Time Series Analysis

Time Series

- Modelling relationships of data collected over a period of time (daily, weekly, monthly, quarterly, yearly).
- **Examples**:
  - ➢ Stock Price
  - ➢ Inflation data
  - ➢ Cost of living etc.
- **Used for**
  - ✓ Identifying trends
  - ✓ Forecasting

- **When lags are ignored**
  - ✓ Stock price of a day depends on the previous day, inflation price depends on previous value, bank balance of a month depends on the previous month's balance etc
  - ✓ Regression does not account for these relationships and overestimates the relationship of X and Y

Univariate Time Series

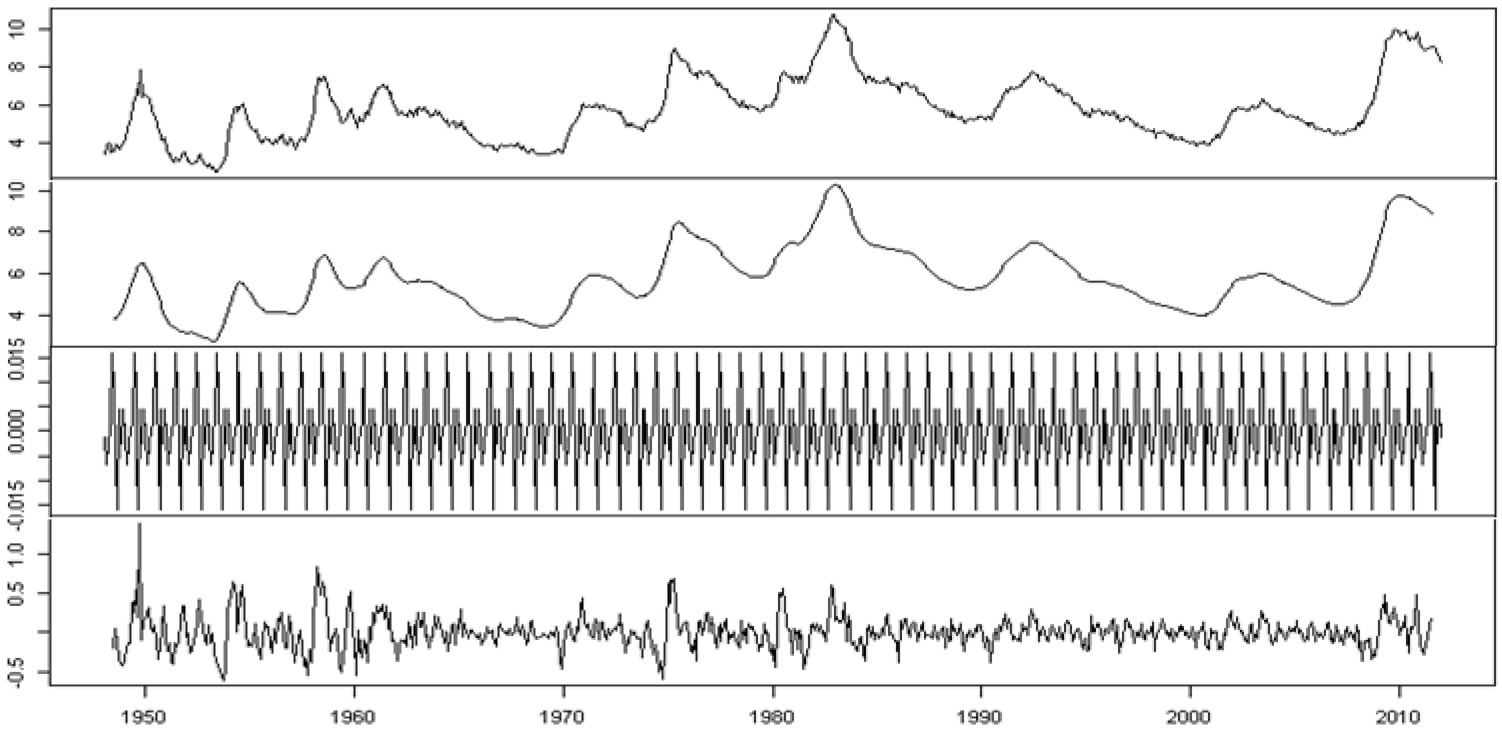
- A time series that has a single time-dependent variable
- Eg: Time ~ Stock Closing Price

