TIME SERIES MODEL-ARIMA

DATA SCIENCE
INSTITUTE

Box-Jenkins (ARIMA) Models

- ARIMA models are statistical models that use lagged values of the dependent variable and/or random disturbance terms as explanatory variables.

- ARIMA models rely heavily on the autocorrelation pattern in the data

- ARIMA models can also be developed in the presence of seasonality in the time series. (SARIMA- to be discussed in the next session)

Box-Jenkins (ARIMA) Models

- ARIMA models thus essentially **ignore domain theory** (by ignoring “traditional” explanatory variables),

- Why use them?

- The use of ARIMA is appropriate when:

  - Little or nothing is known about the dependent variable being forecasted,
  - The independent variables known to be important cannot be forecasted effectively
  - Objective is to obtain short term forecasts

Box-Jenkins (ARIMA) Models

Three basic ARIMA models for a stationary time series y_t :

(1) Autoregressive model of order *p* (AR(*p*))

$$y_t = b_0 + b_1 y_{t-1} + b_2 y_{t-2} + \cdots + b_p y_{t-p} + \varepsilon_t,$$

i.e., y_t depends on its *p* previous values

(2) Moving Average model of order *q* (MA(*q*))

$$y_t = a_o + \varepsilon_t + a_1 \varepsilon_{t-1} + a_2 \varepsilon_{t-2} + \cdots + a_q \varepsilon_{t-q},$$

i.e., y_t depends on *q* previous random error terms

DATA SCIENCE INSTITUTE

Box-Jenkins (ARIMA) Models

(3) Autoregressive-moving average model of order *p* and *q* (ARMA(*p*,*q*))

$$y_t = b_0 + b_1 y_{t-1} + b_2 y_{t-2} + \cdots + b_p y_{t-p}$$
$$+ \varepsilon_t + a_1 \varepsilon_{t-1} + a_2 \varepsilon_{t-2} + \cdots + a_q \varepsilon_{t-q},$$

i.e., … rms

A Five-Step Modeling Procedure

1) Stationarity Checking and Differencing
2) Model Identification
3) Parameter Estimation
4) Diagnostic Checking
5) Forecasting

