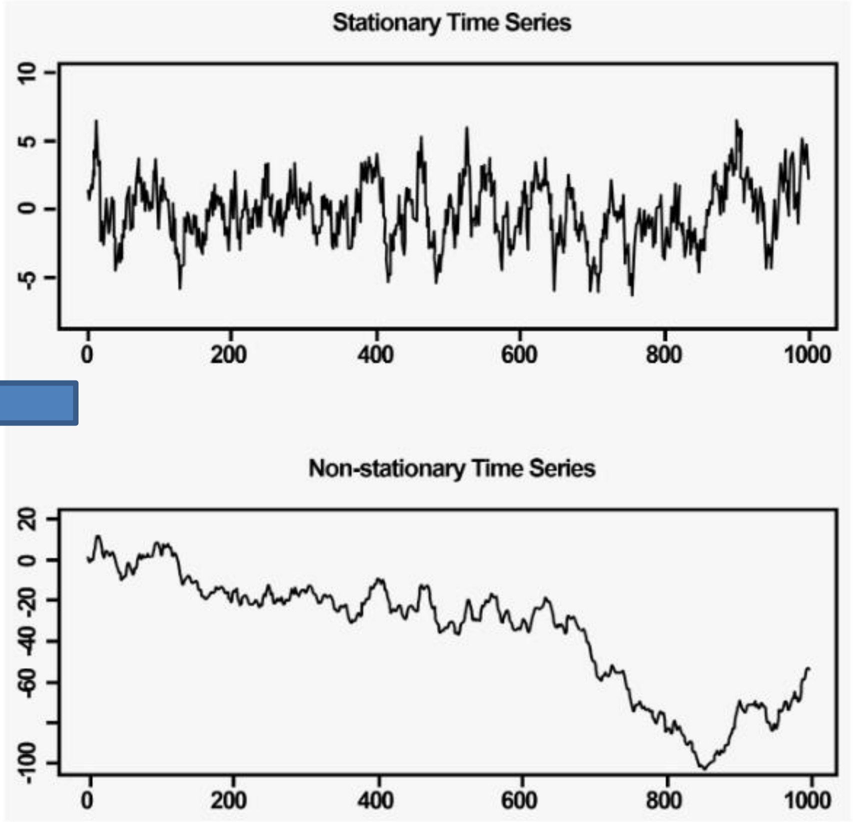
CIVL 7012/8012

Time series modeling

Stationarity

- A stationary time series is the one with statistical properties (mean, variance, etc.) constant
- Assumption of stationarity makes future predictions easier
- Time series need to stationary for modeling and making future predictions



In the first figure the statistical properties are more or less the same for the series which is not true for the second figure

Time series can be weakly stationary or strongly stationary

Image source: https://www.oreilly.com/library/view/hands-on-machine-learning/9781788992282/15c9cc40-bea2-4b75-902f-2e9739fec4ae.xhtml

Strong and Weak stationarity

- **Strong stationarity:** Joint distribution of time series process remains the same over time

Example:

	T_1	T_2	T_3
Y_t	200	220	231
Y_{t+s}	180	195	210

Consider two time series Y_t and Y_{t+s} representing stock prices for a company at time $T_{1,}$ $T_{2,}$ $T_{3.}$

P_1=Probability of (Y_t =200, Y_{t+s}=180); similarly

P_2=P(Y_t =220, Y_{t+s}=195);

P_3=P(Y_t =231, Y_{t+s}=210);

If P_1= P_2= P_3 then strongly stationary

Joint prabbaility distrubiton should be same for all time series

Strong and Weak stationarity

Weak stationarity: Also known as Covariance stationarity

Requirements for weak stationarity

1. $E(Y_t)=\mu$ Constant mean
2. $E(Y_t-\mu)\ (Y_t-\mu)=\sigma^2$ Constant variance
3. $E(Y_{t1}-\mu)\ (Y_{t2}-\mu)= Y_{t1-t2}$ Constant auto-variance

- All these criteria need to be fulfilled for strong stationarity.
- Requirement for joint distribution to be equal for all time series is not required for weak stationarity.

