Data Analysis Course

Time Series Analysis & Forecasting(Version-1)

Venkat Reddy

Data Analysis Course

- Data analysis design document
- Introduction to statistical data analysis
- Descriptive statistics
- Data exploration, validation & sanitization
- Probability distributions examples and applications
- Simple correlation and regression analysis
- Multiple liner regression analysis
- Logistic regression analysis
- Testing of hypothesis
- Clustering and Decision trees
- Time series analysis and forecasting
- Credit Risk Model building-1
- Credit Risk Model building-2

Note

- This presentation is just class notes. The course notes for Data Analysis Training is by written by me, as an aid for myself.
- The best way to treat this is as a high-level summary; the actual session went more in depth and contained other information.
- Most of this material was written as informal notes, not intended for publication
- Please send questions/comments/corrections to venkat@trenwiseanalytics.com or 21.venkat@gmail.com
- Please check my website for latest version of this document

-Venkat Reddy

Contents

- What is a Time Series
- Applications of Time Series Analysis
- Time series model building & Forecasting Methodologies
  - TCSI Method
    - Components of time series
    - Goodness of fit
    - Forecasting using TCSI model
  - ARIMA
    - Main steps in ARIMA
    - Goodness of fit
    - Forecasting using ARIMA model

What is a Time Series

- Time series data is a sequence of observations collected from a process with ***equally*** spaced periods of time.
- Examples
  - Dow Jones Industrial Averages
  - Daily data on sales
  - Monthly inventory
  - Daily Customers ,
  - Monthly interest rates, costs
  - Monthly unemployment rates,
  - Weekly measures of money supply,
  - Daily closing prices of stock indices, and so on

Forecasting Methodologies

- There are many different time series techniques.
- It is usually impossible to know which technique will be best for a particular data set.
- It is customary to try out several different techniques and select the one that seems to work best.
- To be an effective time series modeler, you need to keep several time series techniques in your "tool box."
- Simple ideas
  - Moving averages
  - TSI method
- Complex statistical concepts
  - Box-Jenkins methodology

Building Model Using TSI Method

- Very Simple technique
- Spreadsheet is sufficient
- Less ambiguous
- Easy to understand & Interpret
- Serves the purpose most of the times

Components of a Time series

1. **Secular Trend(T):** Gradual long term movement(up or down). Easiest to detect
   - Eg: Population growth In India
2. **Cyclical Patterns(C):** Results from events recurrent but not periodic in nature. An up-and-down repetitive movement in demand. repeats itself over a long period of time
   - Eg. Recession in US Economy
3. **Seasonal Pattern(S):** Results from events that are periodic and recurrent in nature. An up-and-down repetitive movement within a trend occurring periodically. Often weather related but could be daily or weekly occurrence
   - Eg. Sales in festive seasons
4. **Irregular Component(I):** Disturbances or residual variation that remain after all the other behaviors have been accounted for. Erratic movements that are not predictable because they do not follow a pattern
   - Eg. Earthquake

Building Time Series Model: TCSI

$O(t) = T(t) + S(t) + I(t)$ or $O(t) = T(t) * S(t) * I(t)$

Where O(t) = observed series, T(t) = Trend component, S(t) = Seasonal , I(t) = Irregular component

- **Step 1 :** Smooth the series & de trend the series

$$\frac{O_t}{\hat{T}_t} = \frac{T_t \times S_t \times I_t}{\hat{T}_t} \approx S_t \times I_t$$

- **Step 2 :** Find out Seasonal component and adjust the data for seasonality

$$\frac{O_t}{\hat{S}_t} = \frac{T_t \times S_t \times I_t}{\hat{S}_t} \approx T_t \times I_t$$

- **Step 3 :** See if there is still some trend/seasinality in the data & quantify it

$$\frac{T_t \times I_t}{\hat{T}_t} \approx I_t$$

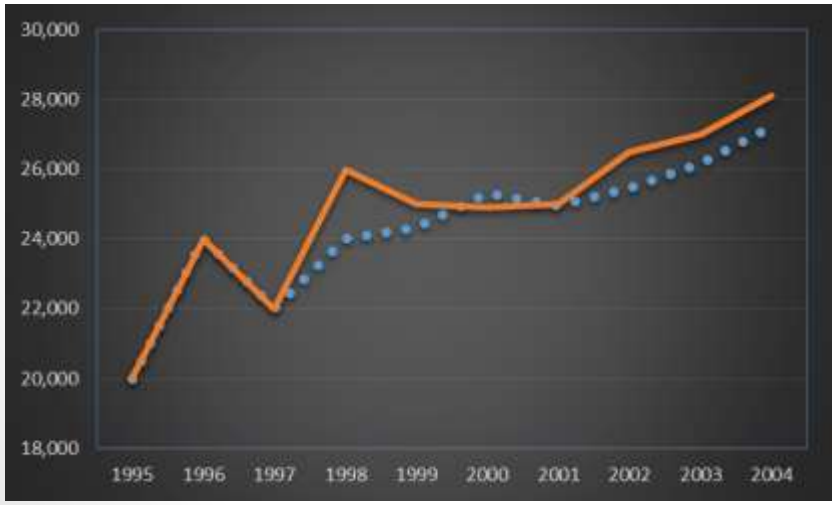
Step-1: Smoothening the series

- Why Smoothening ? Basically to find average value and detrend the series
- After removing the effect of trends series is left with seasonal and irregular components

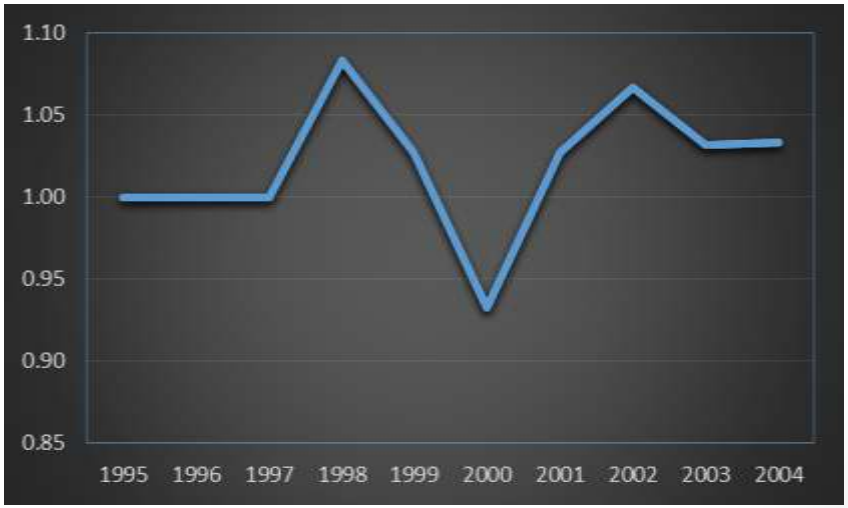
Moving Average

- Series of arithmetic means, Used only for smoothing. Provides overall impression of data over time
- MA(3)= $(y_t + y_{t-1} + y_{t-2})/3$