Step One: Stationarity Checking

Stationary Time Series

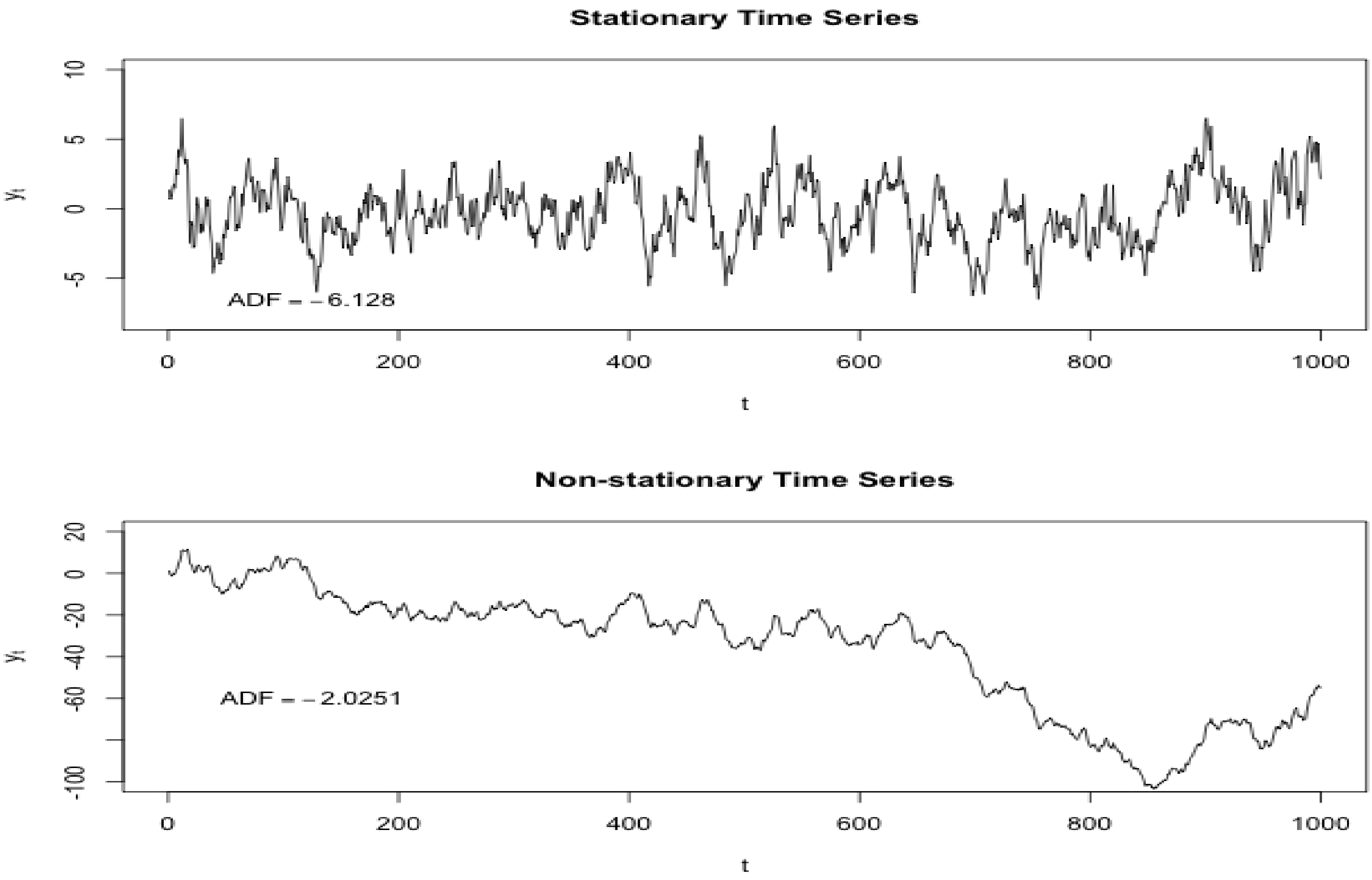
- Time series process is called **stationary** if the **statistical properties** of the process remain unchanged over time.

i.e if Y_t is a time series t=1,2,3,…

- $E(Y_t) = \mu_t = \mu$ (constant) $\forall\, t=1,2,\ldots$
- $Var(Y_t) = \sigma_t^2 = \sigma^2$ (constant) $\forall\, t=1,2\ldots$
- $cov(Y_t, Y_{t-s})$ depends only on **s**(lag),and is independent of **t** (time).

DATA SCIENCE INSTITUTE

Stationary Time Series



Assessing Stationarity of Time Series

Stationarity of a time series can be assessed using:

- Time Series Plot (Time vs. Variable)
- Correlogram
- Dickey-Fuller Test

Non-Stationary time series can be converted into stationary using 'differencing' .

ndiffs() function in forecast package provides number of times time series should be differenced to achieve stationarity.

Differencing

- Differencing continues until stationarity is achieved.

$$\boxed{\Delta y_t = y_t - y_{t-1}} \qquad \boxed{\Delta^2 y_t = \Delta(\Delta y_t) = \Delta(y_t - y_{t-1}) = y_t - 2y_{t-1} + y_{t-2}}$$

  The differenced series has n-1 values after taking the first-difference, n-2 values after taking the second difference, and so on.

- The number of times that the original series must be differenced in order to achieve stationarity is called the <u>order of integration</u>, denoted by d.

- In practice, it is not required to go beyond second difference.