Non-stationary time series to stationary time series

- Non-stationary data can be differentiated to be made stationary by stabilizing the mean
- Differencing maybe:

Ordinary differencing

$Y_t' = Y_t - Y_{t-1}$

i.e. difference between an observation and previous

Second order differencing

$Y_t'' = Y_t' - Y_{t-1}'$

Second order differencing is required when data does not appear to be stationary even after ordinary differencing.

Almost no data go beyond second order differencing

Non-stationary time series to stationary time series

Seasonal differencing

$Y_t' = Y_t - Y_{t-m}$ where m= number of seasons

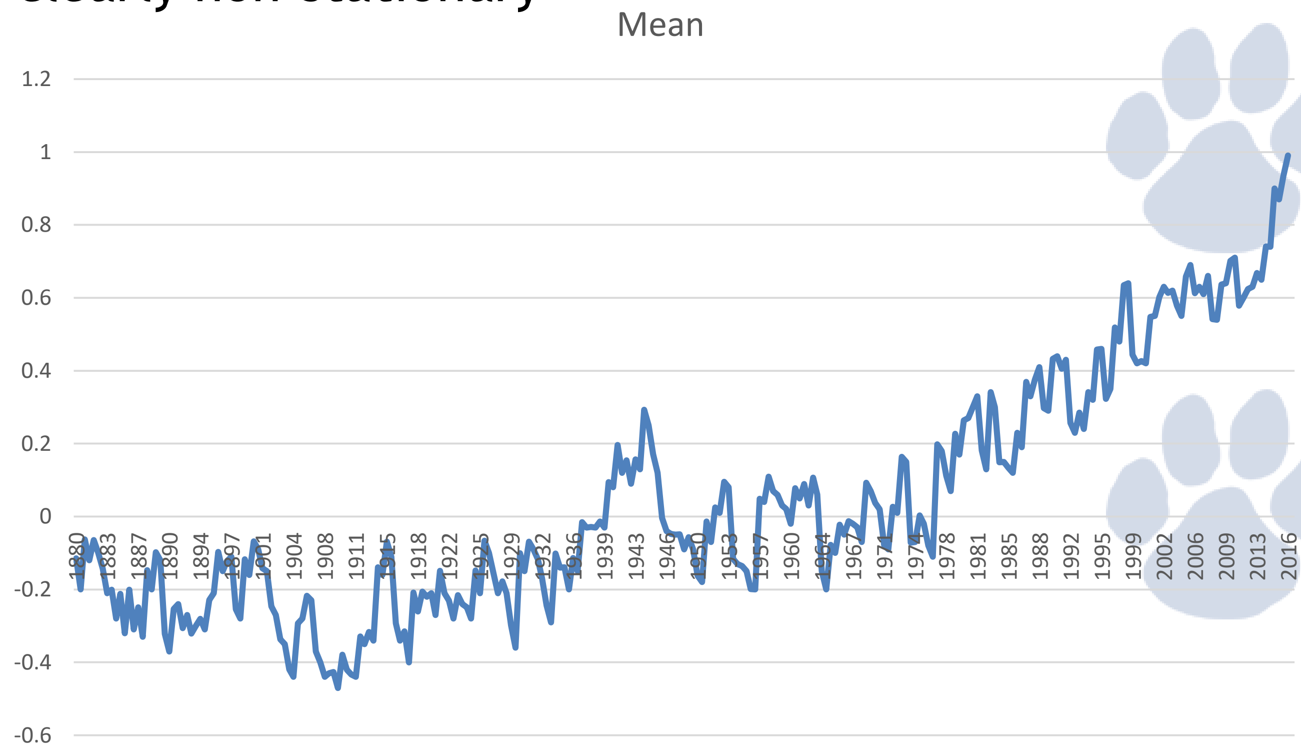
Seasonal differencing is the difference between observation and previous corresponding observation