Multivariate Time Series

- A time series that has more than one time-dependent variable
- Eg: Time ~ Temperature + Humidity + Cloud_Cover + Wind_Speed

Time Series components

- **Time Series data has 4 components**
  - ✓ **Secular**: Variables tend to increase or decrease over a period of time. eg: Cost of living (over a period of time)
  - ✓ **Cyclic**: Ups and downs. eg: Business cycle. Unpredictable pattern
  - ✓ **Seasonal**: A pattern (trend) that gets repeated every year at the same time period
  - ✓ **Irregular**: No definite pattern. Causes aren't exactly known



→ **Secular (Trend)**

→ **Cyclic**

→ **Seasonal**

→ **Irregular**

Stationarity in Time Series

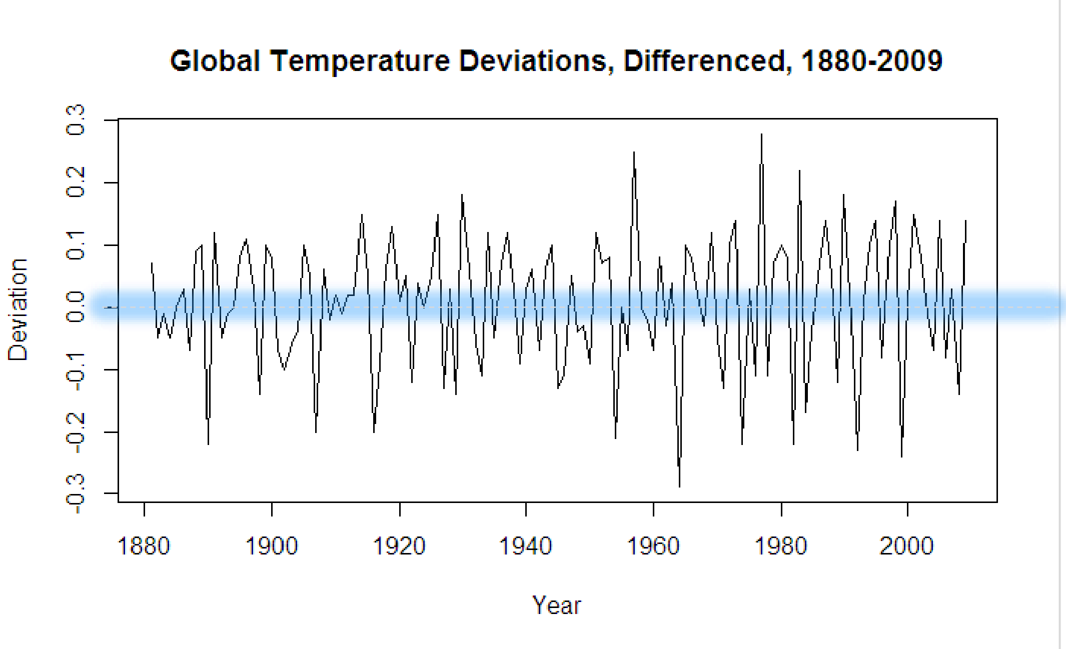
- AR models need to be "Stationary"
- Otherwise, forecasting will not be possible
- If time-series data is not "stationary", then it has to be made "stationary"

Stationarity Time Series

- Joint probability of a series doesn't change over time
  - ✓ i.e. Mean and Variance of data remains constant over time
- There should be no trend

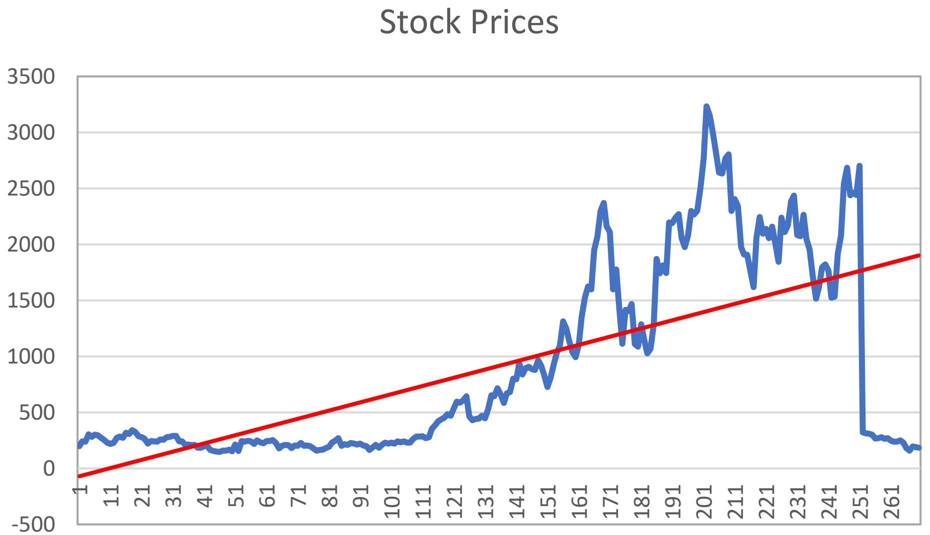
- **Reasons for non-stationarity**
  - Trend in Series
  - Seasonality in Series

