Module 4.

Removing trend -
A trend is a continued increase or decrease in the time series overtime.
A systematic change in the time series that does not appear to be periodic is known as trend.

Types of trends :-
1) Deterministic trend -
Trend that continuously increase or decrease.
2) Stochastic trend -
Trends that increase or decrease inconsistently.

Identifying the trend:-
1) Plot a time series data using HP filter.
2) Plot a time series data using the method of least squares.

Removing trend
- A time series with a trend is called non-stationary.
- A trend can be modelled and removed from the time series data by detrending a time series.
- If a dataset does not have a trend or the trend was successfully removed, then the data can be called as stationary.

Following are the methods to remove/detrend a trend -

1) Detrend by Differencing –
A new series is constructed where the value of the current time step is calculated as the difference between the original observation & the observation of the previous time step.
value (t) = observation (t) – Observation (t-1)

2) Detrend by model fitting :–
Linear trends can be summarized by a linear model and non-linear by using polynomial or any other curve fitting method.

3) Log Transformation –
Apply log transformation on each ~~individual~~ individual value of TS data. This reduces the Variance of TS data in most case. eg., np.log(x)

4) Power Transformation –
Apply power transformation on each individual value of TS data. eg., x**0.5

5) Applying moving window function on log transformed TS –
We can apply more than 1 transformation as, apply log transformation to TS, then take rolling mean over a period of 12 months and then subtract rolled TS from log-transformed TS to get final TS.

5) Applying Moving window function -
Calculate rolling mean over a period of 12 months and subtract it from original data.
eg., rolling (window = 12). mean() .

Unit Roots → characteristics of a

In statistics, a unit root tests whether a time series variable is non-stationary and possesses a unit root.
The null hypothesis is generally defined as the presence of the unit root & the alternative hypothesis tests whether the data is stationary or not.

Dickey Fuller test tests the null hypothesis that a unit root is present in an AR time series model.
The augmented Dickey Fuller test is an extension of the DF test that removes autocorrelation from the series & then tests similar procedure of DF test.
ADF test is a statistical significance test which which means that the test will give results in hypothesis, that is null & alternate.
As a result a p-value will be generated which will be needed to make an inference about the stationarity of the time series data.
Null hypothesis - presence / absence of unit roots.
Alternate hypothesis - whether the data is stationary or not.

To perform the ADF test on any TS, the states model provide the implementation called as ~~ad~~ adfuller().

It provides the following information -

① p-value
② value of the test statistic
③ No. of lags used for testing
④ Critical value such as 1%, 5% & 10%.

| | | |
|---|---|---|
| 1% | :- | -3.435 |
| 5% | :- | -2.863 |
| 10% | :- | -2.567 |

conditions to reject null hypothesis :-

If test statistic < critical value
& p-value < 0.05
then, reject the null hypothesis i.e. TS does not have a unit root, meaning TS is stationary
else otherwise.

(5 mks. compulsory comes)

Q1. Determine from the below observation if the TS data is stationary or not using ADF test & give proper justification.

| | |
|---|---|
| Test statistic | -1.3380 |
| p-value | 0.6115 |
| no. of obs. used | 1258 |
| critical value (1%) | -3.435 |
| critical value (5%) | -2.863 |
| critical value (10%) | -2.567 |

Failed to reject the null hypothesis
Data is non-stationary i.e. it contains unit roots
Justification:
As the test statistic is ~~low~~ greater than critical value and p-value is greater than 0.05.
Therefore, the data is non-stationary using ADF test.
To make the above data stationary, differencing technique can be used.

Q2. Determine from the below observation & find the following.
1. Complete the missing values of critical values.
2. If the time series data contains unit roots or no.
3. Mention the statistical test used to check for unit roots.
4. Provide proper justification.

| | |
|---|---|
| Test statistics. | −36.114 |
| p-value | 0.01 |
| critical value (1%) | — |
| critical value (5%) | −2.863 |
| critical value (10%) | — |
| no. of observation | 1259 |

Solution:

1. critical value (1%) −3.435
   critical value (10%) −2.567

Conditions

2. Reject the null hypothesis.
   Therefore, the data is stationary and it contains no unit roots

3. ADF test is used.

4. The test statistic is less than ~~both p-value~~ ~~and~~ critical value and p-value is less than 0.05.

| 79, 62, 61 |

9/10

give examples
transport
cost of production
sales

## Intervention analysis -

Intervention means a change to a procedure i.e. intended to change the value of the series $x_t$. ~~sup~~

Suppose that at time t, there has been an intervention to a time series, then we need to estimate how much the intervention has changed the series (if at all).

Example, suppose that a region has instituted a new maximum speed limit on it's highway and want to learn, how much the new limit has affected accident rates.