Non-stationary time series to stationary time series

Example

- We have monthly mean anomalies in global temperature shown here.
- The plot is presented on the right.
- The time series is clearly non-stationary

Year	Mean
1880	-0.1148
1880	-0.2
1881	-0.0628
1881	-0.12
1882	-0.0648
1882	-0.1
1883	-0.1424
1883	-0.21
1884	-0.2009
1884	-0.28

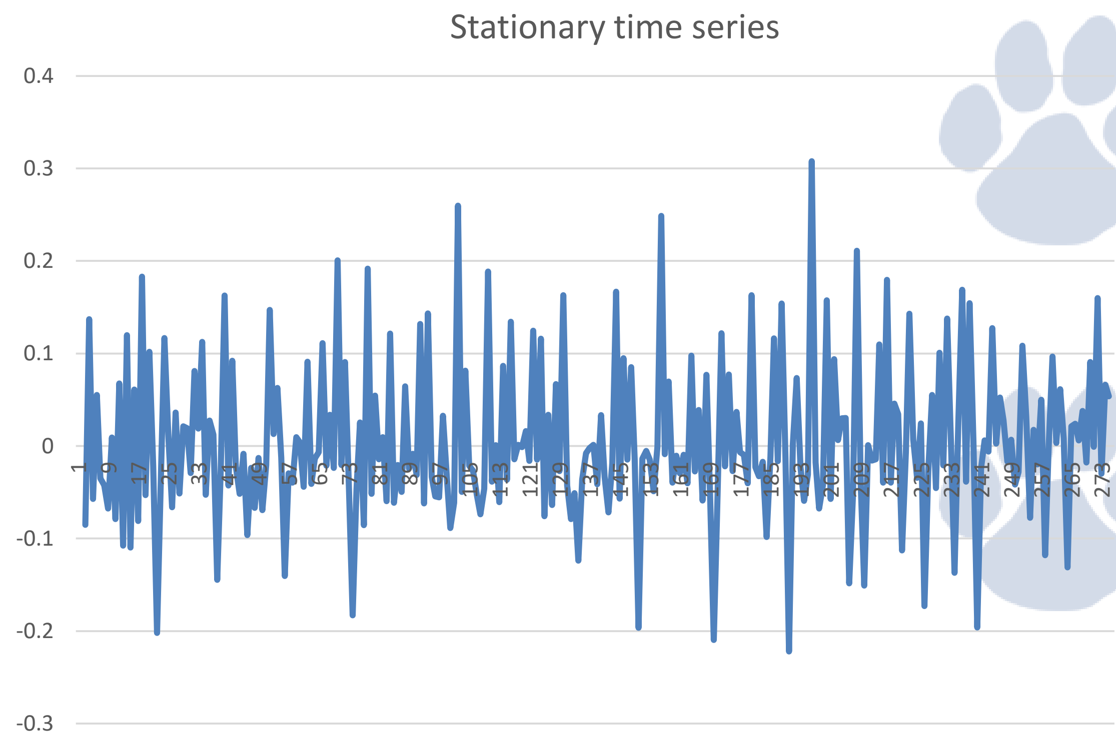


Non-stationary time series to stationary time series

- We differentiate the data by subtracting mean at year t by mean at year t-1 to obtain a stationary time series

-0.2-(-0.1148)

Year	Mean	
1880	-0.1148	
1880	-0.2	-0.0852
1881	-0.0628	0.1372
1881	-0.12	-0.0572
1882	-0.0648	0.0552
1882	-0.1	-0.0352
1883	-0.1424	-0.0424
1883	-0.21	-0.0676
1884	-0.2009	0.0091
1884	-0.28	-0.0791
1885	-0.2125	0.0675



Moving Average (MA) model

- MA model uses past forecast errors instead of past forecast values for regression. MA model can be written as:

$$Y_t = \mu + u_t + \theta_1 u_{t-1} + \theta_2 u_{t-2} \ldots\ldots + \theta_q u_{t-q}$$

$$\Rightarrow Y_t = \mu + \sum_{i=1}^{q} \theta_i u_{t-i} + u_t \quad \ldots\ldots\ldots\ldots\ldots\ldots(1)$$

Here u_t is the error called white noise which is normally distributed with mean zero and SD 1, μ is the intercept.