DATA SCIENCE INSTITUTE

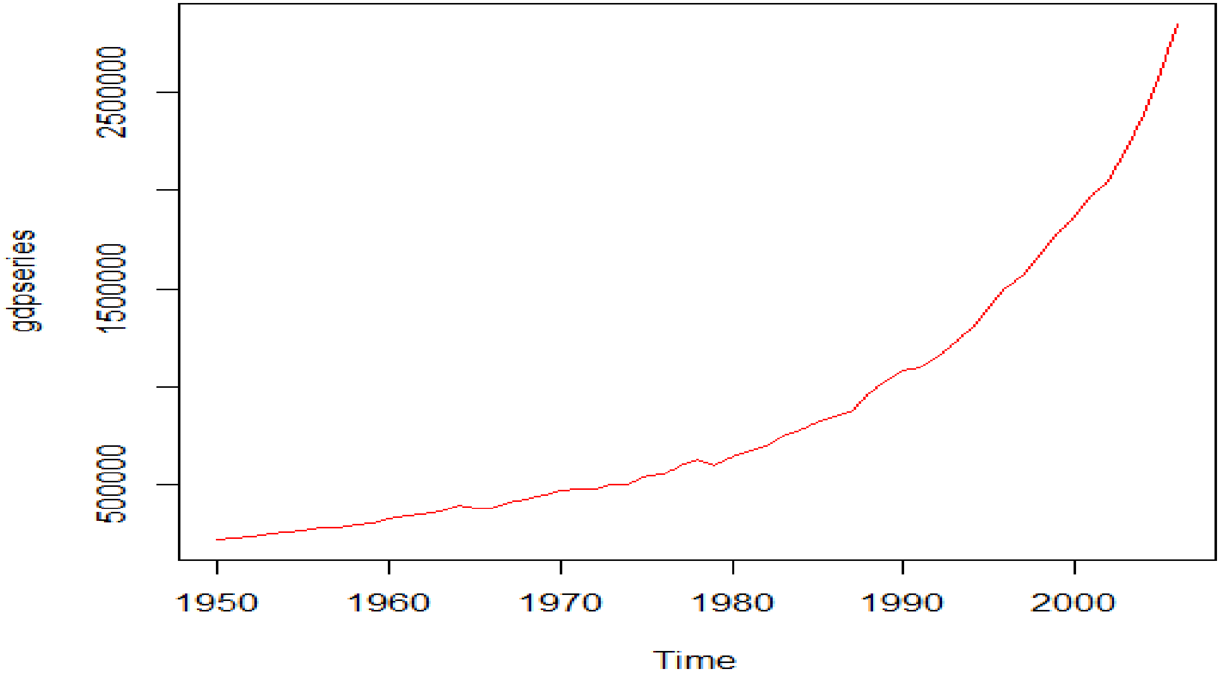
GDP Time Series
Data Snapshot

Year	GDP
1950-51	224786
1951-52	230034
1952-53	236562
1953-54	250960
1954-55	261615
1955-56	268316
1956-57	283589
1957-58	280160
1958-59	301422
1959-60	308018
1960-61	329825
1961-62	340060
1962-63	347253
1963-64	364834
1964-65	392503
1965-66	378157
1966-67	382006
1967-68	413094
1968-69	423874

This is partial data.
The data has GDP values for 1950-51 to 2006-07

Time Series Analysis in R
Plot Time Series

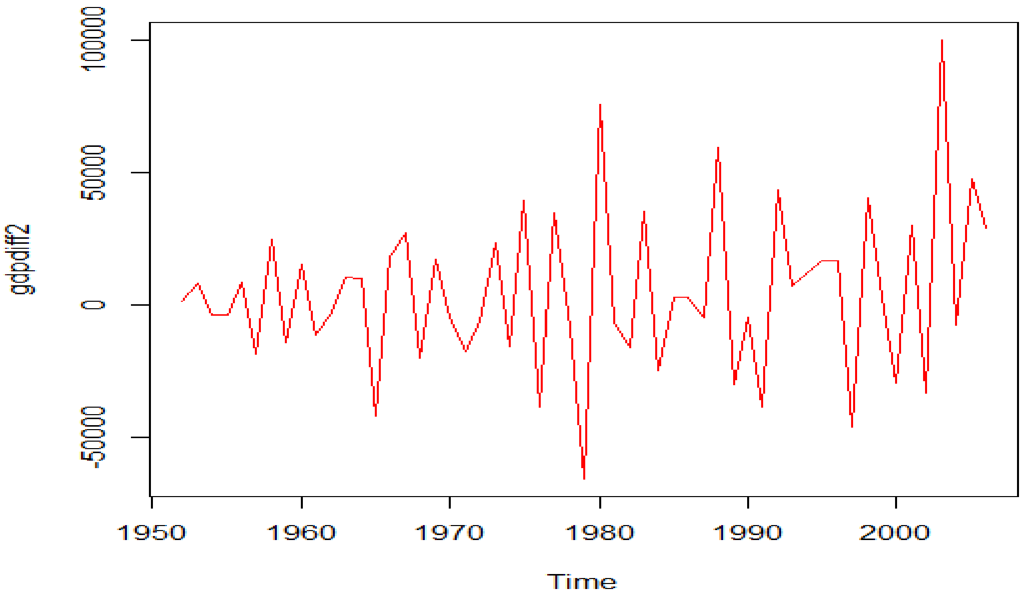
```
gdpdata<-read.csv(file.choose(),header=T)
gdpseries<-ts(gdpdata$GDP,start=1950,end=2006)
plot(gdpseries,col="red")
```