- **Check the stationarity of data**
  - ➢ **Augmented Dickey-Fullter (ADF) test**
  - ▪ If p-value < 0.05 : Data is stationary
  - ▪ If p-value > 0.05 : Data is not stationary

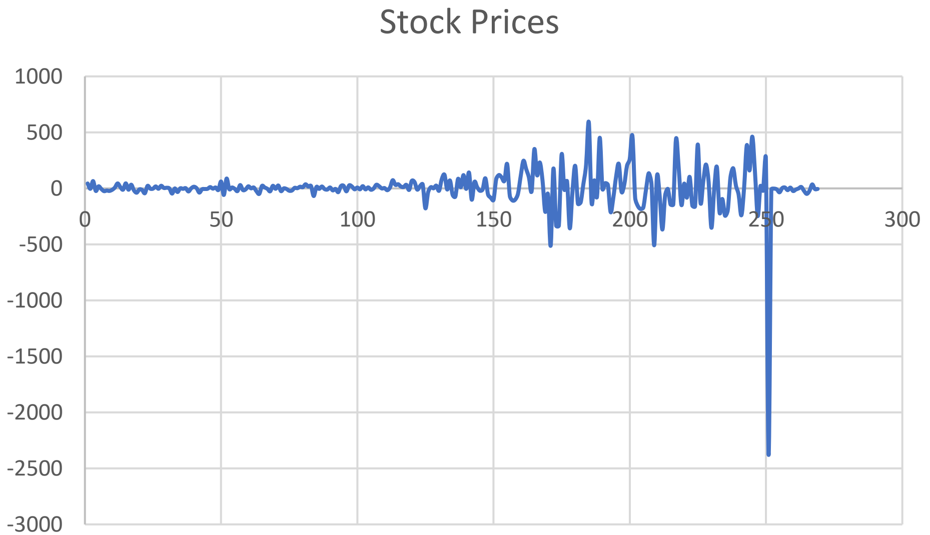


Making a Time-Series Stationary

- Differencing
- Data Transformation
- EDA techniques (adjusting outliers)



Trend (Upward)



Detrended / Stationarity

Differencing (d = 1)

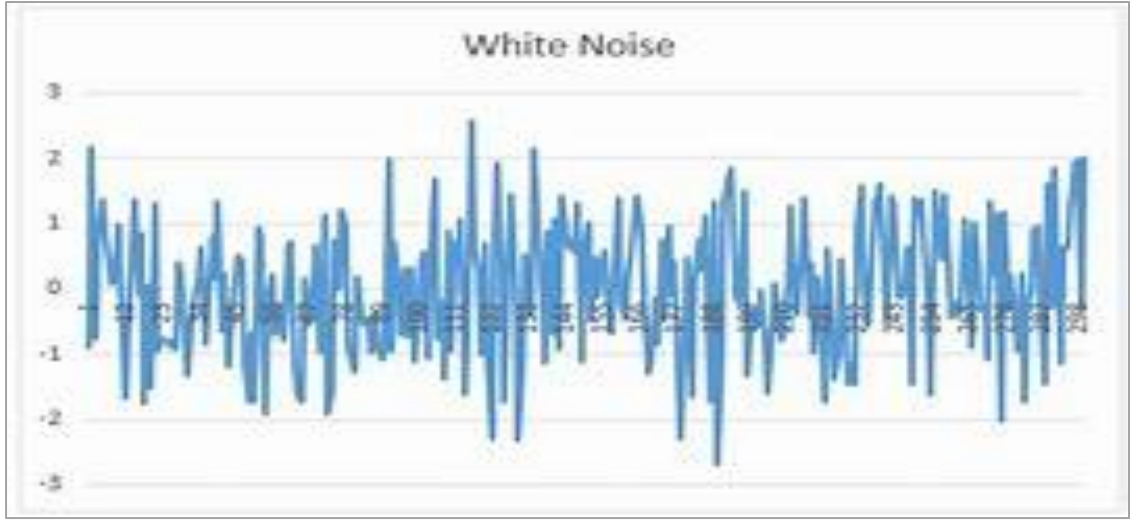
$\mathbf{Y_t - y_{t-1}}$

(First order difference)