Actual Series



Series after Removing the Trend



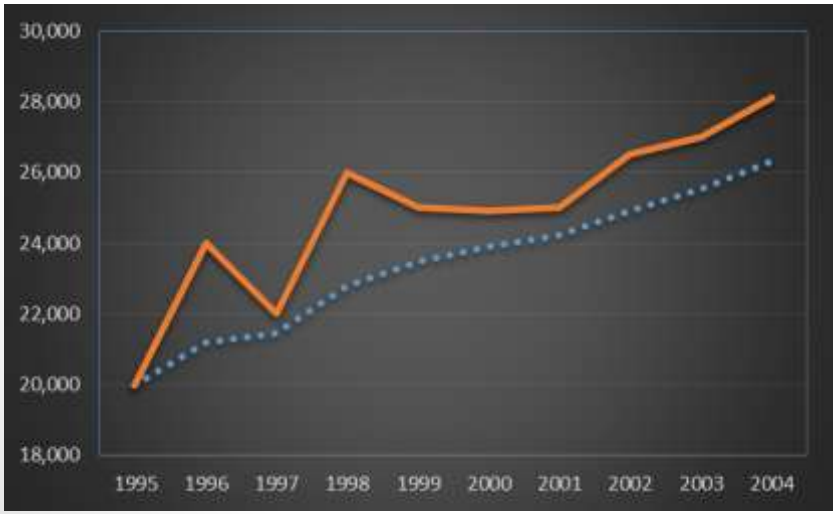
Smoothening the series

Exponential Smoothing

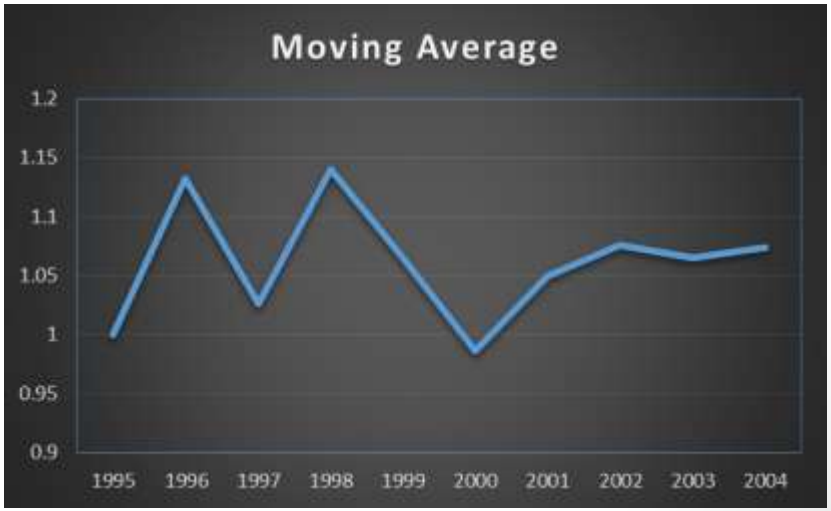
- **Form of weighted moving average.** Weights decline exponentially and most recent data weighted most
- **Requires smoothing constant (*W*).** Ranges from 0 to 1, Subjectively chosen
- Involves little record keeping of past data

$$E_i = W \cdot Y_i + (1 - W) \cdot E_{i-1}$$

Actual Series

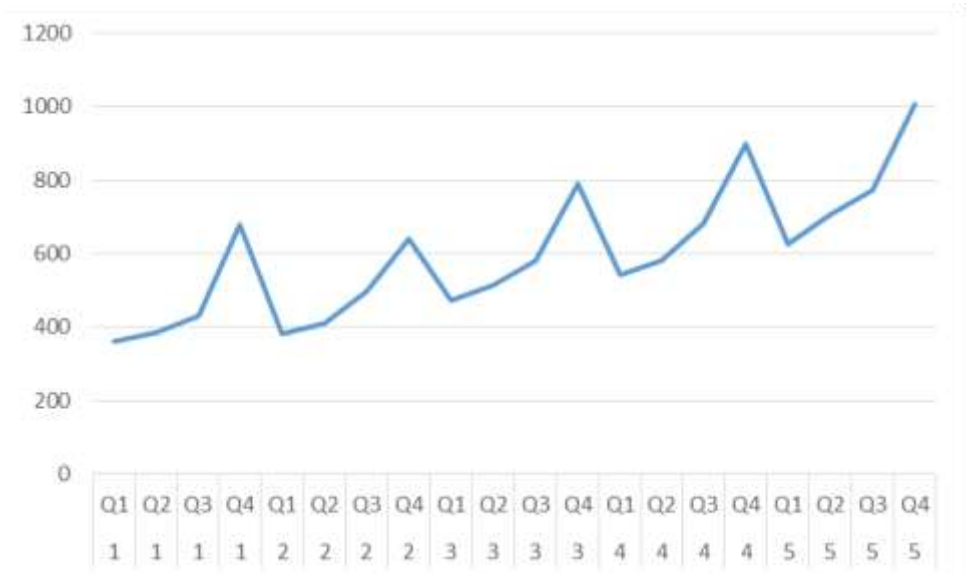


Series after Removing the Trend



Step-2 : Capturing Seasonality - Seasonal indices

Year	Quarter	Sales
1	Q1	362
1	Q2	385
1	Q3	432
1	Q4	678
2	Q1	382
2	Q2	409
2	Q3	498
2	Q4	642
3	Q1	473
3	Q2	513
3	Q3	582
3	Q4	789
4	Q1	544
4	Q2	582
4	Q3	681
4	Q4	899
5	Q1	628
5	Q2	707
5	Q3	773
5	Q4	1008



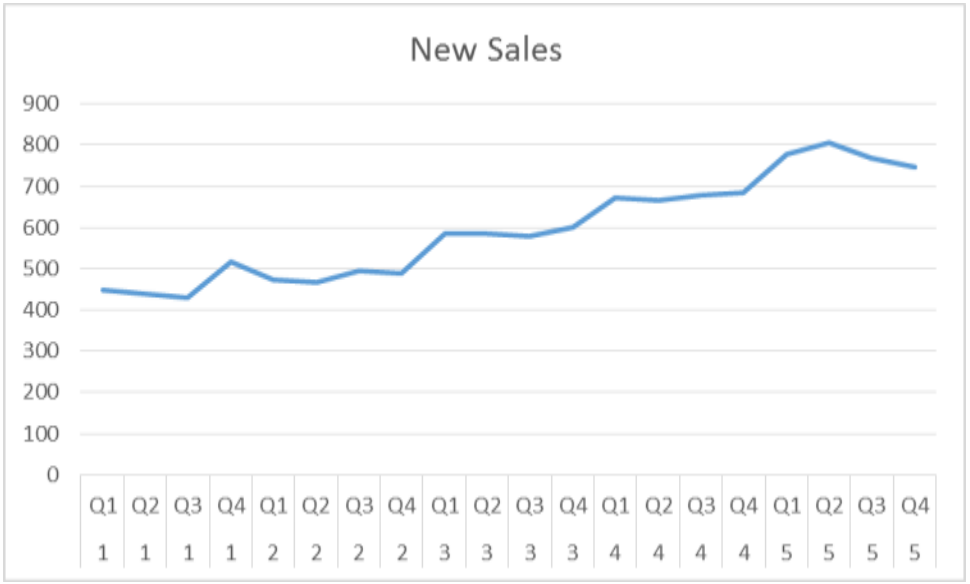
- Quantify the effect of seasonality in each year
- Calculate overall average for each quarter for across all years
- De seasionalize the dataset with by diving with seasonal indices

Seasonal Indices & De Seasionalising

Step 1 : Reformating the Dataset				
	Qtr 1	Qtr 2	Qtr 3	Qtr 4
Y1	362	385	432	678
Y2	382	409	498	642
Y3	473	513	582	789
Y4	544	582	681	899
Y5	628	707	773	1008

Step 2 : Calculation Of Seasonal Indices				
	Qtr 1	Qtr 2	Qtr 3	Qtr 4
Y1	0.78	0.83	0.93	1.46
Y2	0.79	0.85	1.03	1.33
Y3	0.80	0.87	0.99	1.34
Y4	0.80	0.86	1.01	1.33
Y5	0.81	0.91	0.99	1.29
SI	0.80	0.86	0.99	1.35

