Bandhan Bank

Date

Time Series Analysis

AR, MA, ARMA, ARIMA

Time series data :-
A time series data is a set of observation on the values that a variable takes at different times.
Such as monthly, daily, anually etc. They are used in Statistics, econometrics, mathematical finance etc.

Why do time series forecasting? How is it different from a multi-variate regression?

A Regression problem is of the form :-
$y = aX + b$, where X are the set of independent variables.

But say at times it may so happen what these X's are not available, then say only Y values are available to you. That is where TS Analysis comes, where we try to create a predictive model with the trend in output data over the time.

Mathematically, $y_t = \beta y_{t-1} + \varepsilon$.
where y_t is output today, y_{t-1} is output yesterday and ε is the Error, which means out put today depends on output yesterday.

Bandhan Bank

Date

Unlike a Regression, where output is dependent on it's independent variables or features.

The above eqn is that of univariate TS, where

$y_t = \beta \cdot y_{t-1} + \varepsilon$, — (1)

→ here our output of past day is being used as an independent variable.

It is univariate because only one value, that is the output of the past day is being taken into consideration. (only 1 factor)

A note: The time series data needs to be taken at regular intervals, that is $t_3 - t_2 = t_2 - t_1$, it cannot be irregular.

(2) Cross-sectional data: Such type of data is collected by observing many subject (such as individuals, firms, countries, or regions) at the same point of time or during the same time period.

For ex: for the year 2011 we take various aspects of a individual

	Salary	Height	Weight
2011	70000	4ft	40 kg

Bandhan Bank

Date

But in time-series we take data for a single variable at different time periods eg: Weather data.

2011 - 49° 2012 - 50° 2013 - 35° So on.

→ Patterns in Time Series :-
It can be random, constant, increasing, decreasing etc.

→ Components of Time Series :-
The patterns in a time series is sometimes classified into trend, seasonal, cyclical and random components.

- Trend :- A long term relatively smooth pattern that usually persists for more than one year

- Seasonal :- A pattern that appears in a regular interval wherein the frequency of occurence is within a year or even shorter. (Appears seasonally).

- Cyclical :- The repeated pattern that appears in a TS but beyond a frequency of one year. It is a wavelike pattern about a long term trend that is apparent over a number of years. Cycles are rarely regular and appear in combination with other components.

Bandhan Bank

Date

Ex: Business cycles that record periods of Economic recession and inflation, cycles in monetary and financial Sectors.

- Random - The component of TS that is obtained after these three components / patterns have been 'extracted' out of the series is the Random Component.

Therefore, when we plot the residual series then the scatter plot should be devoid of any pattern and would be indicating only a random pattern around a mean value.

Different TS Processes

1) White Noise :-

A series is called white noise if it is purely random in nature. Let $\{\varepsilon_t\}$ denote such a series then it has zero mean $[E(\varepsilon_t) = 0]$, has a constant Variance $[V(\varepsilon_t) = \sigma^2]$ and is an uncorrelated $[E(\varepsilon_t \varepsilon_s) = 0]$ random variable.

The scatter plot of such a series across time will indicate no pattern and hence forecasting the future values of such a series is not possible.

Bandhan Bank

Date

If a time series shows white noise we should stop doing time-series. Only best forecast for white noise TS is Average of all previous outputs.

2) Auto Regressive model :-

An AR model is one in which Y_t depends only on its own past values Y_{t-1}, Y_{t-2}, Y_{t-3}, etc.

$$Y_t = f(Y_{t-1}, Y_{t-2}, Y_{t-3}, \ldots, \varepsilon_t)$$

A common representation of an autoregressive model where it depends on p of its past values called as AR(p) model is represented below :

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \ldots \beta_p Y_{t-p} + \varepsilon_t$$

** AR(p) means AR depending on how many past values, p means 'p' no. of past values.

if AR(0), then $Y_t = \beta_0$, if AR(1) $= \beta_0 + \beta_1 Y_{t-1}$

3) Moving Average Model - is one when Y_t depends only on the random error term which follow a white noise process ie,

$$Y_t = f(\varepsilon_t, \varepsilon_{t-1} \ldots \varepsilon_{t-3} \ldots)$$

Bandhan Bank

Date

Say when we predict Y_t, with a past value Y_{t-1}, we get some error ε_1

$$\therefore Y_t = \beta_1 Y_{t-1} + \beta_0 + \varepsilon_1$$

When say we want Y_{t-1}, we use Y_{t-2} and some error ε_2.

$$\therefore Y_{t-1} = \beta Y_{t-2} + \beta_0' + \varepsilon_2.$$

So, here rather than using past output values as our input we take all past errors and use that as an input

A common representation of a MA model where it depends on 'q' of its values is called MA(q) model and is:

$$Y_t = \beta_0 + \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} \ldots + \phi_q \varepsilon_{t-q}$$

The Error terms ε_t are white noise with zero mean and variance σ^2 (constant).

