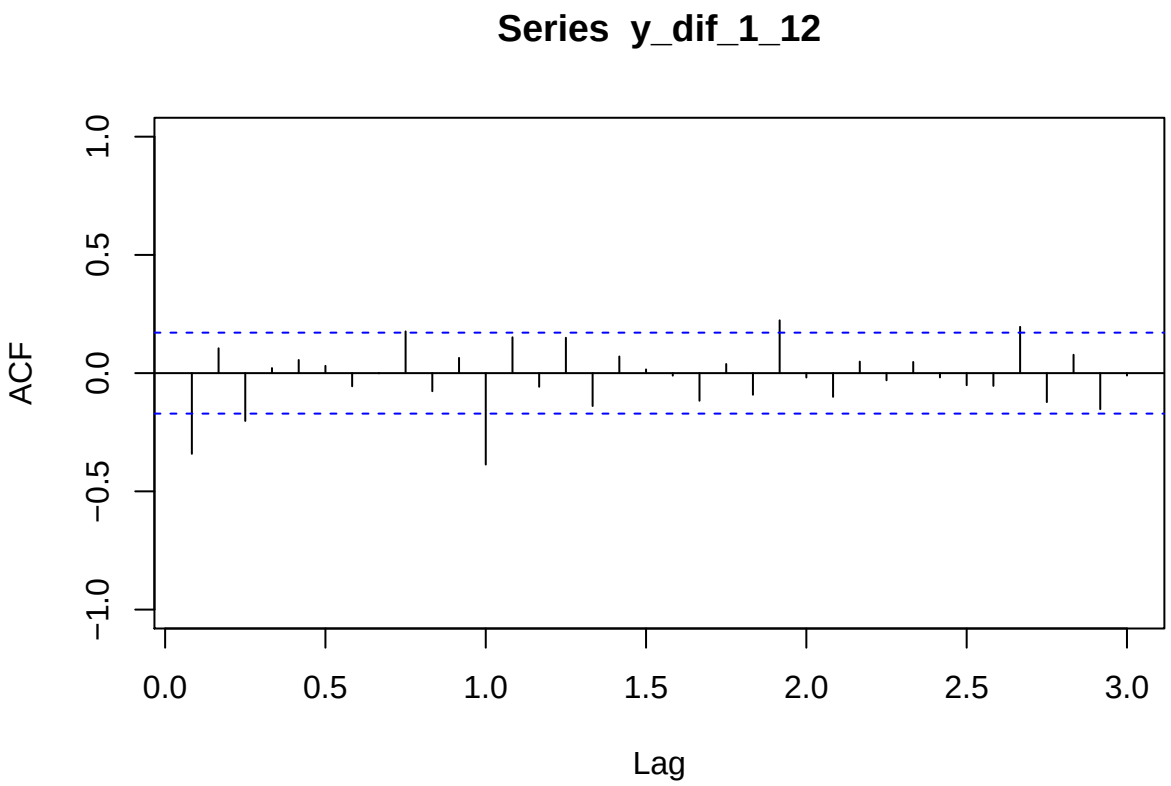
The *ACF* output decreases too slowly to be an autocorrelation function estimation. We may use a differencing transformation at lag 1: $(I - B)$.

```
y_dif1=diff(y,lag=1,differences=1)
plot(acf(y_dif1,lag.max=36,plot=FALSE),ylim=c(-1,1))
```



The ACF output decreases too slowly every 12 multiple lags to be an autocorrelation function estimation. We may use a differencing transformation at lag 12: $\left(I - B^{12}\right)$.

```
y_dif_1_12=diff(y_dif1,lag=12,differences=1)
plot(acf(y_dif_1_12,lag.max=36,plot=FALSE),ylim=c(-1,1))
```



The *ACF* output seems to decrease fast enough to be considered as the estimation of the autocorrelation function.