Step 3: Deseanolised Dataset				
	Qtr 1	Qtr 2	Qtr 3	Qtr 4
Y1	454	446	436	502
Y2	479	474	503	475
Y3	594	594	588	584
Y4	683	674	688	666
Y5	788	819	781	746



Step-3: Irregular Component

- Irregular→ No Trend → Cant be modeled
- After first level of de trending & de seasionalising, if there is still any pattern left in the time series, repeat the above steps once again

Lab

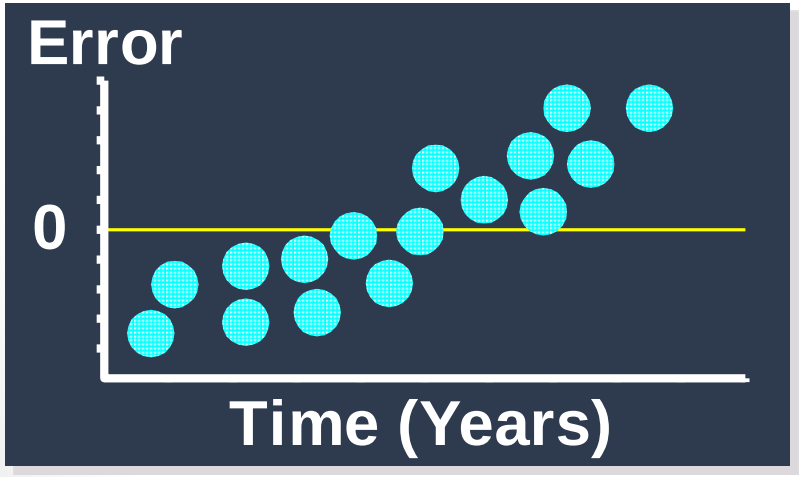
- Download the data from here 

- Draw the trend chart, can you see trend
- Calculate MA2 and MA3
- De-trend the series and create the graph again
- Is there any trend now?
- Can you judge the seasonality in de trended data?
- Find seasonality indices for each quarter
- De-seasinalise the series by diving with seasonality indices
- Forecast sales for next four months, one by one
  - First assume I as 1, multiply with S, multiply with MA

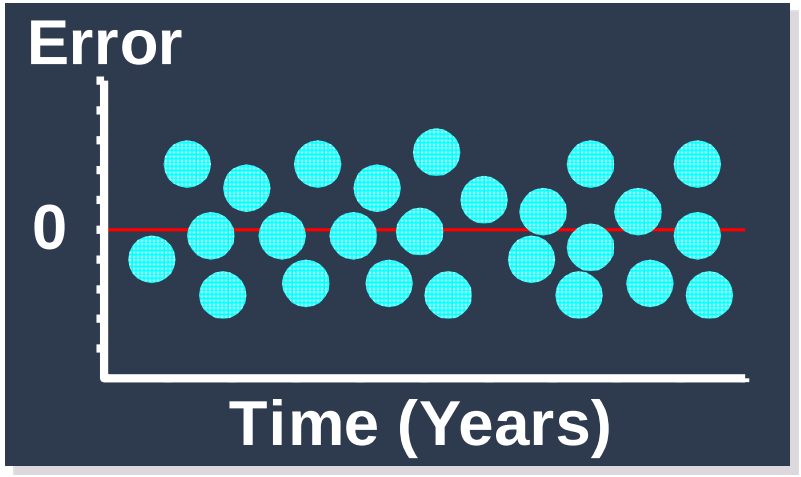
Residual Analysis

- Residual analysis is done after the model is built
- There should be no Pattern (information) left in the residuals.
- Residual should be totally random