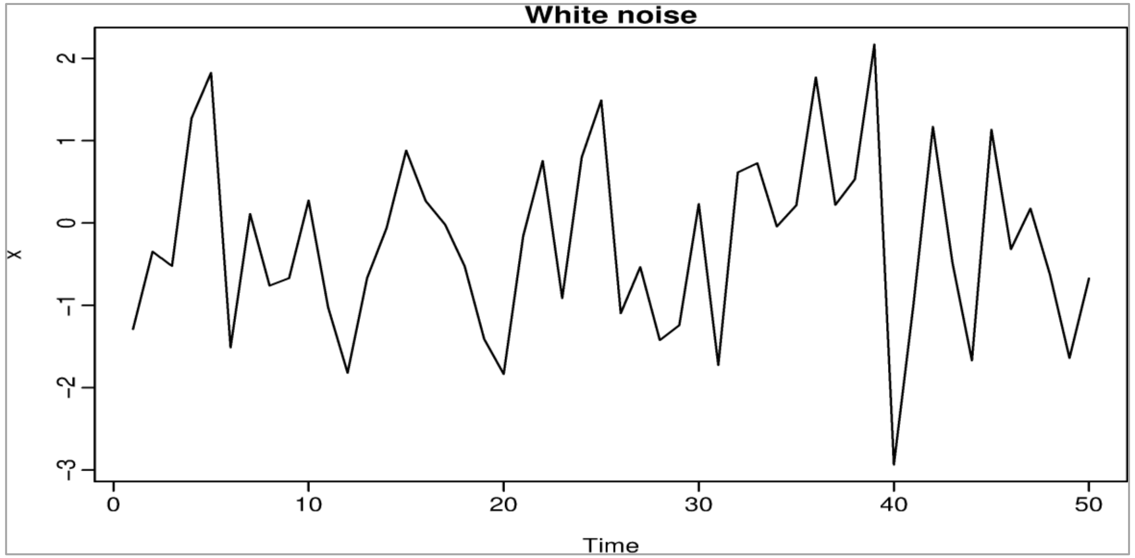
Time Series – White Noise

White noise - characteristics

- It is a type of Time-series
- Variables are random
- Variables exhibit IID
- **Mean = 0; Variance** = Constant
- Each variable has 0 correlation with other (i.e. correlation between lags = 0)



- **White noise is an important concept in TS analysis and forecast because**
  - Cannot predict well with randomness
  - TS errors should ideally be White Noise



White noise

- White Noise has to be checked on the data
  - Plot the data to identify trends

- In case of White noise violations, they have to be corrected before prediction / forecast

- Test for White Noise
  - **Box-Pierce** testing using **Ljung-Box** technique
  - If p_value < 0.05, ***Bad Model*** else ***Good Model***

Smoothing Techniques

Smoothing

- Pre-processing techniques to remove noise from the data (Trends and Seasonality)

- Important patterns are highlighted

- Helps in better predictions / forecasting of data

- Smoothing Methods
  - **MA(Moving Average)**
  - **Exponential Smoothening**
    - **Simple**
    - **Double**
    - **Triple**

(Simple) Moving Average (SMA)

Moving Average

- A series of averages of different subsets and taking the error from the previous time periods

- Moving Average is an MA(q) process

- **Formula**

$$Y_t = C + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots\ldots\ldots + \theta_p \varepsilon_{t-q}$$

Where

C –> Constant; ε_t -> Current Error; ε_{t-1} – Error from the previous Time period

- Technique used to smoothen the data by constantly creating updated average price

- Used in forecasting long-time / short-time trends

- Assumption: Future observations will be similar to past observations

- Typical lags are defined as
  - Short-term MA 5-25 days (very sensitive)
  - Intermediate 25-100 days
  - Long-term 100-250 days (less sensitive)