We will thus select model order based on ACF and PACF outputs of the time series:

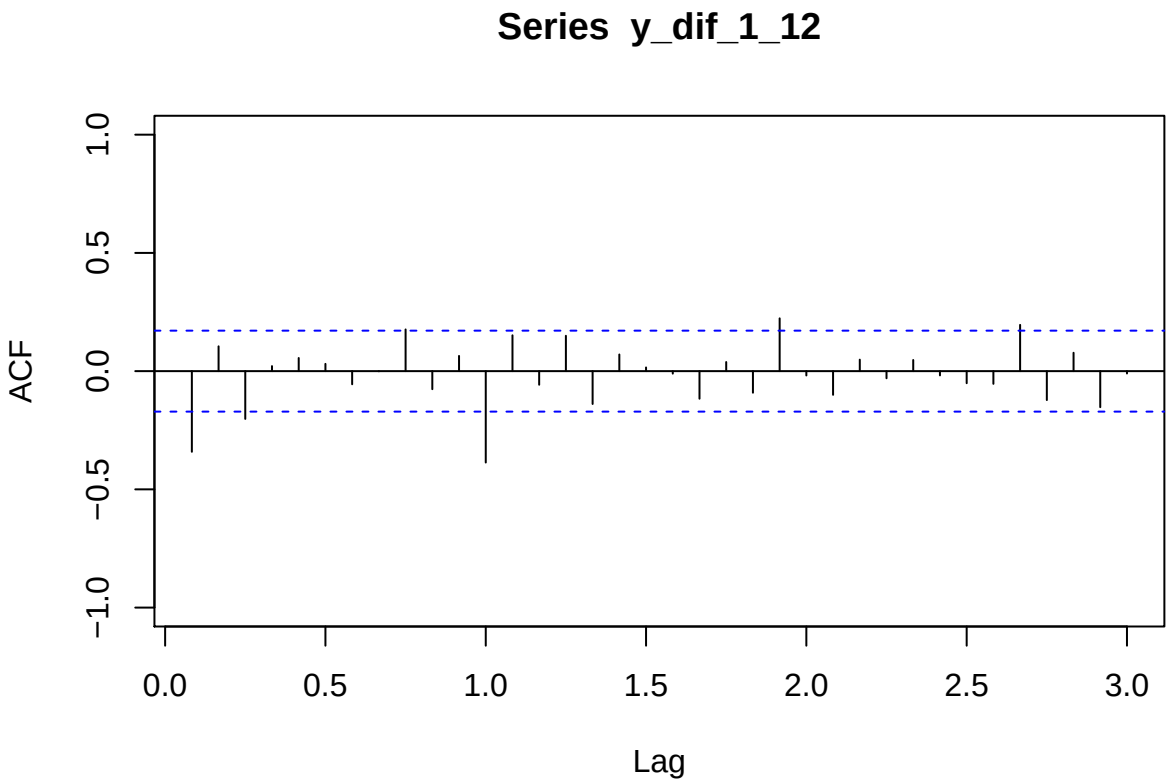
$$(I - B)\left(I - B^{12}\right)\log\left(X_t\right).$$

Models estimations and diagnostic checking

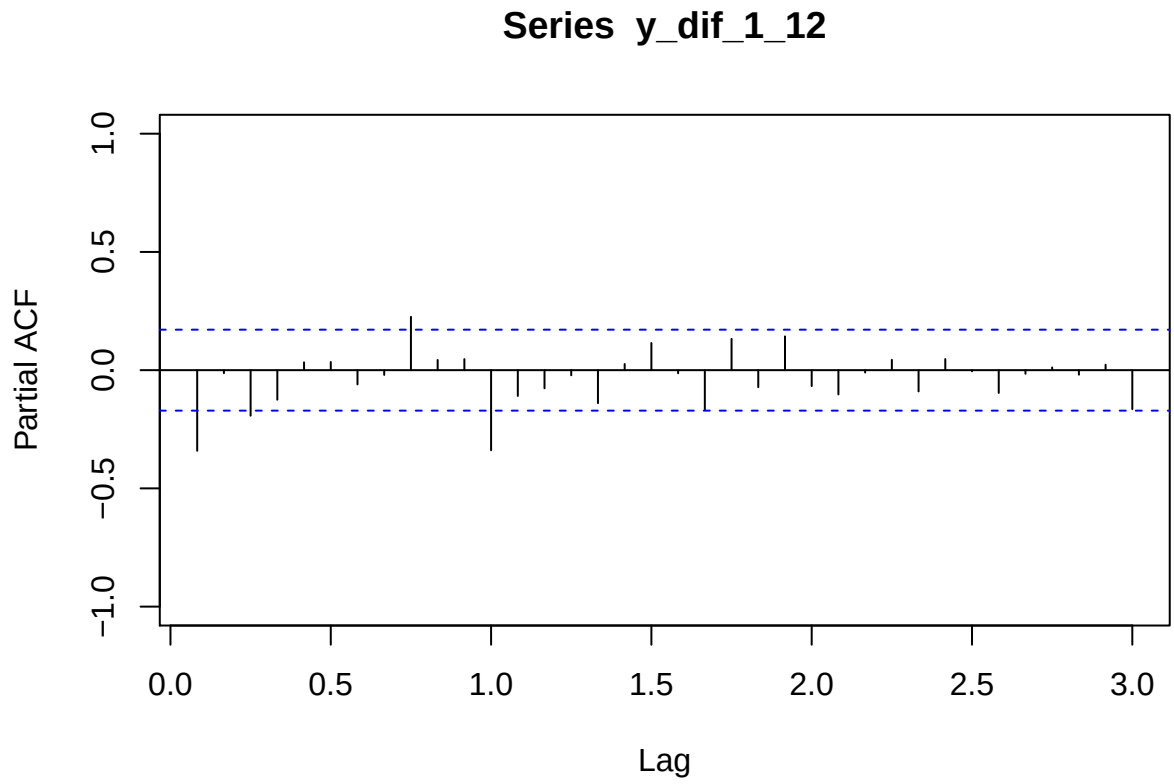
From now test level is 5%.