Clearly a non-stationary time series.

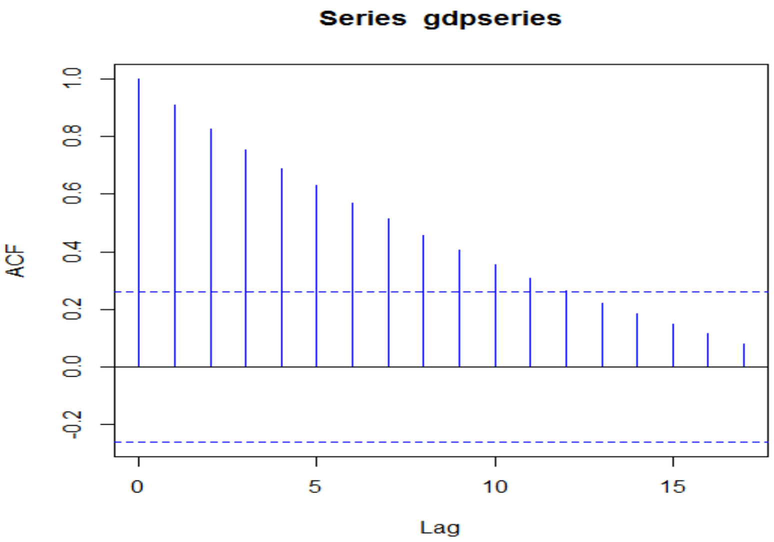
How Many Times Should Time Series Be Differenced to Make Stationary?

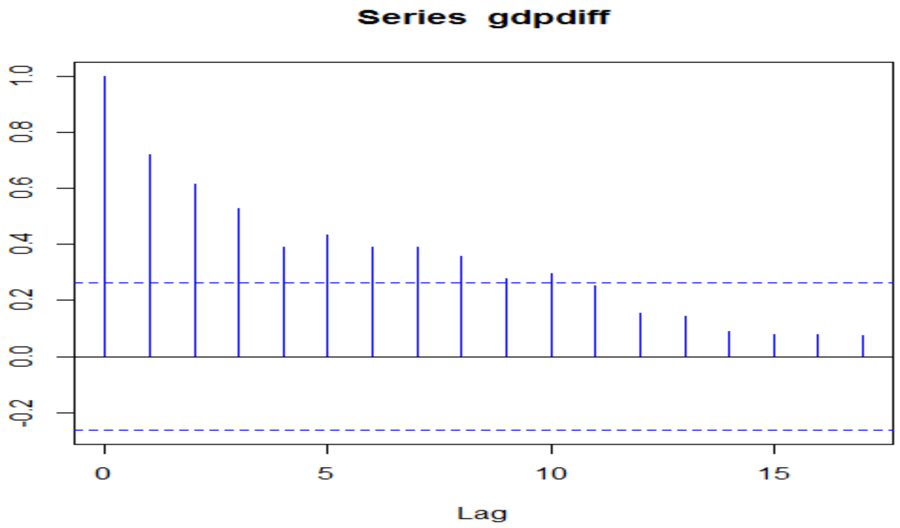
```
install.packages("forecast")
library(forecast)
ndiffs(gdpseries)   # gives 2
gdpdiff2<-diff(gdpseries,differences=2)
plot(gdpdiff2,col="red")
```