Intervention analysis in time series, refers to the analysis of how the mean level of a series changes after an intervention, when it is assumed that the same ARIMA structure for the series $x_t$ holds both before and after the intervention.

The intervention model - suppose that the ARIMA model for $x_t$ (the observed series) with no intervention is as follows,

$$x_t - \mu = \frac{\Theta(B)}{\Phi(B)} w_t$$

where, $\Theta(B)$ is usual MA model.
& $\Phi(B)$ is usual AR model.
& the assumtion about the error i.e. $w_t$.

Let Z be the amount of change at a time t, i.e. attributable to the intervention.

By definition, $Z_t = 0$ before time $t$ (time of the intervention.

The value of $Z_t$ may or may not be 0 after time $t$.

The overall model including the intervention effect may be written as

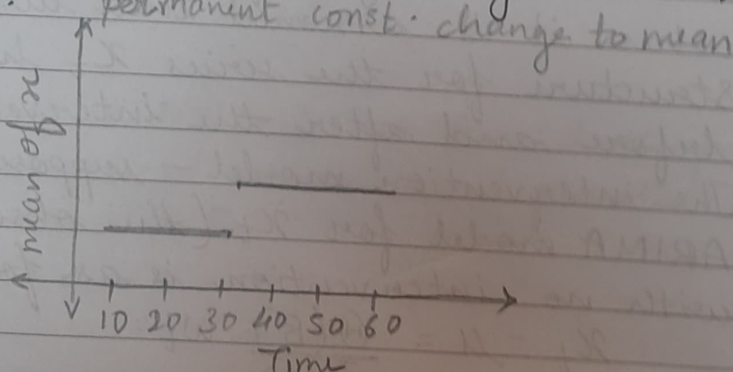
$$x_t - \mu = Z_t + \frac{\Theta(B)}{\Phi(B)} w_t$$

imp. ** Possible patterns for intervention analyis / effect

There are several possible patterns on how an intervention may affect the values of the series.

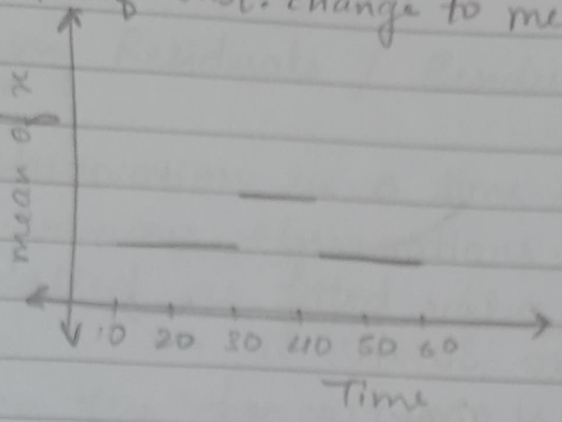
Four possible patterns are as follows —

1. Permanent constant change to the mean level.



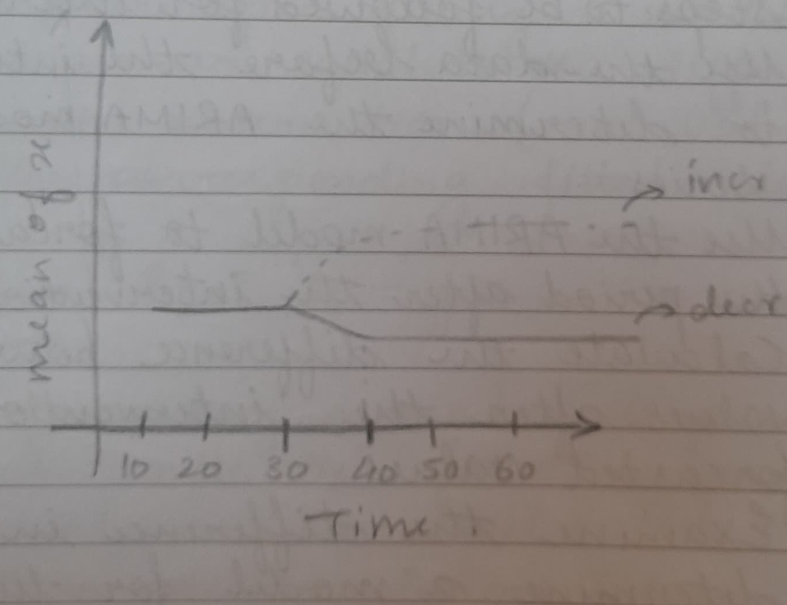

2. Brief constant change to the mean level. There may be temporary change for one or more period after which there is no effect of the intervention.

brief const. change to mean.