The simple and partial autocorrelation estimations are:

```
plot(acf(y_dif_1_12,lag.max=36,plot=FALSE),ylim=c(-1,1))
```




```
plot(pacf(y_dif_1_12,lag.max=36,plot=FALSE),ylim=c(-1,1))
```



Model 1

We estimate first a $SARIMA(1,1,1)(1,1,1)_{12}$ model:

$$\left(I-\varphi_1 B\right)\left(I-\varphi_1' B^{12}\right)\left(I-B\right)\left(I-B^{12}\right)\log\left(X_t\right)=\left(I+\theta_1 B\right)\left(I+\theta_1' B^{12}\right)\varepsilon_t.$$

```
model1=Arima(y,order=c(1,1,1),list(order=c(1,1,1),period=12),include.mean=FALSE,method="CSS-ML")
summary(model1)
```

```
## Series: y
## ARIMA(1,1,1)(1,1,1)[12]
##
## Coefficients:
##          ar1      ma1    sar1     sma1
##       0.1666  -0.5615  -0.099  -0.4973
## s.e.  0.2459   0.2115   0.154   0.1360
##
## sigma^2 estimated as 0.00138:  log likelihood=245.16
## AIC=-480.31   AICc=-479.83   BIC=-465.93
##
## Training set error measures:
##                        ME       RMSE        MAE        MPE      MAPE
## Training set 0.0006239395 0.03489259 0.02595463 0.01199887 0.4696646
##                    MASE        ACF1
## Training set 0.2144266 -0.01250397
```

```
t_stat(model1)
```

```
##             ar1       ma1      sar1      sma1
## t.stat 0.677738 -2.654214 -0.642984 -3.657670
## p.val  0.497938  0.007949  0.520235  0.000255
```

```
Box.test.2(model1$residuals,nlag=c(6,12,18,24,30,36),type="Ljung-Box",decim=5)
```

```
##      Retard p-value
## [1,]      6 0.64051
## [2,]     12 0.81959
## [3,]     18 0.82768
## [4,]     24 0.59646
## [5,]     30 0.75443
## [6,]     36 0.65902
```

The two autoregressive parameters aren't significant, thus we remove the less significant one.

Model 2

We estimate now a $SARIMA(1,1,1)(0,1,1)_{12}$ model:

$$\left(I-\varphi_1' B^{12}\right)\left(I-B\right)\left(I-B^{12}\right)\log\left(X_t\right)=\left(I+\theta_1 B\right)\left(I+\theta_1' B^{12}\right)\varepsilon_t.$$

```r
model2=Arima(y,order=c(1,1,1),list(order=c(0,1,1),period=12),include.mean=FALSE,method="CSS-ML")
summary(model2)
```

```
## Series: y
## ARIMA(1,1,1)(0,1,1)[12]
##
## Coefficients:
##          ar1      ma1     sma1
##       0.1960  -0.5784  -0.5643
## s.e.  0.2475   0.2132   0.0747
##
## sigma^2 estimated as 0.001375:  log likelihood=244.95
## AIC=-481.9   AICc=-481.58   BIC=-470.4
##
## Training set error measures:
##                        ME       RMSE        MAE        MPE      MAPE
## Training set 0.0006214569 0.03495868 0.02587023 0.01205357 0.4682205
##                    MASE        ACF1
## Training set 0.2137293 -0.01530519
```

```r
t_stat(model2)
```

```
##             ar1       ma1      sma1
## t.stat 0.792074 -2.712668 -7.554412
## p.val  0.428318  0.006674  0.000000
```

```r
Box.test.2(model2$residuals,nlag=c(6,12,18,24,30,36),type="Ljung-Box",decim=5)
```

```
##      Retard p-value
## [1,]      6 0.65243
## [2,]     12 0.80854
## [3,]     18 0.82136
## [4,]     24 0.57705
## [5,]     30 0.74768
## [6,]     36 0.67642
```

The autoregressive parameter is still not significant, thus we remove it.

Model 3

We estimate now a $SARIMA(0,1,1)(0,1,1)_{12}$ model:

$$(I-B)\left(I-B^{12}\right)\log\left(X_t\right)=\left(I+\theta_1 B\right)\left(I+\theta_1' B^{12}\right)\varepsilon_t.$$

```r
model3=Arima(y,order=c(0,1,1),list(order=c(0,1,1),period=12),include.mean=FALSE,method="CSS-ML")
summary(model3)
```

```
## Series: y
## ARIMA(0,1,1)(0,1,1)[12]
##
```

```
## Coefficients:
##           ma1     sma1
##       -0.4018  -0.5569
## s.e.   0.0896   0.0731
##
## sigma^2 estimated as 0.001371:  log likelihood=244.7
## AIC=-483.4   AICc=-483.21   BIC=-474.77
##
## Training set error measures:
##                        ME       RMSE        MAE        MPE      MAPE
## Training set 0.0005730622 0.03504883 0.02626034 0.01098898 0.4752815
##                    MASE       ACF1
## Training set 0.2169522 0.01443892
```

```
t_stat(model3)
```

```
##              ma1      sma1
## t.stat -4.482494 -7.618978
## p.val   0.000007  0.000000
```

```
Box.test.2(model3$residuals,nlag=c(6,12,18,24,30,36),type="Ljung-Box",decim=5)
```

```
##      Retard p-value
## [1,]      6 0.51519
## [2,]     12 0.72613
## [3,]     18 0.77822
## [4,]     24 0.50077
## [5,]     30 0.68838
## [6,]     36 0.65352
```

Significance of the parameters and whiteness of the residuals seem correct.

```
shapiro.test(model3$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model3$residuals
## W = 0.98637, p-value = 0.1674
```

Normality hypothesis is also not rejected.

Model 6

One can check that the following model a also correct:

$$(I - \varphi_1 B)(I - B)\left(I - B^{12}\right)\log(X_t) = \left(I + \theta_{12}B^{12}\right)\varepsilon_t.$$

```
model6=Arima(y,order=c(1,1,12),fixed=c(NA,0,0,0,0,0,0,0,0,0,0,0,NA),list(order=c(0,1,0),period=12),inclu
summary(model6)
```

```
## Series: y
## ARIMA(1,1,12)(0,1,0)[12]
##
## Coefficients:
##           ar1  ma1  ma2  ma3  ma4  ma5  ma6  ma7  ma8  ma9  ma10  ma11
##       -0.3395    0    0    0    0    0    0    0    0    0     0     0
## s.e.   0.0822    0    0    0    0    0    0    0    0    0     0     0
##          ma12
##       -0.5619
## s.e.   0.0748
##
## sigma^2 estimated as 0.001521:  log likelihood=243.74
## AIC=-481.49   AICc=-481.3   BIC=-472.86
##
## Training set error measures:
##                         ME       RMSE        MAE         MPE      MAPE
## Training set 0.0004500154 0.03529899 0.02662601 0.008828412 0.4816646
##                    MASE        ACF1
## Training set 0.2199733 -0.02338686
```

```
t_stat(model6)
```

```
##              ar1      ma12
## t.stat -4.129480 -7.510894
## p.val   0.000036  0.000000
```

```
Box.test.2(model6$residuals,nlag=c(6,12,18,24,30,36),type="Ljung-Box",decim=5)
```

```
##      Retard p-value
## [1,]      6 0.24413
## [2,]     12 0.43316
## [3,]     18 0.45925
## [4,]     24 0.23920
## [5,]     30 0.39768
## [6,]     36 0.38129
```

```
shapiro.test(model6$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model6$residuals
## W = 0.98611, p-value = 0.1569
```

Automatic order selection

One can see that the automatic order selection isn't really efficient here...

