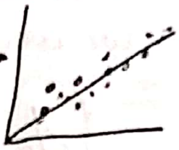
Time Series Analysis ? → definetly something that deals with the timeseries data (data at a Particular time)

i.e, data → every record contains timestamp → mostly this is sensors data

Time Series data

: A time series is a set of observations on values that a variable takes at different times → i.e, data collected at regular time intervals

Jumping back to the linear regression →



$y = mx + b$

y → dependent variable (the one that need to be predicted)

x → Independent variable

So in the single or multivariate linear regression we have some Independent variables & some factors (weights) associated with each independent variables that help us in forecasting future 'y'

But consider a dataset that looks something like this

timestamp	temp_soil	temp_air	soil moisture
28/6/18 08:48	32	42	35
-	-	-	-
-	-	-	-
-			

well damn! → we only have y values here at different 't'

we don't know what are those features or anything particularly about these variables.

→ what if we are asked to forecast future values at a different time intervals ?

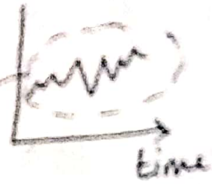
That is where forecasting using the time series analysis

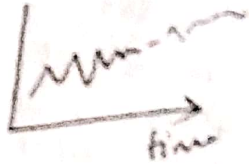
Applications:

- earth quake prediction
- weather forecasting
- stock price prediction

So how the time series analysis is done basically?

- Take the training data & Plot it onto a graph → • Then try to extrapolate the data that looks something like





Pattern

- In contrast with the linear regression we would have something like

$\boxed{y_t = y_{t-1}\beta + \text{error}}$ (AR model)

here we use only one variable and that is why it is called univariate time series
(no dependent & Independent variables literally)

- finally we forecast the values from the past data

Univariate time series

: It refers to a time series that consists of single observations recorded over regular time intervals
(one variable n intervals)

Cross-sectional data

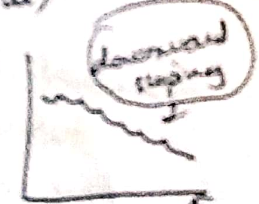
: This data is collected by observing many subjects (Persons, countries etc) at the same point of time (OR) during same time period

eg: population in 2019
(n variables 1 interval)

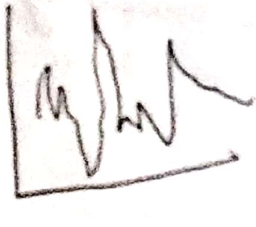
Patterns in Time series data

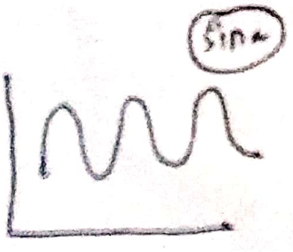
: depending upon the frequency of the data

- time series may increase or decrease over time with a constant slope



eg





Components of a time series

- Sometimes patterns in a time are classified into

trend

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

Seasonal

A pattern that appears in a regular interval

: frequency of occurrency is within a year
eg:- monthly arrival of tourists

Cyclical

repeated pattern that appears in a time series but beyond a frequency of one year

There are rarely regular & appear in combination with other components
eg: Inflation cycles

random → Irregular/residue

After all other three components are extracted the left overs

Modelling univariate Time series

(i) It is used when an appropriate relationship is not available, so we use past data to forecast the future data

(ii) It is used when we donot have sufficient variables to make a regression model

univariate time series Applications

- forecasting unemployment or inflation rate.
- demand of product.
- forecast interest
- forecasting prices

White noise ?

- A series is called white noise if it is purely random in nature → forecasting for such a series is not possible

→ let E_t be that series then

It will have zero mean $E(E_t)=0$

It will have a constant variance $[V(E_t)=\sigma^2]$

And will be uncorrelated $E\{E_t E_s\}=0$ random variable

To forecast white noise Take an average and thats the forecast all you can do.

Auto Regressive model

- AR model is one in which y_t depends only on its past values etc

$\therefore y_t = f(y_{t-1}, y_{t-2}, y_{t-3}, -- \epsilon_t)$

A common representation of AR model would be AR(P)

$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + - + \beta_p y_{t-p}$

and AR(P) means

AR model has P Parameter

eg: If $y_t = \beta_0$

$\therefore$ AR(0) model

If $y_t = y_{t-1}$

$\therefore$ AR(1) model

So Goal is finding $\boxed{P}$ value

ie, how many number of past values we must take

moving Average model

- A moving average model is one when Y_t depends only on the random error terms which follow a white noise process

ie, $y_t = f(\epsilon_t, \epsilon_{t-1}, \epsilon_{t-2} --)$

you first take $y_t = \beta_1 y_{t-1} + \beta_0 + \epsilon_1$ & regress it

ϵ_1 Then you will get an error

Similarly for $y_{t-1} = \beta_1 y_{t-2} + \beta + \epsilon_2$

ϵ_2

After regressing

A common representation is MA(q) q: past values of white noise distributions

$$y_t = \beta_0 + \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + - \phi_q \epsilon_{t-q}$$