(4) Auto Regressive Moving Average Model :-

There are situations where the time-series may be represented as a mix of both AR and MA models referred as ARMA (P,q)

The General form of such a TS model, which depends on 'p' of its past values and 'q' past values of white

Bandhan Bank

Date

noise disturbances, takes the form :-

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-3} + \dots + \beta_p Y_{t-p} + \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \phi_3 \varepsilon_{t-3} + \dots + \phi_p \varepsilon_{t-q}.$$

→ It's Actually (AR + MA) models taken together.

Stationarity of A Time Series

A Series is said to be "strictly stationary" if the marginal distribution of Y at time t $[p(Y_t)]$ is the same as at any other point in time.

$p(Y_t) = p(Y_{t+k})$ and $p(Y_t, Y_{t+k})$ does not depend on t.

(Here, $t \geq 1$ and k is any integer)

This implies that the mean, variance and co-variance of the series Y_t are time invariant.

Stationarity means the mean, variance at any point of time will be same, time invariant

Bandhan Bank

Date

An example of Marginal distribution :-

	Baseball	football	Total	
Male	13	20	33	Marginal dist of Gender.
Female	23	13	36	
	36	33		

Marginal dist of Sports

However, a series is said to be 'weakly' stationary or "covariance" stationary if the following conditions are met:

(a) $E(Y_1) = E(Y_2) = E(Y_3) \cdots E(Y_t) = \mu$ (constant)

(b) $Var(Y_1) = Var(Y_2) = Var(Y_3) = \ldots = Var(Y_t)$
$= Y_0$ (const)

(c) $Cov(Y_1, Y_{1+k}) = Cov(Y_2, Y_{2+k}) = Cov(Y_3, Y_{3+k})$
$= \gamma_k$, depends only on lag k.

For example, say t_5 and t_1 we consider so their covariance only depends on the difference of k value
$\therefore$ 5-1 = 4 (we are subtracting the two 'k' values).

Bandhan Bank

Date

A series which is non-stationary can be made stationary after differencing.

A series which is stationary after being differentiated once is said to be integrated of order 1 and is denoted by I(1)

In general a series which is stationary after being differentiated 'd' times is said to be integrated of order d, denoted I(d)

∴ a series which is stationary without differencing is said to be I(0).

→ When this integrated is included in ARMA, it becomes ARIMA.

→ Why only use Stationary Series

1) Econometrics theory derived under that Assumption.
2) Standard techniques invalid if non-stationary series.
3) May result in Auto-Correlation.
4) May result in spurious regression if not stationary, i.e two features showing some relation which actually should not.

Bandhan Bank

Date

Univariate TS Analysis is done using Box-Jenkins (B-J) Methodology, which has the following steps.

1) Identification
2) Estimation
3) Diagnostic Checking

→ The B-J method is applicable only to stationary variables.

* Identification :

a) Auto correlation function (ACF) - It refers to the way the observations in a TS are related to each other and is measured by a simple correlation between current obs (Y_t) and the observation p periods from the current one (Y_{t-p})

$$\rho_k = Corr(Y_t, Y_{t-p})$$

$$= \frac{Cov(Y_t, Y_{t-p})}{\sqrt{Var(Y_t)\, Var(Y_{t-p})}} = \frac{\gamma_p}{\gamma_0}$$

Bandhan Bank

Date

(b) Partial Auto-Correlation Function (PACF).

PACF are used to measure the degree of Association b/w Y_t and Y_{t-p} when the effects at other time lags 1, 2, 3 ... ... (p-1) are removed.

Inference of ACF, PACF : It helps us in understanding the correlation between the obs in a TS, which allows us to decide upto which lag we must consider that obs for modelling. For ex: the Stock price of today may be related to yesterday's value, day before yesterday and beyond that there is hardly any relation, ACF PACF helps in understanding those relations.

* Correlograms - A plot of ACF's vs lags.

MODEL	ACF	PACF
AR(p)	Spike decays toward 0	Spikes cutoff to 0.
MA(q)	Spike cutoff to zero.	Spikes decays toward 0.
ARMA(p,q)	Spike decays toward 0	Spikes decay toward 0.

→ Above table helps us in choosing right model (AR or MA or ARMA)

Bandhan Bank

Date

Steps in TS-Analysis :-

① Check for Stationary, if not make it Stationary.
② Using ACF, PCF find values of p and q
③ Find pattern for choosing model from table (in previous Pg).

(2) Several methods are available for estimating the parameters of an ARMA models depending on the Assumptions one make on the Error terms. They are -

a) Yule Walker procedure
b) Method of Moments.
c) Maximum likelihood.

3) Diagnostic Checking :- Different methods can be obtained for various combinations of AR and MA individually and collectively. The Bst model is obtained by following the diagnostic testing procedure.

The various test are :-

a) lowest value of AIC / BIC / SBIC - The Model with the lowest value of the above criterion is chosen as the best model.

(b) Plot of the residual ACF :- On fitting the appropriate ARIMA model, the goodness of fit can be estimated by plotting the ACF of residuals of the fitted models.

Bandhan Bank

Date

If the most of the sample auto-correlation co-efficients of the residuals lie within the limits.

$(-1.96/\sqrt{N}, +1.96/\sqrt{N})$, where N is the no. of observations,

then the residuals are white noise indicating that the model fit is appropriate.

** Errors must be random must not show any pattern.

$x = (3,1)$