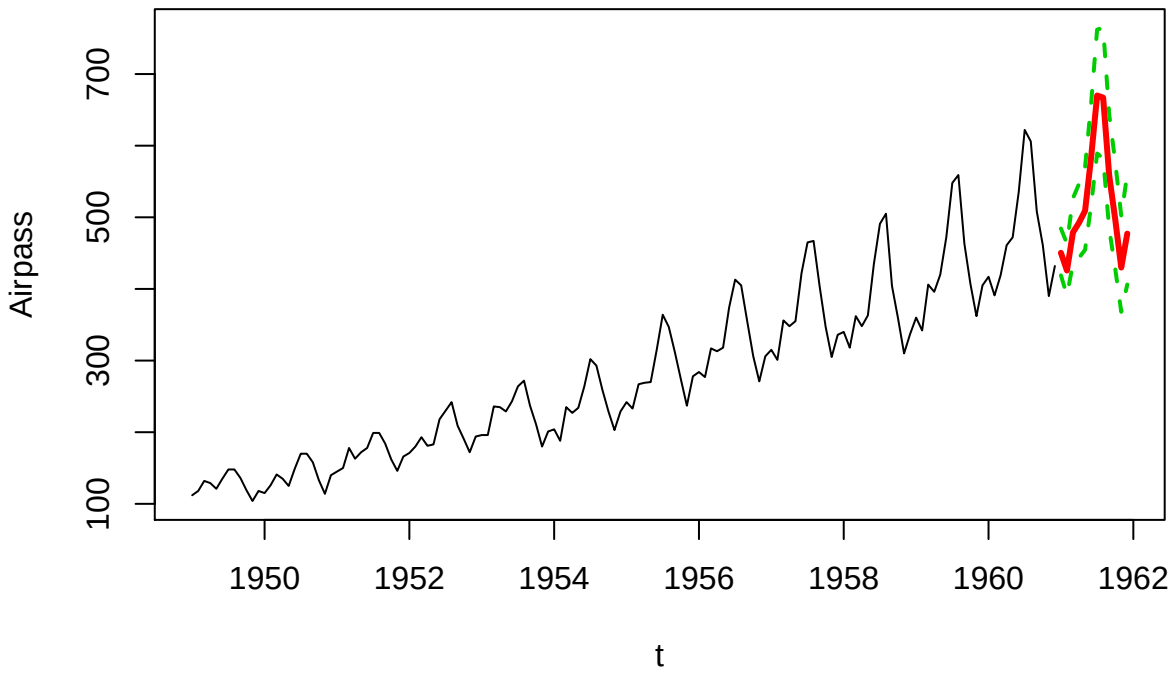
```
armaselect(y_dif_1_12,max.p=20,max.q=20,nbmod=10)
```

```
##       p  q       sbc
##  [1,] 1  1 -829.6517
##  [2,] 0  1 -827.8830
##  [3,] 1  2 -826.4204
##  [4,] 2  1 -824.7766
##  [5,] 0  2 -824.5979
##  [6,] 3  1 -824.1854
##  [7,] 0 12 -823.8556
##  [8,] 1 20 -822.1977
##  [9,] 2  2 -822.1365
## [10,] 1  3 -822.1087
```