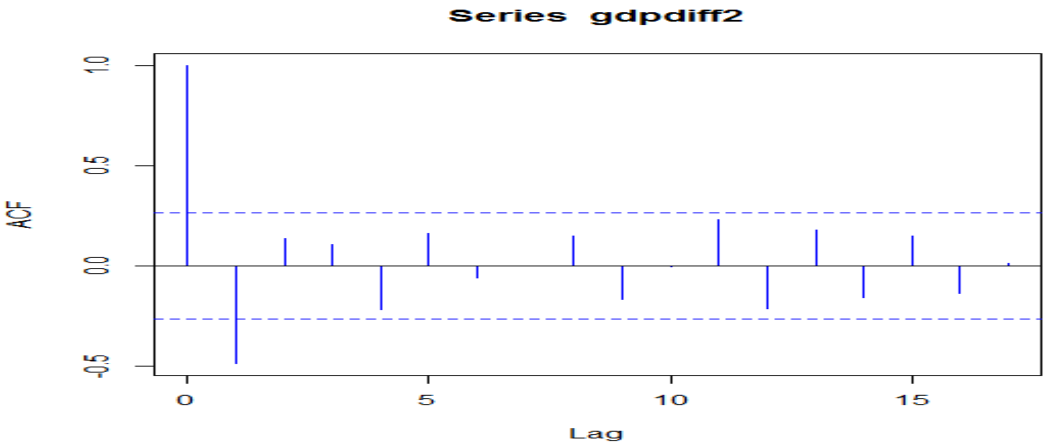


DATA SCIENCE
INSTITUTE

Time Series Analysis in R
Correlograms







Stationarity is achieved with second order difference

DATA SCIENCE
INSTITUTE

Time Series Analysis in R
Dickey Fuller Test…

```
library(urca)
df<-ur.df(gdpseries,lag=0)
summary(df)
```

Value of test-statistic is: 19.2745

Critical values for test statistics:

	1pct	5pct	10pct
tau1	-2.6	-1.95	-1.61

Inference: Time series is non-stationary. Value of test statistic is greater than 5% critical value.

Time Series Analysis in R
Dickey Fuller Test

```
library(urca)
df<-ur.df(gdpdiff2,lag=0)
summary(df)
```

```
Value of test-statistic is:  -11.9083

Critical values for test statistics:

        1pct   5pct  10pct
tau1    -2.6  -1.95  -1.61
```

Inference: Time series is stationary. Value of test statistic is less than 5% critical value.

Step Two: Model Identification

Model Identification

- When the data are confirmed stationary, one may proceed to tentative identification of models through visual inspection of correlogram and partial correlogram.

- Some guidelines exist to identify models using correlogram and partial correlogram.

- In practice, it is not always easy to identify model using visualization.

  However, R /Python has built in function to identify best model which can be used for forecasting.

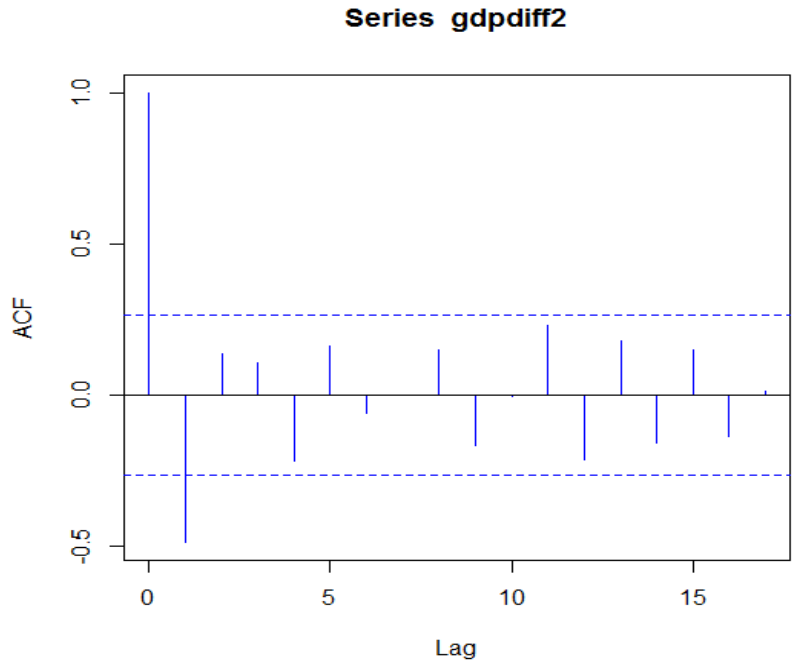
Summary of the Behaviour of Autocorrelation and Partial Autocorrelation Functions