Consider the following data

Given the **month** and a **Y-value** (let’s assume it is the total sales done)

Calculate the Moving Average

- To calculate the SMA, we can consider any number of lags
- For this example, lets assume the lags = 3

month	Y
Q1-2010	147772
Q2-2010	154400
Q3-2010	166188
Q4-2010	170202
Q1-2011	173264
Q2-2011	175371
Q3-2011	184957
Q4-2011	186935
Q1-2012	191130
Q2-2012	191213
Q3-2012	195749
Q4-2012	198262
Q1-2013	199980
Q2-2013	209566
Q3-2013	212529
Q4-2013	213754
Q1-2014	222124
Q2-2014	224372
Q3-2014	229871
Q4-2014	236260

month	Y	PredY	err
Q1-2010	147772		
Q2-2010	154400		
Q3-2010	166188		
Q4-2010	170202	156120	14082
Q1-2011	173264		
Q2-2011	175371		
Q3-2011	184957		
Q4-2011	186935		
Q1-2012	191130		
Q2-2012	191213		
Q3-2012	195749		
Q4-2012	198262		
Q1-2013	199980		
Q2-2013	209566		
Q3-2013	212529		
Q4-2013	213754		
Q1-2014	222124		
Q2-2014	224372		
Q3-2014	229871		
Q4-2014	236260		

month	Y	PredY	err
Q1-2010	147772		
Q2-2010	154400		
Q3-2010	166188		
Q4-2010	170202	156120	14082
Q1-2011	173264	163596.7	9667.33
Q2-2011	175371		
Q3-2011	184957		
Q4-2011	186935		
Q1-2012	191130		
Q2-2012	191213		
Q3-2012	195749		
Q4-2012	198262		
Q1-2013	199980		
Q2-2013	209566		
Q3-2013	212529		
Q4-2013	213754		
Q1-2014	222124		
Q2-2014	224372		
Q3-2014	229871		
Q4-2014	236260		

month	Y	PredY	err
Q1-2010	147772		
Q2-2010	154400		
Q3-2010	166188		
Q4-2010	170202	156120	14082
Q1-2011	173264	163596.7	9667.3
Q2-2011	175371	169884.7	5486.3
Q3-2011	184957	172945.7	12011.3
Q4-2011	186935	177864	9071
Q1-2012	191130	182421	8709
Q2-2012	191213	187674	3539
Q3-2012	195749	189759.3	5989.6
Q4-2012	198262	192697.3	5564.6
Q1-2013	199980	195074.7	4905.3
Q2-2013	209566	197997	11569
Q3-2013	212529	202602.7	9926.3
Q4-2013	213754	207358.3	6395.6
Q1-2014	222124	211949.7	10174.3
Q2-2014	224372	216135.7	8236.3
Q3-2014	229871	220083.3	9787.6
Q4-2014	236260	225455.7	10804.3

Exponential Smoothening

Exponential Smoothening

- Technique used to make short-term forecasts

- Recent observations given more weightage compared to older values

- There are 3 main types of Exponential Smoothening
  - Simple Exponential Smoothening
  - Double Exponential Smoothening
  - Triple Exponential Smoothening

Simple Exponential Smoothening

- Implemented to a univariate dataset that has no trend or seasonality
- Past data get smaller weights compared to recent ones
- Short term forecasting
- Requires a smoothing factor **α (alpha) { 0 (insentitive) <= α <= 1 (sensitive) }**
- $\mathbf{F_{t+1} = \alpha(A_t) + (1-\alpha)(F_t)}$

  where

  F_{t+1} = forecast at Time t

  A_t = Actual value at Time t

Check XL for exercise

F:\work\2 myPresentation\2 ml\2 algorithms\4 time series\xl\ smoothing_techniques.xls

Double Exponential Smoothening

- Holt's Trend method

- Data has trend, but no seasonality

- Requires 2 smoothing factors **α (alpha)** and **β (beta)** **{ 0 – 1 }**

- $\mathbf{Y_{t+1} = S_t + (h)T_t}$
- $\mathbf{S_t = \alpha(Y_t) + (1-\alpha)(S_{t-1} + T_{t-1})}$

  $\mathbf{T_t = \beta(S_t - S_{t-1}) + (1-\beta)(T_{t-1})}$

where

S_t : smoothed (Levelled) forecast at time t

A_t: Actual value at time t

T_t: Trend forecast value at time t

Check XL for exercise

F:\work\2 myPresentation\2 ml\2 algorithms\4 time series\xl\ smoothing_techniques.xls

Triple Exponential Smoothening

- Holt Winter's Exponential smoothing
- Data has trend and seasonality
- Requires 3 smoothing factors **α (alpha), β (beta) and g (gamma) { 0 – 1 }**

Time-Series Models

Auto Regressive (AR)

- Autocorrelation is a mathematical representation of the degree of similarity between a given time series and a lagged version of itself over successive time intervals.