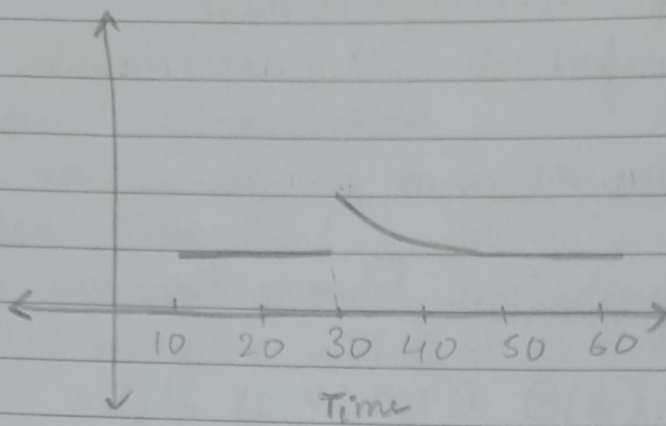



3. Gradual increase or decrease to a new mean level.

There may be a gradually increasing amount that is added which eventually levels off at a new level compared to before the intervention.




4. Initial change followed by gradual return to no change.




## Estimating the intervention effect.

Two parts of the overall model have to be estimated.

1. The basic ARIMA model for the series.
2. The intervention effect.

Steps to be followed for intervention analysis.

1. Use the data before the intervention point to determine the ARIMA model for the series.
2. Use the ARIMA model to forcast values for the period after the intervention.
3. Calculate the difference between actual values after the intervention & the forcasted values.
4. Examine the difference in step 3, to determine a model for the intervention effect.

classmate
Date
Page

# Regression Residuals / Residual Diagnostics.

Each observation in a time series can be forcasted using previous observations. These observations are called as fitted value. Denoted by $\hat{y}_{t|t-1}$.

Meaning, the forcast of $y_t$ based on the observations $y_1 \ldots y_{t-1}$ and also written as $\hat{y}_t$

$y_t$ = real values
$\hat{y}_t$ = fitted values.

The residuals in a time series model are what is left over after fitting a model.
The residuals are equal to the difference between the observations (real value) and the corresponding fitted values.

Residual = $\varepsilon_t = y_t - \hat{y}_t$

A good forcasting method will provide residuals with the following properties.

1. Residuals are uncorrelated.
   If there are correlations between residuals, then there is information left in the residuals which should be used in computing forcast.
2. The residuals have 0 mean -
   If the residuals have mean other than 0, then the forcast is either under or

Date
Page

over predicted.
3. The residuals have constant variance.
4. The residuals are normally distributed.

30/10

Monte-Carlo method. (used to analyse complex systems) that involve randomness & uncertainity) volatility

A monte-carlo simulation / method / model is used to predict the probability of a variety of outcomes when the potential for random variables is present.
It is also referred as a multiple probability simulation.
A monte-carlo method requires assigning multiple values to an uncertain variable to achieve the results to obtain an estimate. The probability of varying outcomes cannot be firmly pinpointed because of random variable inference. Therefore, a monte-carlo method focuses on constantly repeating random samples.

Step by Step process of how monte-carlo simulation is applied in time series analysis.
1. Data Generation :- collect / obtain historical time series data that needs to be used as a basis for simulation.
2. Model Selection - choose an appropriate time series model such as ARIMA, GARCH,

or any other stochastic model.

3. Parameter estimation - selecting appropriate model parameter such as p, d, q, etc.
4. Mote-carlo simulation process -
   a) Generate multiple random simulation using the estimated model & parameters.
   b) Vary the key factors or assumptions in each simulation as need to explore different scenarios.
   c) Repeat the simulation process multiple times to create a distribution of possible outcomes.
5. Analysis - Draw conclusion, make forcasts, assess the risk or evaluate the impact of different scenarios.

Autoregressive distributed lag (ADL) model.

1. The ADL model in time series analysis is used to analyse, the long run relationship and short run dynamics between variables in a time series data set.
2. It is particularly useful for examining the relationship between a dependent variable and one or more independent variables in a dynamic context.
3. The ADL model combines elements of autoregressive (AR) and Distributed Lag (DL) as follows,
   a) Autoregressive component - It captures the short term dynamics of the dependent variable (i.e., current value dependent on past values).

b) Distributed lag component - It accounts for the lag effect of the independent variable on dependent variable (i.e. it allows you to model how the change on independent variable impact the dependent variable over time)

4. The general equation / ADL model can be written as (AR)

$$y_t = \alpha + \beta_0 x_t + \beta_1 x_{t-1} + \dots + \beta_p x_{t-p}$$
$$\text{dep } + \gamma y_{t-1} + \delta_1 y_{t-1} + \delta_2 y_{t-2} + \dots$$
$$+ \delta_q y_{t-q} + \varepsilon_t \quad \text{DL}$$

where, $y_t$ is dependent variable at time $t$

$x_t, x_{t-1}, \dots x_{t-p}$ is lagged values of the independent variable.