And these ϵ_t: are assumed to be white noise i.e, Processes with Mean 0 & variance σ^2 (constant)

Autoregressive moving average model (ARMA)

It combines both past values & error terms

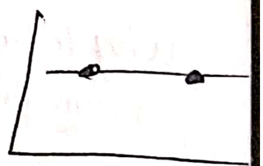
ie, ARMA (P, q)

$\therefore\ y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \beta_3 y_{t-3} + \beta_4 y_{t-4} + \cdots$

$\beta_p y_{t-p} + \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + \cdots \phi_q \epsilon_{t-q}$

Test

Stationarity of a time series (not changing with time)



- A series is said to be strictly Stationary if the marginal distribution of Y at time time period $t\,[P(Y_t)]$ is same as any other point in time

ie, $P(y_t) = P(y_{t+k})$

&

$P(y_t, y_{t+k})$ does not depend on t

Note:

- A series is said to be weakly stationary or covariance stationary

that means mean, variance, & covariance of series y_t are time invarient.

then

(i) $E(y_1) = E(y_2) = E(y_3) \ldots = E(y_t) = \mu$ (constant)

(ii) $Var(y_1) = Var(y_2) = \cdots = Var(y_t) = \gamma_0$ (constant)

(iii) $cov(y_1, y_{1+k}) = cov(y_2, y_{2+k}) = cov(y_3, y_{3+k}) = \gamma_k$, depends on k only

Make a non stationary series Stationary through differencing

- A Series that is made stationary after making it differentiated once is said to be integrated of order 1 & 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 $I(d)$

And a series which is stationary & not differentiated then it is $I(0)$

ARIMA model → another kind of model which Speaker about

I → Integrated

$$ARIMA(p, 1, q)$$

p → number of Past values

1 → no. of / order of Integration

q → moving average

Why to take the stationary assumption

- Standard techniques are invalid if the data is non stationary
- Regression may not also work good on non stationary data

Box-Jenkins (B-J) methodology

- The estimation & forecasting of univariate time series models is carried out using BJ method
- which contains 3 steps
  1. Identification
  2. estimation
  3. Diagnostic checking
- This methodology is only applicable to the stationary variables

Identification

a) Auto correlation function (ACF)

- It refers to a way the observations are related to each other
- It is measured by a simple correlation between current one (y_t) & the observations 'p' periods from current one (y_{t-p})

$$\therefore \boxed{\rho_k = Corr(y_t, y_{t-p}) = \frac{Cov(y_t, y_{t-p})}{\sqrt{Var(y_t)}\sqrt{Var(y_{t-p})}} = \frac{\gamma_p}{\gamma_0}}$$

using this we can compute the Parameter 'p' → number of lags (y_t corr y_{t-1} corr —)

b) Partial Autocorrelation function PACF

- It is used to measure the degree of association between y_t & y_{t-p} when the effects at other time lags 1, 2, 3, --- (p-1) are removed.

eg: $y_t = y_5$ $y_{t-p} = y_1$

then we remove y_4, y_3, y_2 between them

- It is used to remove the intermediate lags

c) Inference from ACF & PACF :-

- A comparision between sample ACF's v/s Lags (correlograms) will help us choose the appropriate ARIMA(p, q) model

Theoretical charecteristics of ACF's & PACF: / Normal Patterns

Model	ACF	PACF
AR(p)	spikes decay towards zero	spikes cutoff to zero
MA(q)	spikes cutoff to zero	spikes decay towards zero
ARMA(p,q)	spikes decay towards zero	spikes decay towards zero

Modelling in a glance

- check stationarity of time series - If non stationary, transform to stationary

↓

- find initial values of 'p' & 'q' by Autocorrelation & Partial autocorrelation coefficients

↓

- estimation

↓

- diagnostic checking

Estimation

- we can estimate the parameters of the ARMA model depending upon the assumptions one makes on the error terms.

a) Yule walker procedure

b) Method of movements

c) Maximum likelihood method

→ For this we must seek computer aided methods & statistical packages.

Diagnostic checking

a) lowest value of AIC/BIC/SBIC

- The lowest value model is choosen as the best model.

b) Plot of residual ACF :-

- on fitting an ARIMA model, the goodness of fit can be estimated by plotting ACF of residuals of the fitted model.

- If most of sample autocorrelation coefficients of the residuals lie within the limits

$$\left(-1.96/\sqrt{N}, +1.96\sqrt{N}\right)$$

$N \rightarrow$ number of observations.

then it is clear that the residuals are white noise Indicating that the model fit is appropriate.