Lag operator:

$u_{t-2} = Lu_{t-1}$

or, $u_{t-s} = L^s u_t$

Moving Average (MA) model

Lag operator:

$u_{t-2}=Lu_{t-1}$

or, $u_{t-s}=L^{s}u_{t}$

Using the lag operator equation (1) becomes:

$$Y_t= \mu+\sum_{i=1}^{q} \theta_i L^i u_t + u_t$$

$\Rightarrow$ $Y_t= \mu+ \theta(L)u_t$ where $\theta(L)=1+\theta_1 L+\theta_2 L^2+ \cdots +\theta_q L^q$

Properties

1. $E(Y_t)=\mu$ Constant mean
2. $Var(Y_t)= (1+\theta_1^2+\theta_2^2+\theta_3^2+.......+\theta_q^2)\sigma^2$ Constant variance
3. Autocovariance is non-zero till p lags

Auto Regressive (AR) model

Auto Regressive model is when value of time series is regressed from previous values of same time series.

- If we want to predict Y_t using previous values (Y_{t-1}, Y_{t-2}) then the model is written as;

$Y_t = \mu + u_t + \Phi_1 Y_{t-1} + \Phi_1 Y_{t-2}$; Where μ is the intercept and u_t is error term

$\Longrightarrow Y_t = \mu + \sum_{i=1}^{p} \Phi_i Y_{t-i} + u_t$

This second order autoregression is written as AR(2)

- Similar to MA model, AR model can be written using lag operator as:

$Y_t = \mu + \sum_{i=1}^{q} \Phi_i L^i u_t + u_t$

Auto Regressive (AR) model

Testing stationarity

AR model represented as;

$Y_t = \mu + \sum_{i=1}^{p} \Phi_i L^i u_t + u_t$ can be written as:

⇒ $\Phi(L)\, Y_t = \mu + u_t$;

where $\Phi(L) = 1 - \Phi_1 L - \Phi_2 L^2 \ldots - \Phi_p L^p$ is called autoregressive polynomial of Y_t

$Y_t = \Phi(L)^{-1}(\mu + u_t)$

- For AR model to be stationary $\Phi(L)$ should converge to zero
- When $\Phi(L)=0$; this is called characteristic equation

Checking stationarity of time series model

Consider a time series

$Y_t = Y_{t-1} + u_t$ where u_t is the error term

Next value in the sequence if a modification of previous value in the sequence.
This is called random walk.

⇒ $Y_t = LY_t + u_t$ where L is the lag of Y_t

⇒ $Y_t(1-L) = u_t$

⇒ $Y_t(1-z) = 0$; Set polynomial = 0 for characteristic equation

⇒ $z=1$

Therefore;

Roots of the characteristic equation should be greater than 1 for the time series to be stationary.

Checking stationarity of time series model

Example:

Check stationarity for the following time series

$Y_t = 3Y_{t-1} + 2.75Y_{t-2} + 0.75Y_{t-3} + u_t$

⇒ $Y_t = 3LY_t + 2.75L^2Y_t + 0.75L^3Y_t + u_t$

⇒ $Y_t(1-3L + 2.75L^2 + 0.75L^3) = u_t$

⇒ $1-3z + 2.75z^2 + 0.75z^3 = 0$

⇒ $(1-z)(1-1.5z)(1-0.5z) = 0$

⇒ $z = 1, 2/3, 2$

All roots are not greater than 1 so it is a non-stationary time series

Auto Regressive Moving Average (ARMA) model

ARMA Model is a combination of AR and MA models

The process can be written as:

$$Y_t = \mu + u_t + \underbrace{\theta_1 u_{t-1} + \theta_2 u_{t-2} + \ldots.. + \theta_q u_{t-q}}_{\text{AR terms}} + \underbrace{\Phi_1 Y_{t-1} + \Phi_2 Y_{t-2} + \ldots + \Phi_p Y_{t-p}}_{\text{MA terms}}$$