- $Y_t = a + b_1 y_{t-1} + b_2 y_{t-2} + \ldots\ldots\ldots + b_p y_{t-p} + \varepsilon_t$

  Y_t → Current time period for which prediction is made
  a → Intercept (constant) term
  b → Coefficient of the lagged term
  Y_{t-p} → Previous time period(s)
  ε_t → Error / disturbance term (white noise: mean=0, variance is constant)

- Lies between -1 -> +1

$$\text{Autocorrelation} = \frac{\Sigma[(y_t - \bar{y})(y_{t-k} - \bar{y})]}{\sum(y_t - \bar{y})^2}$$

y_t = current time
Y_{t-k} = previous time at lag k
k = lag number
$\bar{y}$ = mean

Calculate Autocorrelation in Excel

- Autocorrelation is a **AR(p)** model, where **p** -> **lags**
  - t-1 -> lag=1 -> AR(1) model
  - t-2 -> lag=2 -> AR(2) model etc.

- It is the same as calculating the correlation between two different time series, except that the same time series is used twice: once in its original form and once lagged one or more time periods.

  e.g: Stock price of Day 15 depends on the price of Day 14, 13, 12 etc..and so on. Eventually , dependency will decrease with increase of lags

- The resulting output can range from +1 (positive correlation) to -1 (negative correlation)

- Autocorrelation measures linear relationships; even if the autocorrelation is miniscule, there may still be a nonlinear relationship between a time series and a lagged version of itself.

- Technical analysts can use autocorrelation to see how much of an impact past prices for a stock has on its future price

The following data represents the sales done (in lacs) for the given days.
Calculate the Auto Correlation

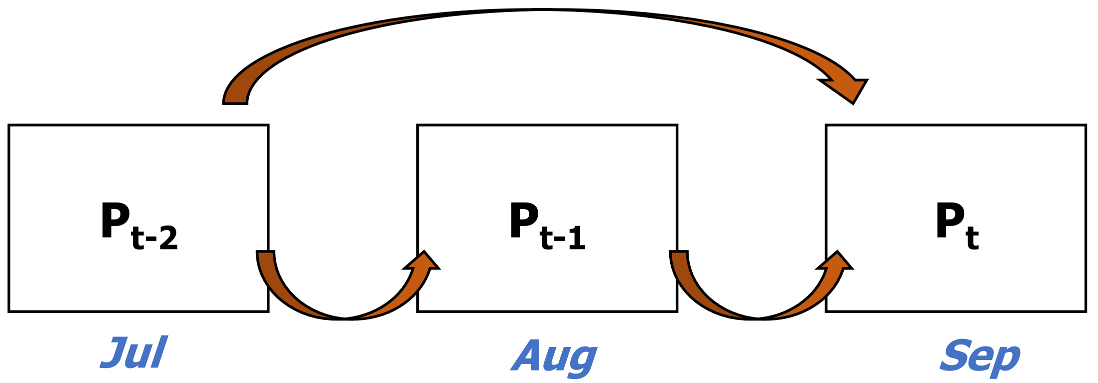
sales	
t	**Y_t**
1	10
2	20
3	24
4	30
5	40
6	50
7	60

Autocorrelation = $\dfrac{\Sigma[(y_t - \bar{y})(y_{t-k} - \bar{y})]}{\sum(y_t - \bar{y})^2}$

- **AutoCorrelation Function (ACF)**
- **Partial AutoCorrelation Function (PACF)**

Consider a situation where we need to predict the price of an item today as compared to the price last month or the month before or any prior months