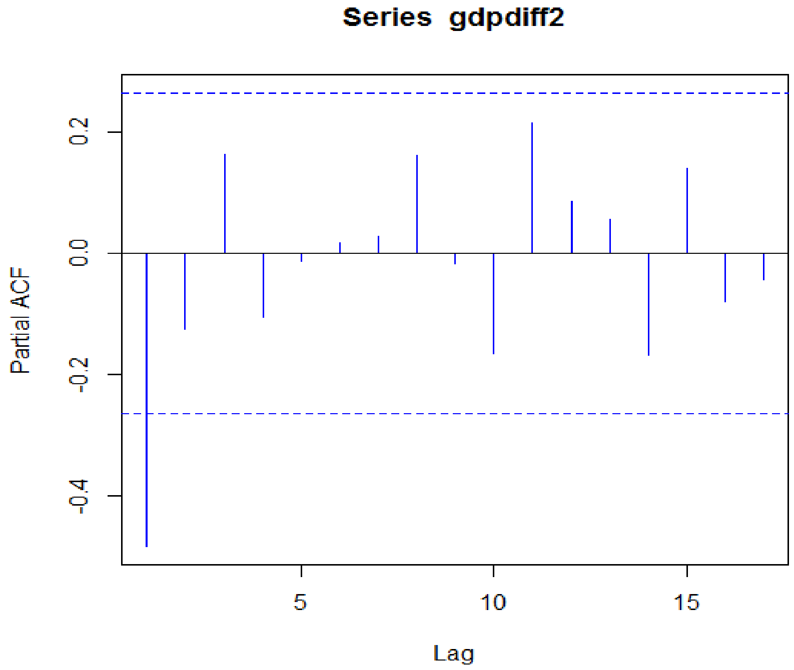
Model	AC	PAC
Autoregressive of order p	Dies down	Cuts off after lag p
Moving Average of order q	Cuts off after lag q	Dies down
Mixed Autoregressive-Moving Average of order (p,q)	Dies down	Dies down

The meaning of dies down is “gradual decrease”

DATA SCIENCE INSTITUTE

Model Identification





Indicative Statistical Model – ARIMA(1,2,1)
Where,
number of autoregressive terms= 1
order of differencing= 2
number of moving average terms= 1

Step Three: Parameter Estimation

Parameter Estimation

- The method of least squares can be used. However, for models involving an MA component MLE is used.
- Given n observations $y_1, y_2, \ldots, y_n$, the likelihood function L is defined to be the probability of obtaining the data actually observed.
- The maximum likelihood estimators (M.L.E.) are those value of the parameters for which the data actually observed are most likely, that is, the values that maximize the likelihood function L.

Time Series Analysis in R
ARIMA Model in R

```
gdpmodel<-arima(gdpseries,order=c(1,2,1))
coef(gdpmodel)
```

```
> coef(gdpmodel)
 ar1        ma1
-0.3654655 -0.1202087
```

```
AIC(gdpmodel)
```

```
> AIC(gdpmodel)
[1] 1285.361
```

Smaller the AIC value, better is the model. We need to try out various combinations of AR and MA terms to arrive at final model.