Autocorrelation Function (ACF) and Partial Autocorrelation Function

ACF:

Correlation between current observation time and observation at previous time

- Number of terms in MA model is determined by ACF

PACF:

Correlation between a time series with its own lagged values considering that they are correlated with other lagged values

- Number of terms in AR is determined using PACF

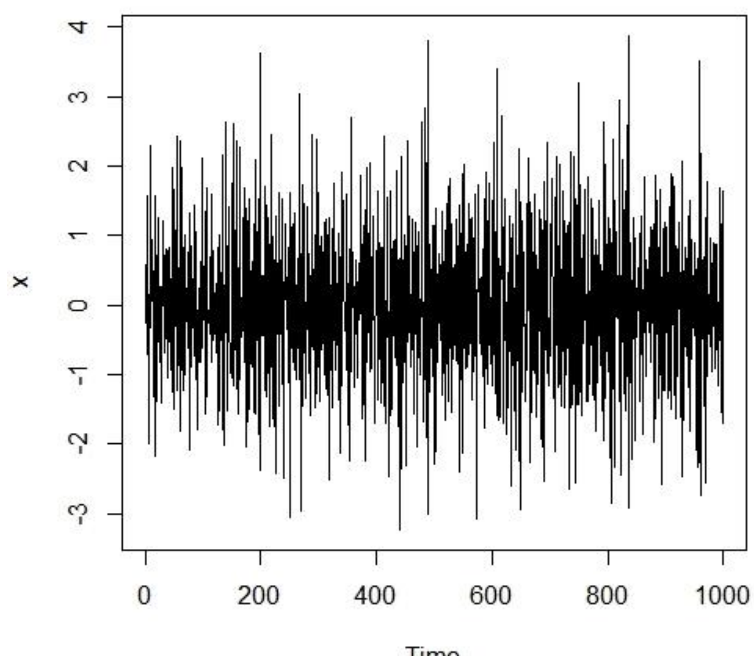
Fitting MA, AR and ARMA models using R

Step1: Identification

- Fitting an ARMA model in R is quite straight forward using R functions.
- In this example we use R functions to simulate a time and then estimate AR, MA and ARMA models

```
set.seed(1)
x <- arima.sim(n=1000, model=list(ar=0.25, ma=-0.75))
```

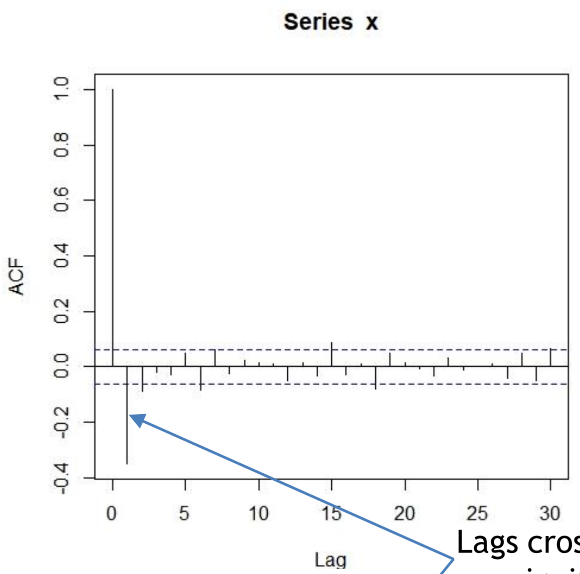
First, we simulate an ARMA model with the coefficient of AR model and MA models as 0.25 and -0.75 as an example



- We plot the generated model using the function plot(x) to visualize and identify the data