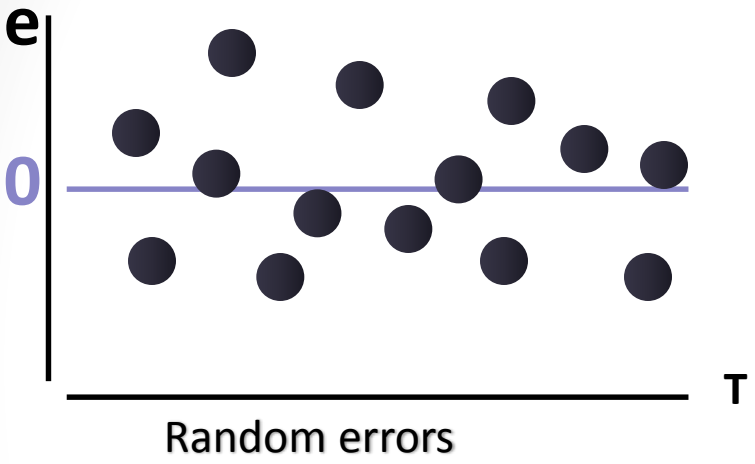
Trend Not Fully Accounted for



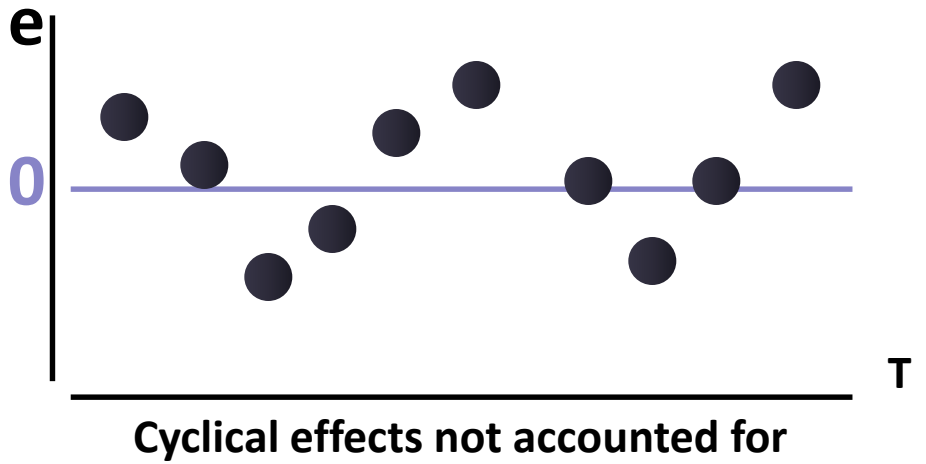
Desired Pattern



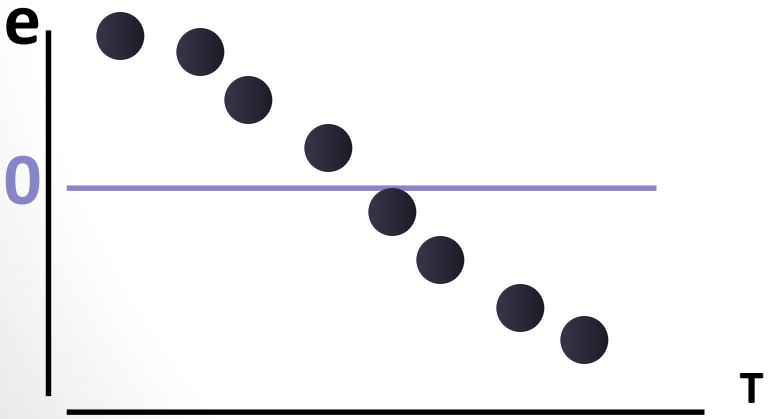
Residual Analysis



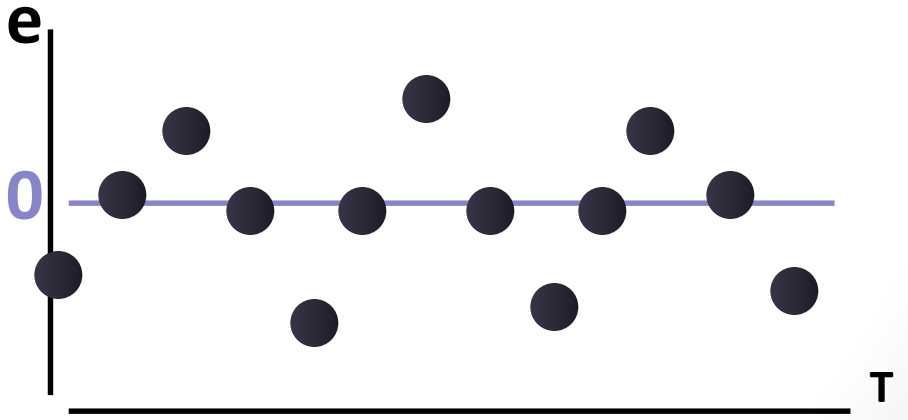
Random errors



Cyclical effects not accounted for



Trend not accounted for



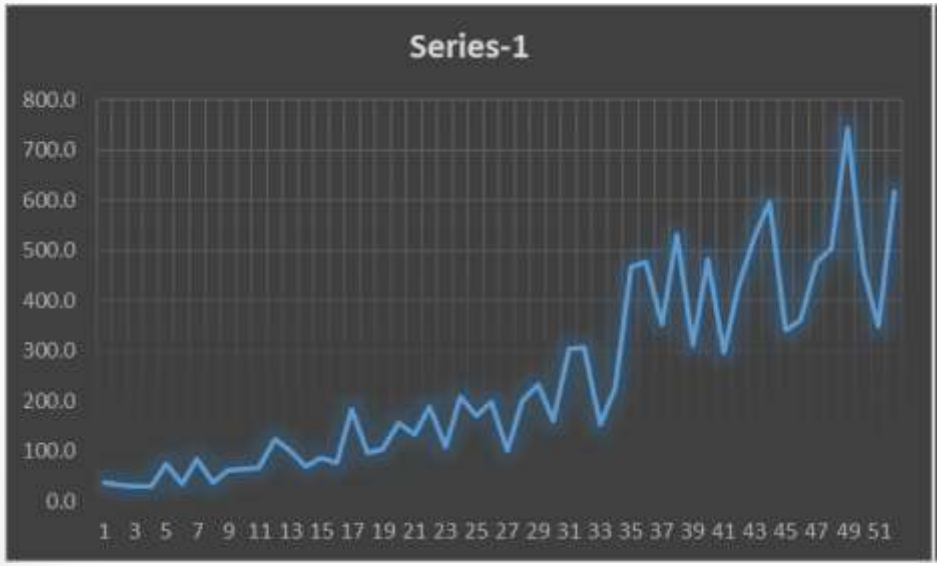
Seasonal effects not accounted for

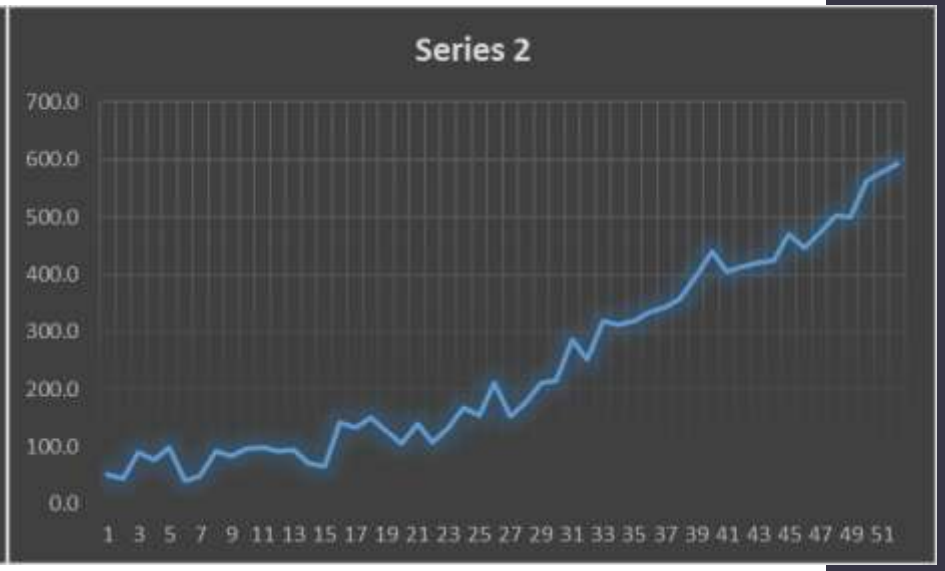
Goodness of fit

- We need a way to compare different time series techniques for a given data set.
- Four common techniques are the:
  - Mean absolute deviation, $$\mathrm{MAD} = \sum_{i=1}^{n} \frac{\left|\mathrm{Y}_i - \hat{\mathrm{Y}}_i\right|}{n}$$
  - Mean absolute percent error $$\mathrm{MAPE} = \frac{100}{n} \sum_{i=1}^{n} \frac{\left|\mathrm{Y}_i - \hat{\mathrm{Y}}_i\right|}{\mathrm{Y}_i}$$
  - Mean square error, $$\mathrm{MSE} = \sum_{i=1}^{n} \frac{\left(\mathrm{Y}_i - \hat{\mathrm{Y}}_i\right)^2}{n}$$
  - Root mean square error. $$\mathrm{RMSE} = \sqrt{\mathrm{MSE}}$$

Additive and Multiplicative. What model to use?

- Is there any striking difference between these two time series?





Additive and Multiplicative. What model to use?

- In many time series, the amplitude of both the seasonal and irregular variations increase as the level of the trend rises. In this situation, a multiplicative model is usually appropriate.

- In some time series, the amplitude of both the seasonal and irregular variations do not change as the level of the trend rises or falls. In such cases, an additive model is appropriate.

Lab-2

- Find the error at each point in above problem
- Conduct the residual analysis for problem in lab-1

Data Set-2: Predicting the Stock Price

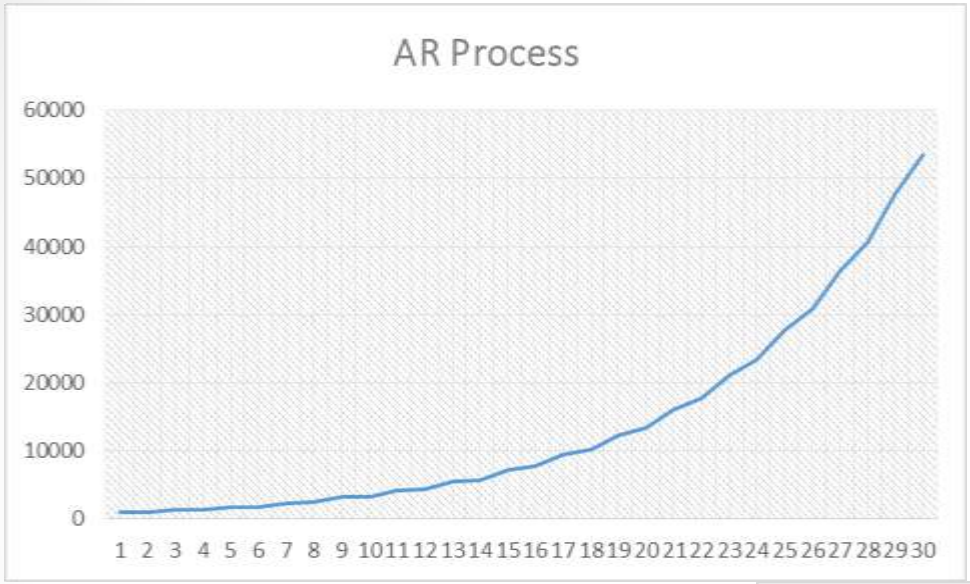
- Download Coal India Limited Stock data from Yahoo finance
- De trend it by calculating MA5(one week)
- Use day function and pivot table to identify seasonality interval
- Are 5-26 & 26 to 4 are two seasons in every month ?
- Find the seasonal indices and remove the seasonality
- Conduct residual analysis
- Forecast future stock price values, buy and sell at the right time and **become rich**