P_t = price this month
P_{t-1} = price last month
P_{t-2} = price 2 months back

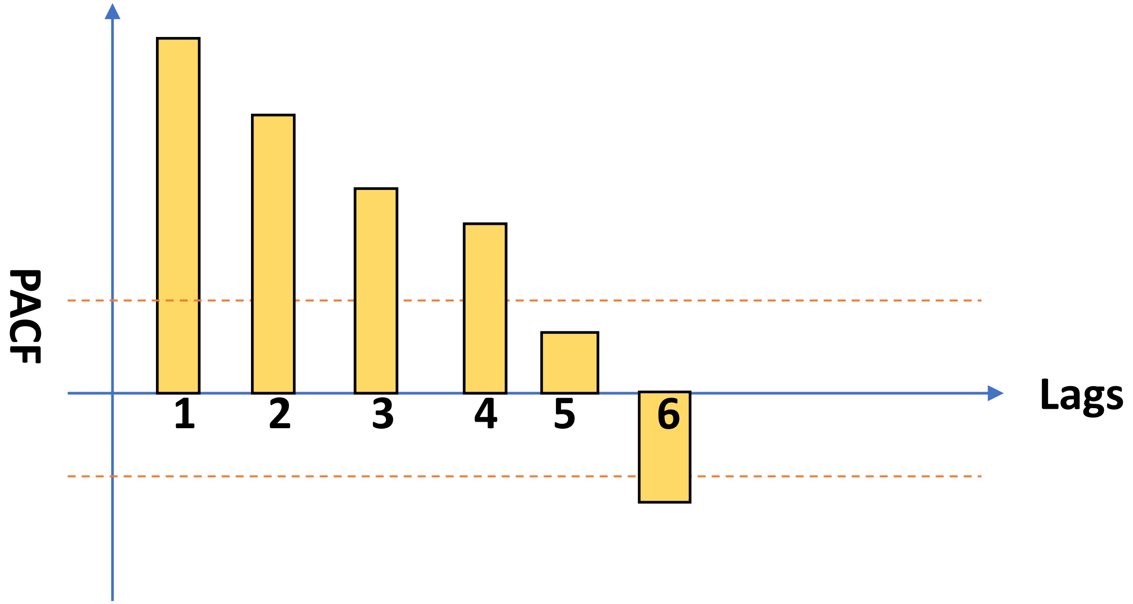


Partial AutoCorrelation Function (PACF)

- Direct effect of P_{t-2} and P_t without bothering about intermediate effects (other time periods)
- Formula(for n lags)
  $P_t = \beta_1 * P_{t-1} + \beta_2 * P_{t-2} + \ldots.\ \beta_n * P_{t-n} + \varepsilon$
- $\boldsymbol{\beta_n}$ gives the direct effect of the price now and the lagged time
- β is the PACF for the given lag
- PACF can be negative

Error Bands

- **Within the bands indicate 0 PACF**
- **Should cross the bands**



$$P_t = \beta_1 * P_{t-1} + \beta_2 * P_{t-2} + \beta_3 * P_{t-3} + \beta_4 * P_{t-4} + \beta_5 * P_{t-5} + \varepsilon$$

AutoCorrelation Function (ACF)