Choosing between MA and AR using ACF and PACF

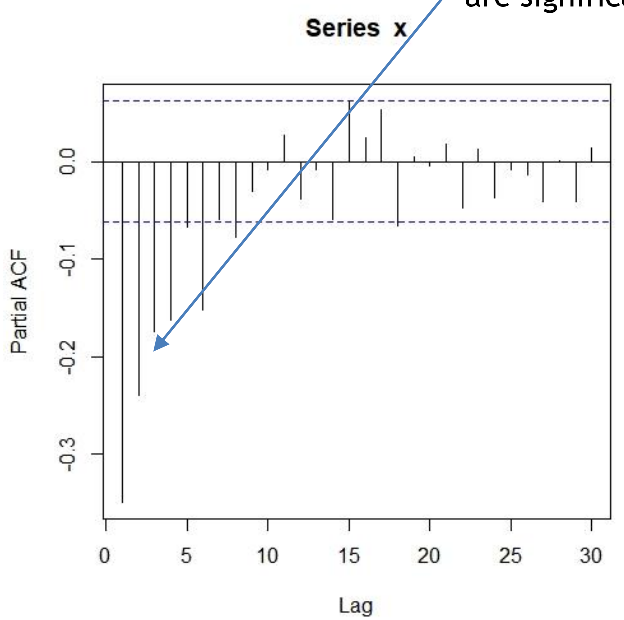
- To choose between MA and AR models we use the ACF and PACF plots
- We look for less complex model with a smaller number of variables that can fit the time series data
- ACF and PACF help identify the number of significant variables in a time series
- In the next slide we present ACF, PACF charts




```
#ACF PACF plots
acf(x)
pacf(x)
```

- Plot the ACF and PACF for the model using the functions:
  acf() and *pacf()* respectively

Lags crossing the dotted line (confidence interval) represent the lags that are significant.



- Look at the number of lags crossing the dotted line, lower the number less complex is the resulting model used.
- Here, ACF has a smaller number of significant lags, meaning using MA model it will result in a simpler model with less number of variables.
- By visual observation we see lag(1), lag(2), lag(6), lag(16), lag(18) are significant in ACF

Fitting MA, AR and ARMA model using R: Estimating MA, AR and ARMA

```
arima(x,order=c(1,0,0))   (1)
arima(x, order=c(0, 0, 1))   (2)
arima(x, order=c(1, 0, 1))   (3)
```

(1) ➡ Fits first order AR model

(2) ➡ Fits first order MA model

(3) ➡ Fits first order ARMA model

- Based on ACF and PACF we saw that the MA model would be the better one,
- Nevertheless here we model single order MA, AR and ARMA using the function *arima()* to give you an idea of the modeling approach
- The function outputs the coefficients of the model along with log-likelihood and AIC values
- Note that all models used here are first order models.

Fitting ARMA model using R: Estimating MA, AR and ARMA

```
> arima(x, order=c(1, 0, 0))

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

Coefficients:
          ar1  intercept
      -0.3499    -0.0042
s.e.   0.0296     0.0260

sigma^2 estimated as 1.233:  log likelihood = -1523.59,  aic = 3053.19
```

AR model

```
> arima(x, order=c(0, 0, 1))

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

Coefficients:
          ma1  intercept
      -0.6524    -0.0044
s.e.   0.0309     0.0115

sigma^2 estimated as 1.099:  log likelihood = -1466.46,  aic = 2938.93
```

MA model

```
> arima(x, order=c(1, 0, 1))

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

Coefficients:
         ar1      ma1  intercept
      0.2517  -0.7969    -0.0044
s.e.  0.0436   0.0265     0.0089

sigma^2 estimated as 1.066:  log likelihood = -1451.11,  aic = 2910.22
```