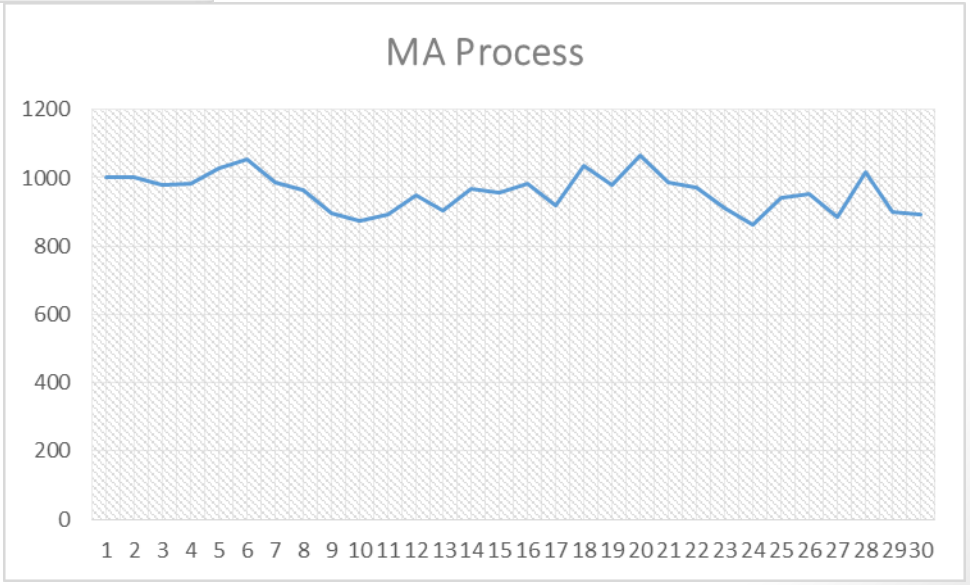
Time Series Models : Type 2

- Autoregressive (AR) process: series current values depend on its own previous values
- Moving average (MA) process: The current deviation from mean depends on previous deviations
- Autoregressive Moving average (ARMA) process
- Autoregressive Integrated Moving average (ARIMA)process.

- ARIMA is also known as Box-Jenkins approach. It is popular because of its generality; It can handle any series, with or without seasonal elements, and it has well-documented computer programs

AR Process and MA Process





ARIMA Model

Autoregressive Integrated Moving Average

 → AR filter → Integration filter → MA filter → ε_t

(long term) (stochastic trend) (short term) (white noise error)

ARIMA (2,0,1) $y_t = a_1 y_{t-1} + a_2 y_{t-2} + \varepsilon_t + b_1 \varepsilon_{t-1}$

ARIMA (3,0,1) $y_t = a_1 y_{t-1} + a_2 y_{t-2} + a_3 y_{t-3} + \varepsilon_t + b_1 \varepsilon_{t-1}$

ARIMA (1,1,0) $\Delta y_t = a_1 \Delta y_{t-1} + \varepsilon_t$, where $\Delta y_t = y_t - y_{t-1}$

ARIMA (2,1,0) $\Delta yt = a_1 \Delta y_{t-1} + a_2 \Delta y_{t-2} + \varepsilon_t$ where $\Delta yt = yt - yt\text{-}1$

ARIMA (2,1,1) =? ARIMA (2,1,2) =? ARIMA (3,1,2) =?

ARIMA(1,0,0) =? ARIMA(2,0,0) =?

ARIMA(0,0,1) =? ARIMA(0,0,2) =?

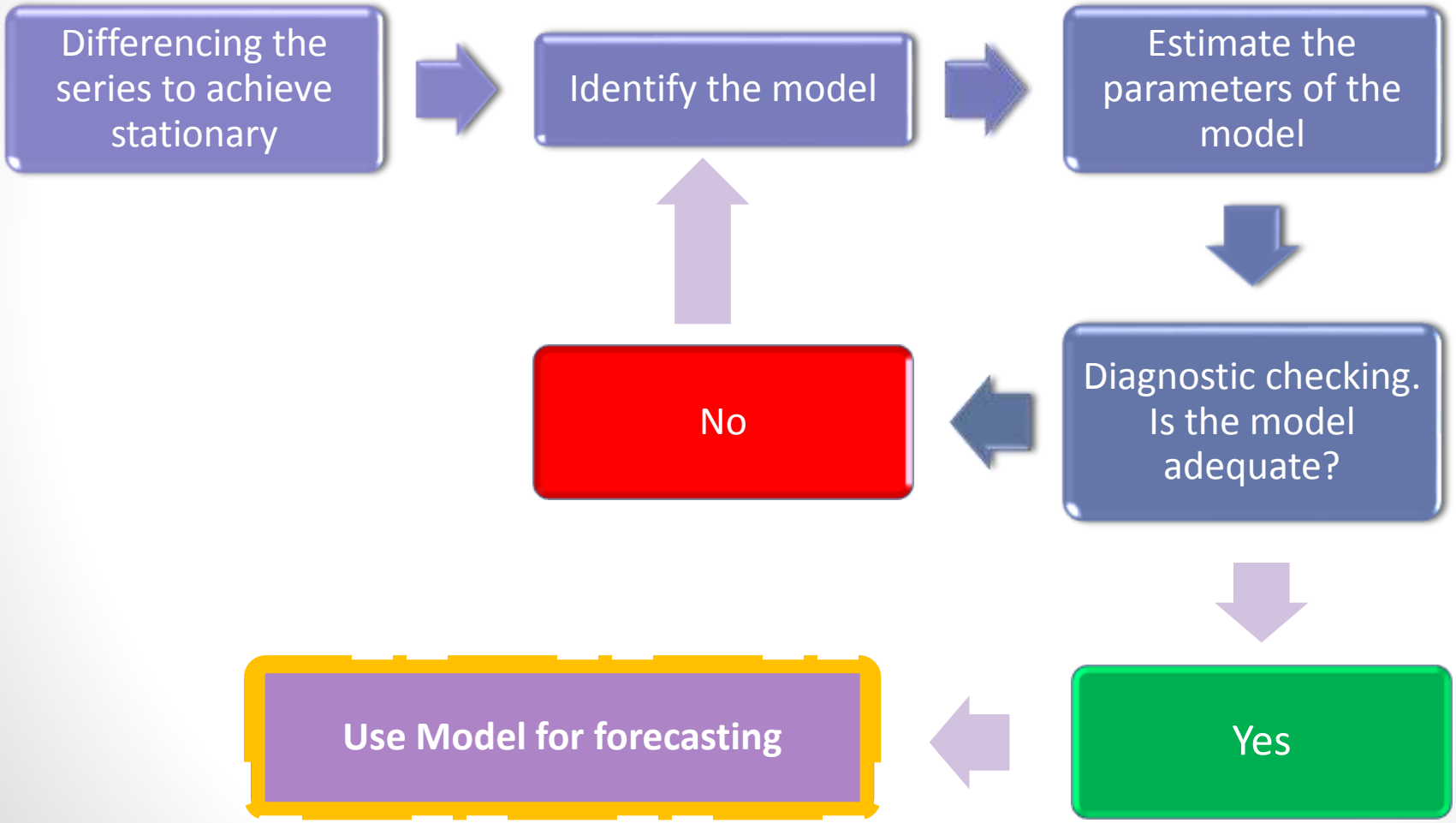
To build a time series model issuing ARIMA, we need to study the time series and identify p,d,q…….simple :P

ARIMA (p,d,q) modeling

To build a time series model issuing ARIMA, we need to study the time series and identify p,d,q

- **Identification:**
  - Determine the appropriate values of p, d, & q using the ACF, PACF, and unit root tests
  - p is the AR order, d is the integration order, q is the MA order
- **Estimation :**
  - Estimate an ARIMA model using values of p, d, & q you think are appropriate.
- **Diagnostic checking:**
  - Check residuals of estimated ARIMA model(s) to see if they are white noise; pick best model with well behaved residuals.
- **Forecasting:**
  - Produce out of sample forecasts or set aside last few data points for in-sample forecasting.