$\alpha$ is mean

$\beta_0, \beta_1, \dots \beta_p$ are coefficients of the independent variables at different lags

$\gamma$ coefficient of ~~lag~~ the dependent variable at different lags.

$\delta_1, \delta_2, \dots \delta_q$ are coefficients of lagged values of dependent variable.

$\varepsilon_t$ is error

Qus on convertion from ADL to AR & vice versa

Transfer functions (studying cause and effect of the relationship between variables) (Multivariate data)

only theory

A transfer function model is a mathematical model used to determine the relationship between two or more time series variables.

It is a common tool to analyse ~~&~~ model, how one time series affects / influences other.

The transfer function can be represented as -

$$Y(t) = H(B) * X(t) + E(t)$$

where, $Y(t)$ is dependent time series which needs to be predicted.
$X(t)$ independent time series (series that influences / causes changes in $Y$)
$H(B)$ is the transfer function which is typically a function of lag operator $B$ specifying how $X$ affects $Y$ over time.
$E(t)$ is error term.

- Financial Markets:

Stock prices → Dependent.
Interest rates → Independent

$$Y_t = \frac{B(L)}{A(L)} X_t + \frac{C(L)}{D(L)} \varepsilon_t$$

# Vector Autoregressive Model (VAR) :-

① 
$$\begin{tikzcd} A_t \arrow[r, leftrightarrow] & B_t \\ A_{t-1} \arrow[u, bend left] \arrow[ur] & B_{t-1} \arrow[u, bend right] \arrow[ul] \end{tikzcd}$$

② 
$$A_t = \mu + \overset{c_{11}}{\wedge} A_{t-1} + \overset{c_{12}}{\wedge} B_{t-1} + \varepsilon_t$$

$$B_t = \mu + c_{21}\overset{A}{B_{t-1}} + c_{22}\overset{B}{A_{t-1}} + \varepsilon_t$$

A → No. of Apples sold.
B → No. of Banana's sold

$$\begin{bmatrix} A_t \\ B_t \end{bmatrix} = \mu + \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} \begin{bmatrix} A_{t-1} \\ B_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{A,t} \\ \varepsilon_{B,t} \end{bmatrix}$$

one time period from past of fruit vector

vector multiplication

$$F_t = \mu + C \cdot F_{t-1} + \varepsilon_t$$

component variables

How many apples & bananas are we going to sell in a given month?
2 variables $A_{t-1}$ & $B_{t-1}$ are interactive which will help to predict the value of $A_t$ & $B_t$.
① ← where, $A_t$ is the apple sales this month, $A_{t-1}$ is apple sales last month. Similarly for $B_t$ & $B_{t-1}$.

② ← The vector method is used for a compact model, if the time period increases.
where,

No of eqns = no. of variables

VAR is a multivariate forcasting algorithm that is used when 2 or more time series influence each other.

The basic requirements in order to use VAR model are -
1) Need atleast 2 time series variable.
2) The time series should influence each other.

Why is it called Autoregressive?
Each variable is modelled as a function of the past values, which are lags of the series.

How VAR is different from AR, ARMA or ARIMA?
The primary difference in those models are unidirectional where the predictors influence the value of y and not vice versa. Whereas, VAR is bidirectional and the variables influence each other.

In VAR model, each variable is modelled as a linear combination of past values of itself & the past values of other variables in the system.

As we consider only 1 time period in the past, therefore the above equation is a VAR(1) model.

no. of lags      no. of variables

Apply a VAR(2) model with 2 coefficients influencing each other.

$$A_t = \mu + C_{11} A_{t-1} + C_{12} B_{t-1} + C_{13} A_{t-2} + C_{14} B_{t-2} + \varepsilon_{A,t}$$

$$B_t = \mu + C_{21} A_{t-1} + C_{22} B_{t-1} + C_{23} A_{t-2} + C_{24} B_{t-2} + \varepsilon_{B,t}$$

$$\begin{bmatrix} A_t \\ B_t \end{bmatrix} = \mu + \begin{bmatrix} C_{11} & C_{12} & C_{13} & C_{14} \\ C_{21} & C_{22} & C_{23} & C_{24} \end{bmatrix} \begin{bmatrix} A_{t-1} \\ B_{t-1} \\ A_{t-2} \\ B_{t-2} \end{bmatrix} + \begin{bmatrix} \varepsilon_{A,t} \\ \varepsilon_{B,t} \end{bmatrix}$$