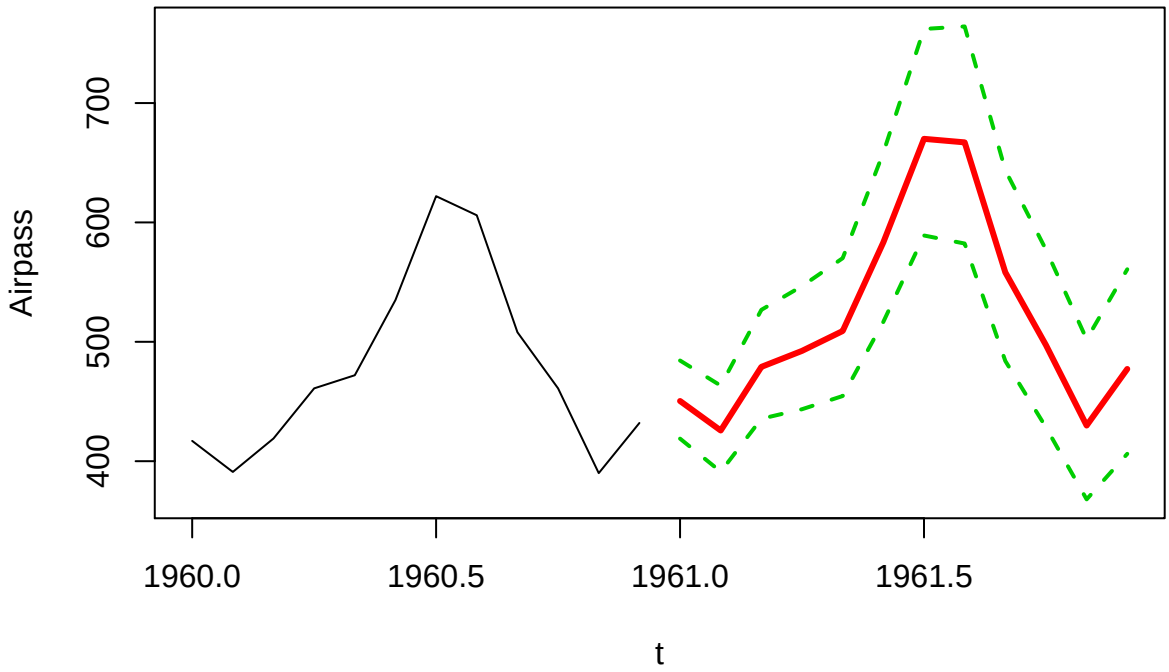
Final order selection and forecasting

The Schwarz criterium of the model 3 is -474.77 against -472.86 for the model 4, so we choose here the model 3, and then forecast the year 1961.

```
forecast_model3=forecast(model3,h=12,level=95)
pred=exp(forecast_model3$mean)
forecast_l=ts(exp(forecast_model3$lower),start=c(1961,1),frequency=12)
forecast_u=ts(exp(forecast_model3$upper),start=c(1961,1),frequency=12)
ts.plot(x,pred,forecast_l,forecast_u,xlab="t",ylab="Airpass",col=c(1,2,3,3),lty=c(1,1,2,2),lwd=c(1,3,2,
```



```
ts.plot(window(x,start=c(1960,1)),pred,forecast_l,forecast_u,xlab="t",ylab="Airpass",col=c(1,2,3,3),lty=
```



Ex post analysis

We truncate the series by removing the year 1960 and we forecast this year with the model 3 based on 1949-1959:

```
x_trunc=window(x,end=c(1959,12))
y_trunc=log(x_trunc)
x_to_forecast=window(x,start=c(1960,1))
```

We check that the model 3 is acceptable on the truncated time series:

```
model3trunc=Arima(y_trunc,order=c(0,1,1),list(order=c(0,1,1),period=12),include.mean=FALSE,method="CSS-
summary(model3trunc)
```

```
## Series: y_trunc
## ARIMA(0,1,1)(0,1,1)[12]
##
## Coefficients:
##           ma1     sma1
##       -0.3484  -0.5623
## s.e.   0.0943   0.0774
##
## sigma^2 estimated as 0.001338:  log likelihood=223.63
## AIC=-441.26   AICc=-441.05   BIC=-432.92
##
## Training set error measures:
##                      ME       RMSE        MAE        MPE      MAPE
```

```
## Training set 0.00104934 0.03443221 0.02590904 0.01899277 0.4738142
##                    MASE        ACF1
## Training set 0.2113963 0.004637637
```

```
t_stat(model3trunc)
```

```
##              ma1      sma1
## t.stat -3.695894 -7.262873
## p.val   0.000219  0.000000
```

```
Box.test.2(model3trunc$residuals,nlag=c(6,12,18,24,30,36),type="Ljung-Box",decim=5)
```