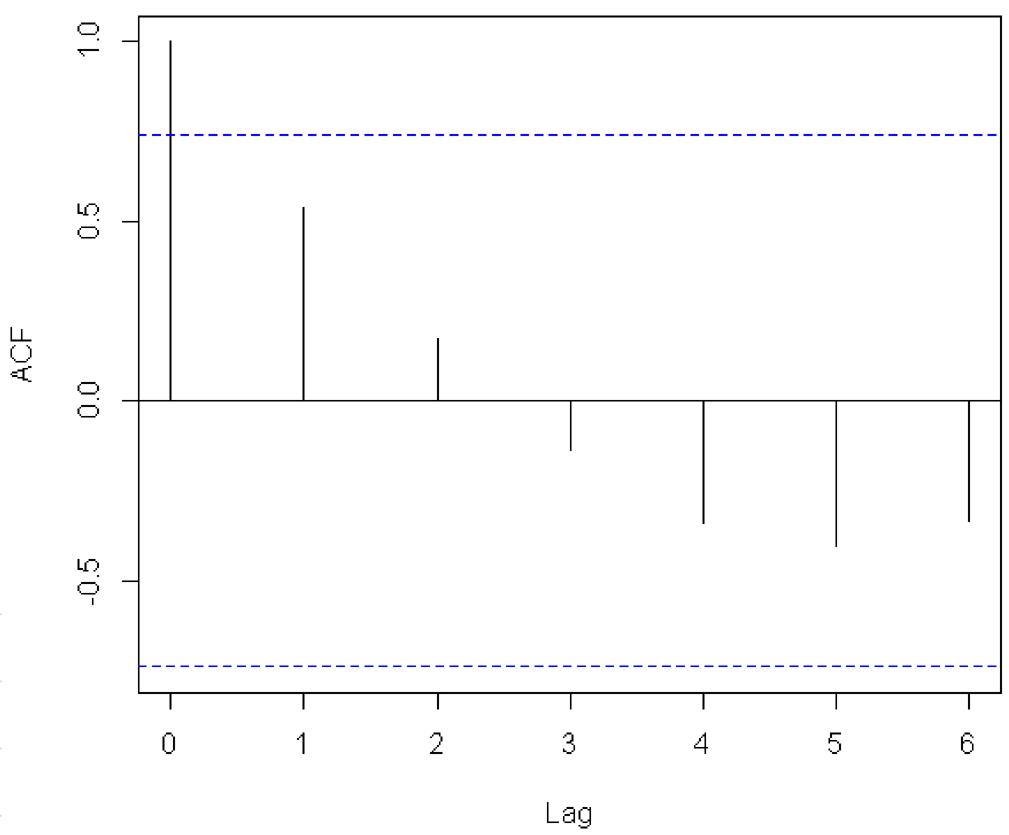
Sales	10	20	24	30	40	50	60

- Correlation between P_{t-2} and P_t
  - Direct effect (P_{t-2}->P_t)
  - Indirect effect (P_{t-2}->P_{t-1}->P_t)

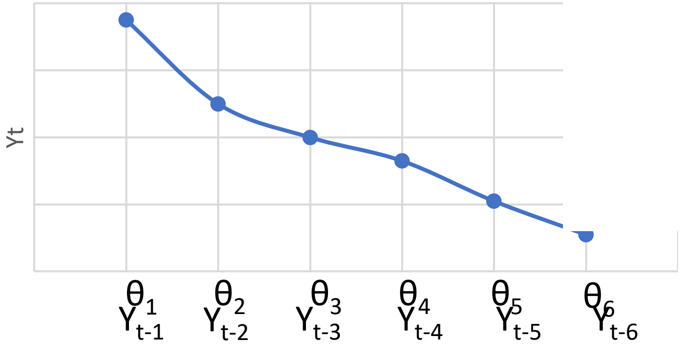
- Formula
  - Pearson's Correlation coefficient formula

- Correlation may be high due to these indirect effects

Correlogram




```
          [,1]
[1,]  1.0000000
[2,]  0.5395896
[3,]  0.1741237
[4,] -0.1377048
[5,] -0.3382508
[6,] -0.4019288
[7,] -0.3358288
```



Error Terms

A	P	Err	| Err |	APE	MAD	MAPE
100	105	-5	5	5.0	10.3	10.0
80	104	-24	24	30.0		
110	99	11	11	10.0		
115	101	14	14	12.2		
105	104	1	1	1.0		
110	104	6	6	5.5		
125	105	20	20	16.0		
120	109	11	11	9.2		
110	111	-1	1	0.9		

A	Actual Value
P	Predicted Value
Err (Error)	A - P
|Err|	Absolute Error
APE (Absolute Percent Error)	(|Err| / A) * 100
MAD (Mean Absolute Deviation)	Average |Err|
MAPE (Mean Absolute Percent Error)	Average APE

ARIMA model

- **ARIMA(p,d,q)** is AR and MA integrated where:
  - ✓ **p** → autoregressive lags
  - ✓ **q** → moving average lags
  - ✓ **d** → difference in the order

- **ARIMA** requires **Stationarity**

- **Seasonality** needs to be corrected before implementing ARIMA

$$y_t = \mu + \sum_{i=1}^{p} \theta_i y_{t-i} + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$$

ARIMA model implementation

- Read dataset

- Read the column that needs to be forecasted

- Convert Dataframe into a time-series object

- Check the **Stationarity** of data
  - Augmented Dickey-Fullter test determines stationarity
    - If p-value < 0.05 : Data is stationary
    - If p-value > 0.05 : Data is not stationary

- If Data is not stationary, difference the data and check for Stationarity on differenced data

- Use the optimum values for p,d,q to build the ARIMA model

✓ BIC (Bayesian Information Criterion)