The Box-Jenkins Approach

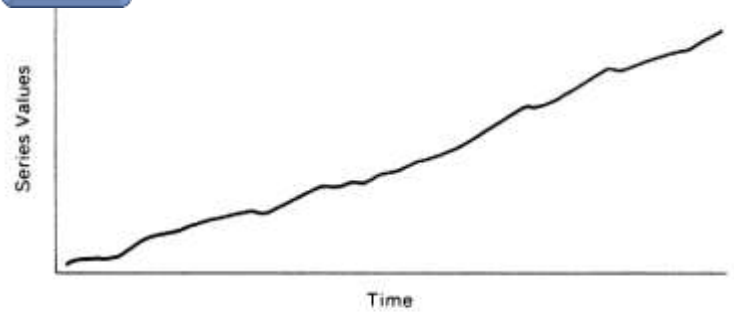


The time series has to be Stationary Processes

- In order to model a time series with the Box-Jenkins approach, the series has to be stationary
- In practical terms, the series is stationary if tends to wonder more or less uniformly about some fixed level
- In statistical terms, a stationary process is assumed to be in a particular state of statistical equilibrium, i.e., $p(x_t)$ is the same for all t
- In particular, if z_t is a stationary process, then the first difference $\nabla z_t = z_t - z_{t-1}$ and higher differences $\nabla^d z_t$ are stationary
- BTW …..Most time series are nonstationary

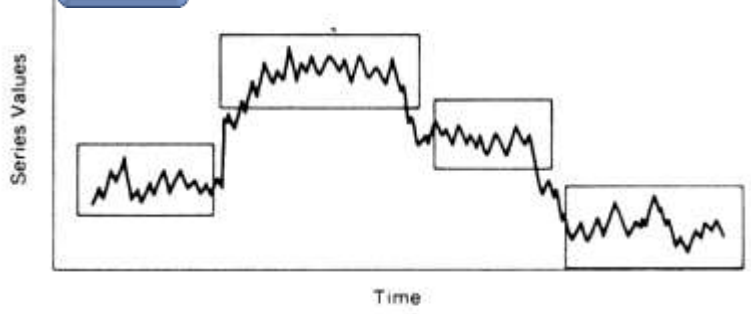
Some non stationary series

1



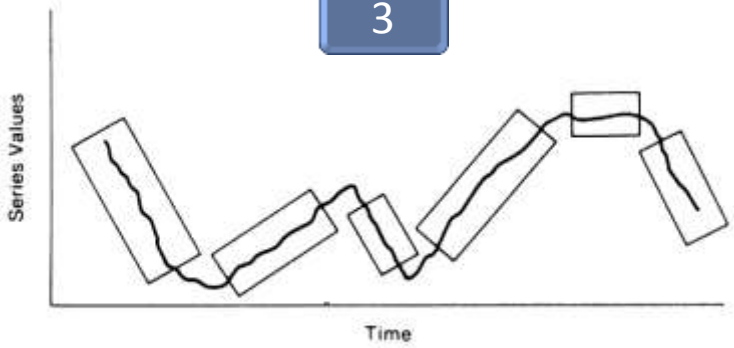
A Nonstationary Series: Overall Trend

2



A Nonstationary Series: Random Changes in Level

3



A Nonstationary Series: Random Changes in Both Level and Slope

Achieving Stationarity

- Regular differencing (RD)

(1st order) $\nabla x_t = (1 - B)x_t = x_t - x_{t-1}$

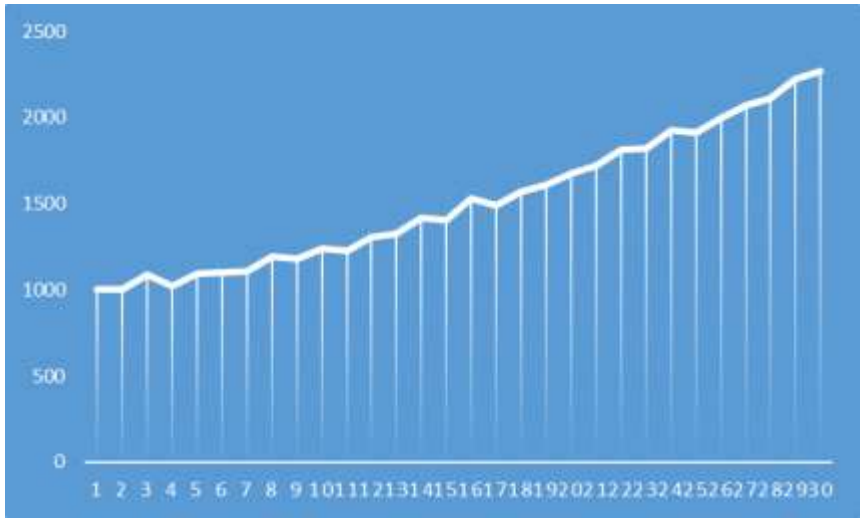
(2nd order) $\nabla^2 x_t = (1 - B)^2 x_t = x_t - 2x_{t-1} + x_{t-2}$

"B" is the backward shift operator

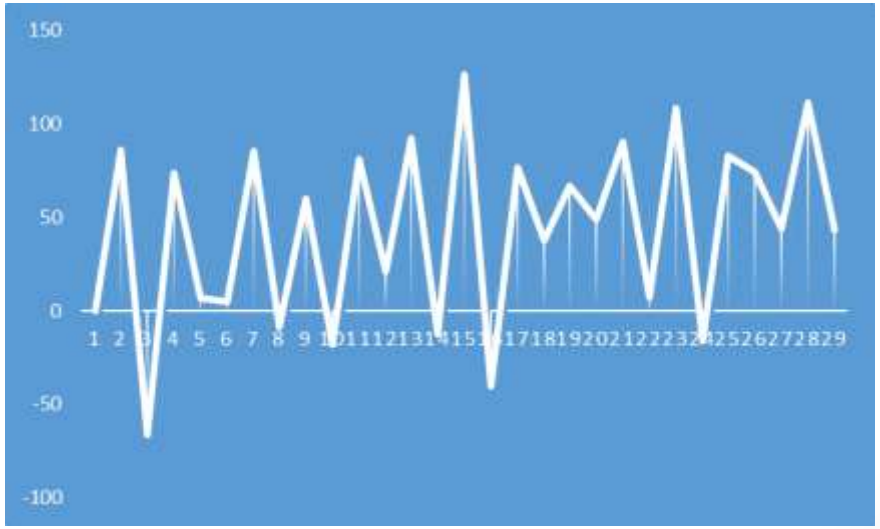
- It is unlikely that more than two regular differencing would ever be needed
- Sometimes regular differencing by itself **is not** sufficient and prior transformation is also needed

Differentiation

Actual Series



Series After Differentiation



Identification of orders p and q

- Identification starts with d
- ARIMA(p,d,q)
- What is Integration here?

- First we need to make the time series stationary

- We need to learn about ACF & PACF to identify p,q
- Once we are working with a stationary time series, we can examine the ACF and PACF to help identify the proper number of lagged y (AR) terms and ε (MA) terms.