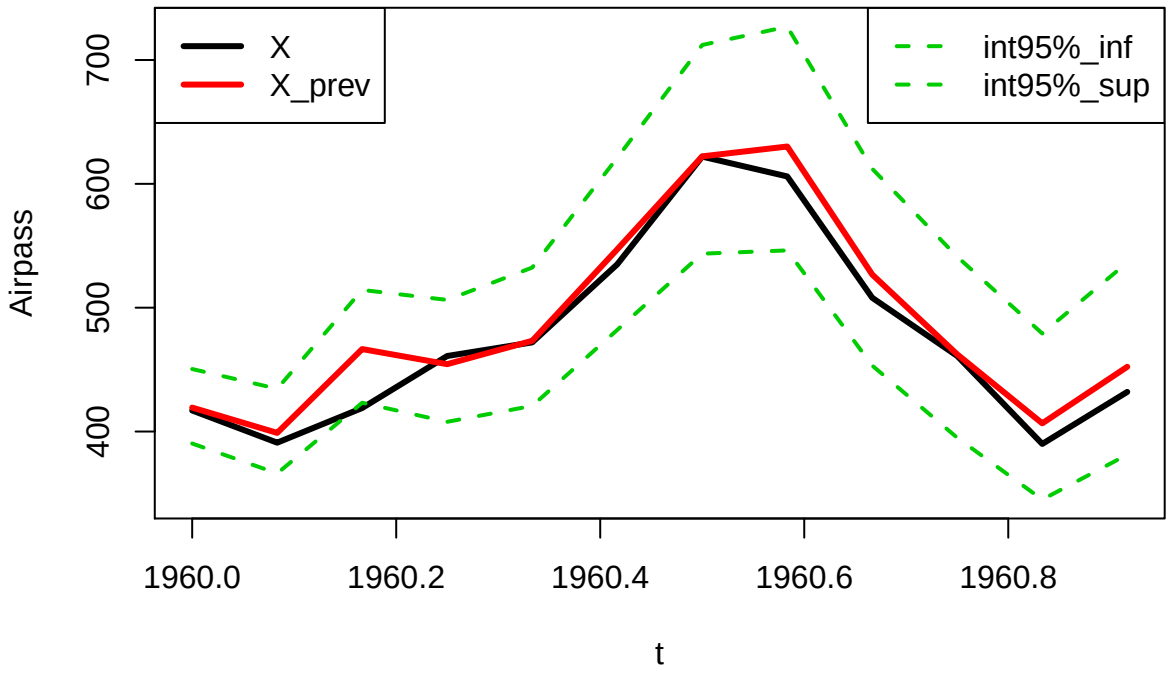
```
##      Retard p-value
## [1,]      6 0.52539
## [2,]     12 0.85631
## [3,]     18 0.87341
## [4,]     24 0.78327
## [5,]     30 0.90181
## [6,]     36 0.84635
```

```
shapiro.test(model3trunc$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model3trunc$residuals
## W = 0.988, p-value = 0.3065
```

The 1960 data are in the forecast interval of level 95% (based on the truncated time series):

```
forecast_model3trunc=forecast(model3trunc,h=12,level=95)
forecast_trunc=exp(forecast_model3trunc$mean)
forecast_l_trunc=ts(exp(forecast_model3trunc$lower),start=c(1960,1),frequency=12)
forecast_u_trunc=ts(exp(forecast_model3trunc$upper),start=c(1960,1),frequency=12)
ts.plot(x_to_forecast,forecast_trunc,forecast_l_trunc,forecast_u_trunc,xlab="t",ylab="Airpass",col=c(1,1
legend("topleft",legend=c("X","X_prev"),col=c(1,2,3,3),lty=c(1,1),lwd=c(3,3))
legend("topright",legend=c("int95%_inf","int95%_sup"),col=c(3,3),lty=c(2,2),lwd=c(2,2))
```



We compute RMSE and MAPE:

```
rmse=sqrt(mean((x_to_forecast-forecast_trunc)^2))
rmse
```

```
## [1] 18.59359
```

```
mape=mean(abs(1-forecast_trunc/x_to_forecast))*100
mape
```

```
## [1] 2.904473
```