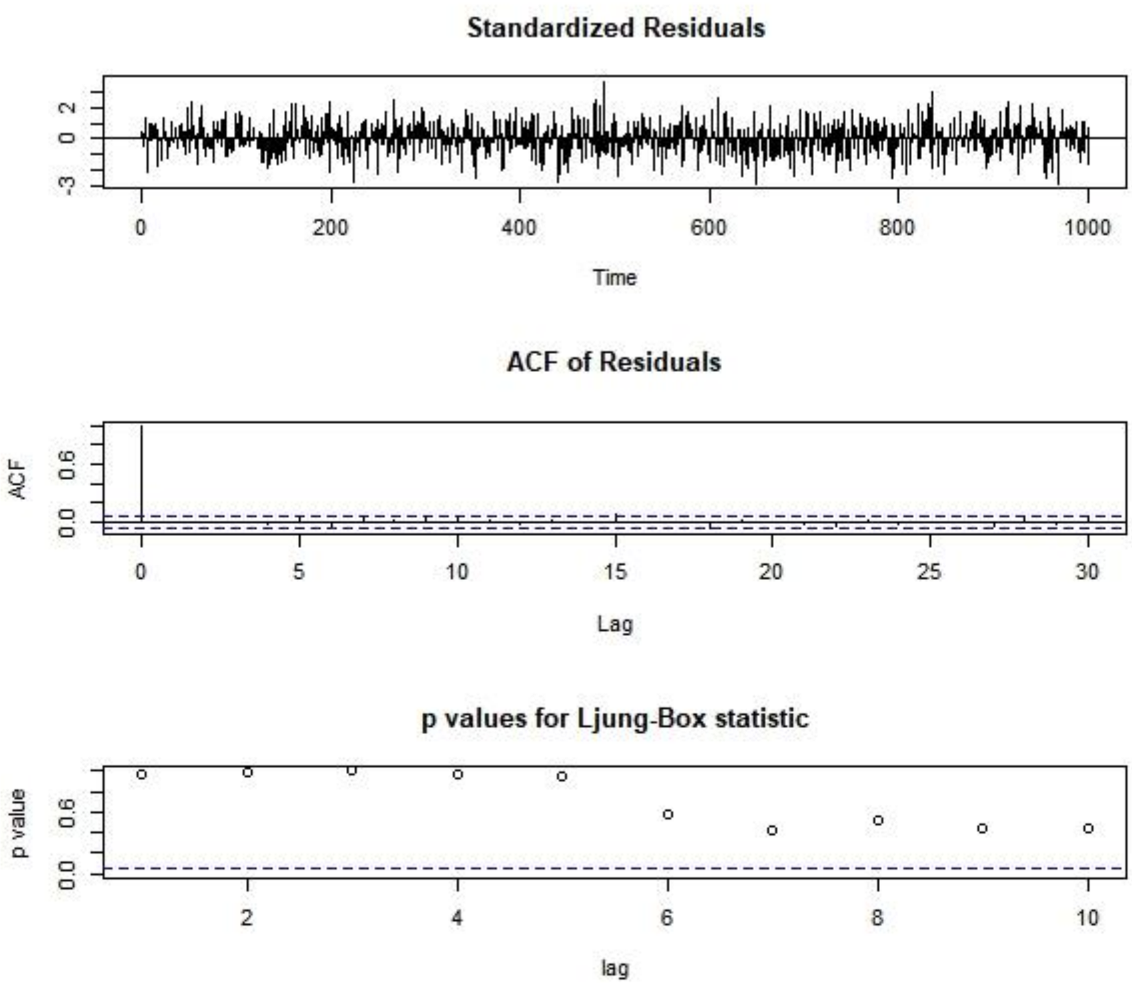
ARMA model

- Model selection can be based on AIC values with low AIC being selected
- Here ARMA (1,1) seems to be the better model based on AIC values
- Furthermore higher order models can be experimented with to find the best model for the time series

Fitting ARMA model using R: Diagnostic checking

```
tsdiag(arima(x, order=c(1, 0, 1)))
```

R function *tsdiad* can be used to create residual plots to assess model accuracy



For ARMA(1,1) we see no clustering of residuals.

ACF of residuals is also within the confidence interval

ARMA(1,1) is fit for this time series