Autocorrelation Function (ACF)

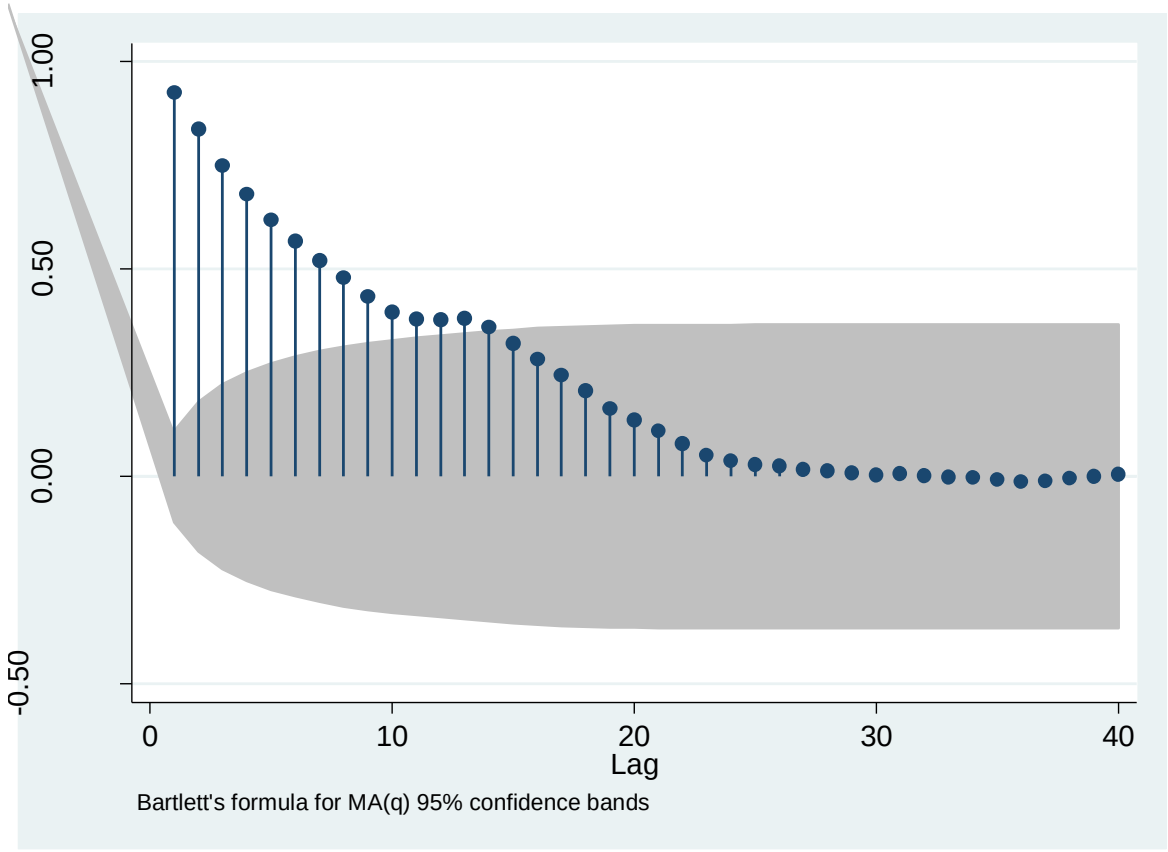
- Correlation with lag-1, lag2, lag3 etc.,
- The ACF represents the degree of persistence over respective lags of a variable.

$\rho_k = \gamma_k / \gamma_0$ = covariance at lag k/ variance

$$\rho_k = \frac{E[(y_t - \mu)(y_{t-k} - \mu)]^2}{E[(y_t - \mu)^2]}$$

ACF (0) = 1, ACF (k) = ACF (-k)

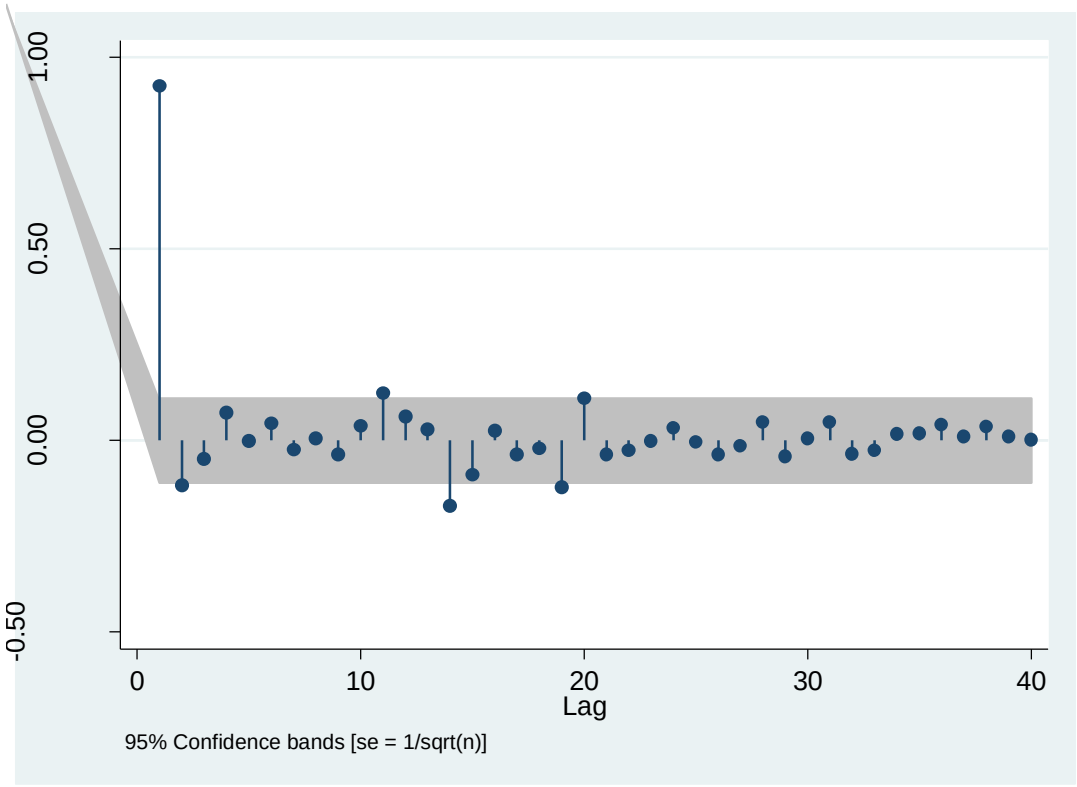
ACF Graph


1.00
0.50
0.00
-0.50
0
10
20
30
40
Lag
Bartlett's formula for MA(q) 95% confidence bands

Partial Autocorrelation Function (PACF)