DATA SCIENCE INSTITUTE

Time Series Analysis in R
ARIMA Model in R…

```
library(forecast)
gdpmodel<-auto.arima(gdpseries,d=2,max.p=1,max.q=1,trace=TRUE,ic="aic")
```

```
> auto.arima(gdpseries,d=2,max.p=1,max.q=1,trace=TRUE,ic="aic")

 ARIMA(1,2,1)                    : 1285.361
 ARIMA(0,2,0)                    : 1294.497
 ARIMA(1,2,0)                    : 1283.644
 ARIMA(0,2,1)                    : 1285.212

 Best model: ARIMA(1,2,0)
```

DATA SCIENCE INSTITUTE

Time Series Analysis in R
ARIMA Model in R…

coef(gdpmodel)

```
> coef(gdpmodel)
       ar1
-0.4555743
```

AIC(gdpmodel)

```
> AIC(gdpmodel)
[1] 1283.644
```

Brief Note
Model Selection Criteria

- Akaike Information Criterion (AIC)

$$AIC = -2 \ln(L) + 2k$$

- Schwartz Bayesian Criterion (SBC)

$$SBC = -2 \ln(L) + k \ln(n)$$

where L = likelihood function
k = number of parameters to be estimated,
n = number of observations.

- Ideally, the AIC and SBC should be as small as possible

DATA SCIENCE INSTITUTE

Step Four: Diagnostic Checking

Residual Analysis

- If an ARMA(p,q) model is an adequate representation of the data generating process, then the residuals should be 'White Noise'.

- A white noise process is a serially uncorrelated, zero-mean, constant and finite variance process.

- Under the null hypothesis that y_t is a white noise process, the **Box-Pierce Q-statistic (based on autocorrelations upto lag m and T observations in a time series)**

$$\boxed{Q_{BP} = T\sum_{\tau=1}^{m} \hat{\rho}^2(\tau) \sim \chi^2(m)}$$ for large T.

- Another closely connected statistical test is Ljung-Box test.

DATA SCIENCE INSTITUTE

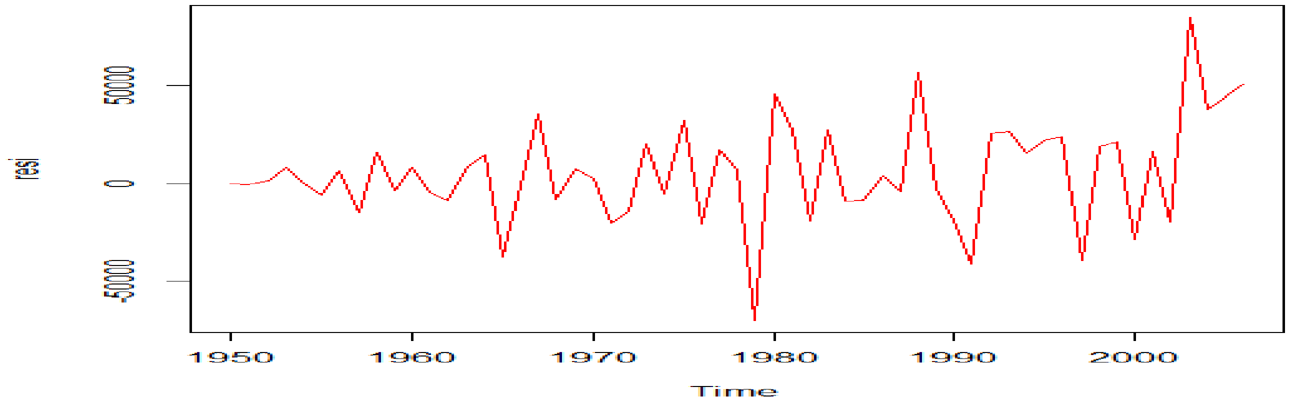
Time Series Analysis in R
ARIMA Model in R…

```
resi<-residuals(gdpmodel)
Box.test(resi)
plot(resi,col="red")
```

```
Box-Pierce test

data:  resi
X-squared = 0.5391, df = 1, p-value = 0.4628
```

Do not reject Ho.
Errors follow white noise



DATA SCIENCE INSTITUTE

Step Five: Forecasting

Time Series Analysis in R
ARIMA Model in R…

```
predict(gdpmodel,n.ahead=3)
```

Time Series:
Start = 2007
End = 2009
Frequency = 1
[1] 3078683 3315176 3548951

****
****

THANK YOU!!

DATA SCIENCE
INSTITUTE