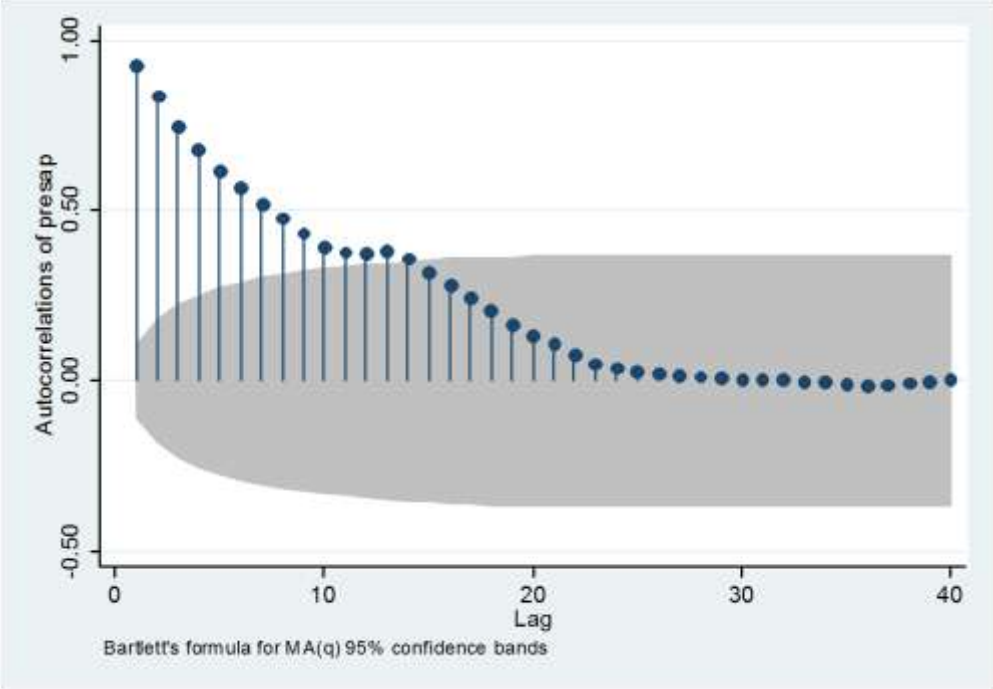
- Partial regression coefficient -
- The lag k partial autocorrelation is the partial regression coefficient, θ_{kk} in the k^{th} order autoregression
- $y_t = \theta_{k1} y_{t-1} + \theta_{k2} y_{t-2} + \ldots + \theta_{kk} y_{t-k} + \varepsilon_t$
- **Partial correlation** measures the degree of association between two random variables, with the effect of a set of controlling random variables removed.

PACF Graph



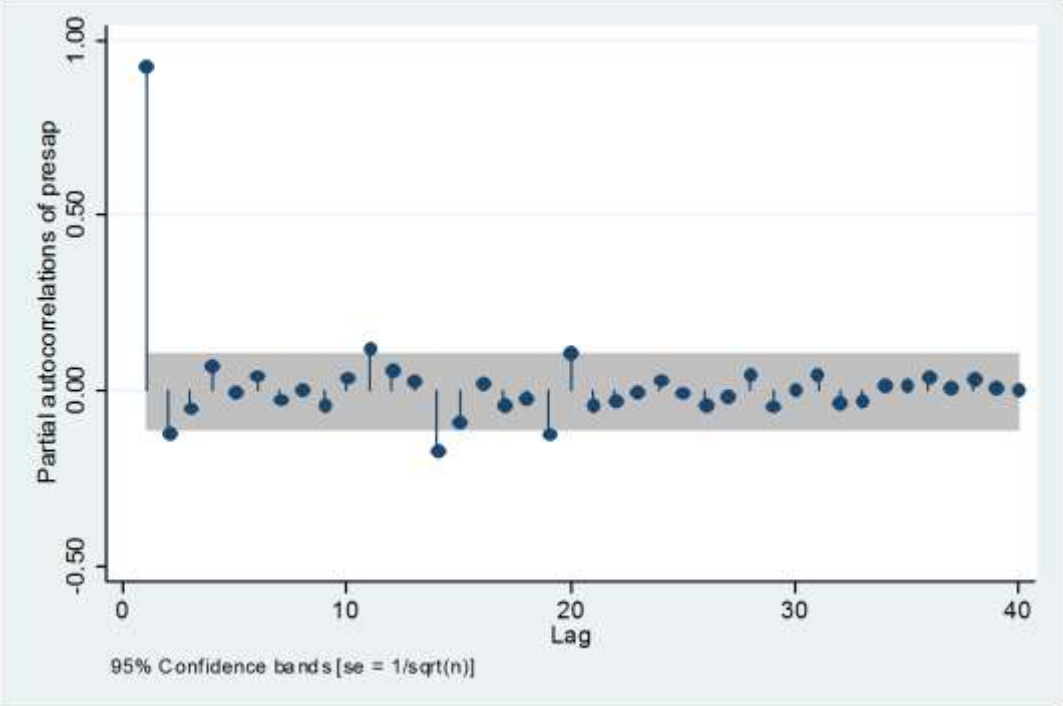
Identification of AR Processes & its order -p

- What is AR procerss
- For AR models, the ACF will dampen exponentially
- The PACF will identify the order of the AR model:
  - The AR(1) model ($y_t = a_1 y_{t-1} + \varepsilon_t$) would have one significant spike at lag 1 on the PACF.
  - The AR(3) model ($y_t = a_1 y_{t-1} + a_2 y_{t-2} + a_3 y_{t-3} + \varepsilon_t$) would have significant spikes on the PACF at lags 1, 2, & 3.



Identification of MA Processes & its order - q

- Recall that a MA(q) can be represented as an AR(∞), thus we expect the opposite patterns for MA processes.
- The PACF will dampen exponentially.
- The ACF will be used to identify the order of the MA process.
- MA(1) (yt = εt + b1 εt-1) has one significant spike in the ACF at lag 1.
- MA (3) (yt = εt + b1 εt-1 + b2 εt-2 + b3 εt-3) has three significant spikes in the ACF at lags 1, 2, & 3.



Parameter Estimate

- We already know the model equation. AR(1,0,0) or AR(2,1,0) or ARIMA(2,1,1)
- We need to estimate the coefficients using Least squares. Minimizing the sum of squares of deviations

$$min \qquad \sum_t \epsilon_t^2$$

$$min \qquad \sum_{t=2}^{T} (y_t - \phi y_{t-1})^2$$

Interpreting Coefficients

- If we include lagged variables for the dependent variable in an OLS model, we cannot simply interpret the β coefficients in the standard way.
- Consider the model, $Yt = a0 + a1Yt-1 + b1Xt + \varepsilon t$
- The effect of Xt on Yt occurs in period t, but also influences Yt in period t+1 because we include a lagged value of Yt-1 in the model.
- To capture these effects, we must calculate multipliers (impact, interim, total) or mean/median lags (how long it takes for the average effect to occur).

How good is my model?

- Does our model really give an adequate description of the data
- Two criteria to check the goodness of fit
  - Akaike information criterion (AIC)
  - Schwartz Bayesiancriterion (SBC)/Bayesian information criterion (BIC).

Goodness of fit

- Remember… **Residual analysis** and Mean deviation, Mean Absolute Deviation and Root Mean Square errors?
- Four common techniques are the:
  - Mean absolute deviation, $$\mathrm{MAD} = \sum_{i=1}^{n} \frac{\left|\mathrm{Y}_i - \hat{\mathrm{Y}}_i\right|}{n}$$
  - Mean absolute percent error $$\mathrm{MAPE} = \frac{100}{n} \sum_{i=1}^{n} \frac{\left|\mathrm{Y}_i - \hat{\mathrm{Y}}_i\right|}{\mathrm{Y}_i}$$
  - Mean square error, $$\mathrm{MSE} = \sum_{i=1}^{n} \frac{\left(\mathrm{Y}_i - \hat{\mathrm{Y}}_i\right)^2}{n}$$
  - Root mean square error. $$\mathrm{RMSE} = \sqrt{\mathrm{MSE}}$$

Lab

- Download time_slales data
- Draw the trend graph
- Test the stationary using stationarity=(DICKEY) option
- Fit a time series model
  - Correlograms
  - Do we need to differentiate the model?
  - Test the stationary using stationarity again
  - Identify p, q
- Draw a residual plot
- Predict the sales for Mar-2013

Venkat Reddy Konasani

Manager at Trendwise Analytics

venkat@TrendwiseAnalytics.com

21.venkat@gmail.com

+91 9886 768879