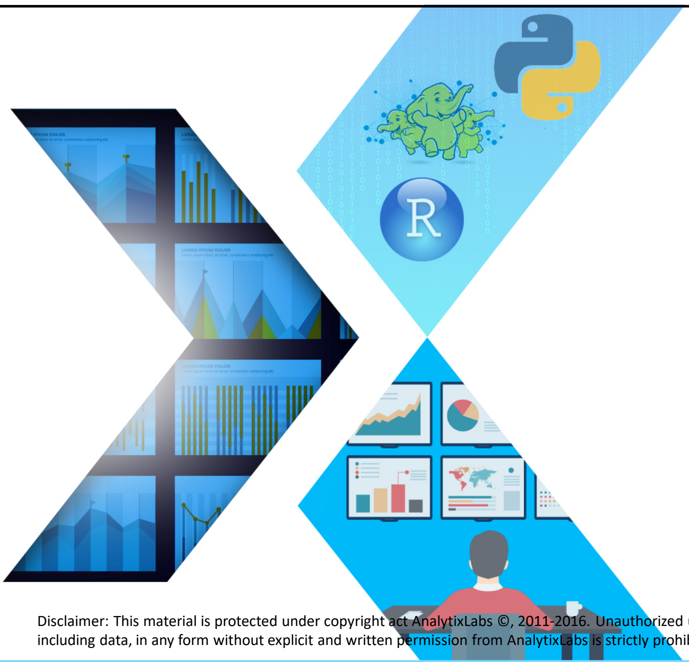


ANALYTIXLABS

Time Series Analysis - Forecasting

Introduction to Time Series Forecasting

ANALYTIXLABS

Time Series Introduction

- A time series is a set of observations generated, sequentially in time, on a single variable
- Time periods are of equal length (days, weeks, months, quarters, annual)
- Time series analysis accounts for the fact that data points taken over time may have an internal structure (such as auto correlation, trend or seasonal variation) that should be accounted for
- Time series data is indexed by time and No missing values
- For example
  - daily stock data,
  - monthly unemployment data,
  - annual sales data
  - Macro Economic indicators

ANALYTIXLABS

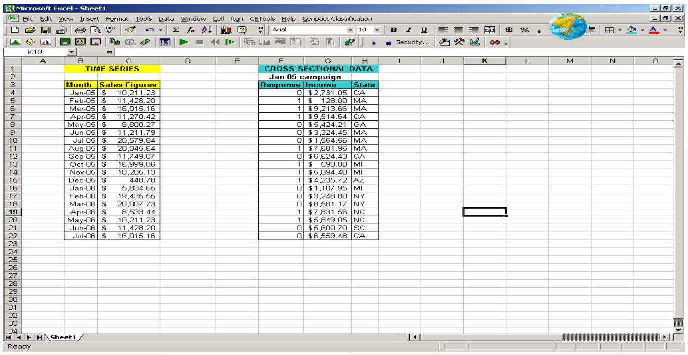
Data Types

The success of any analysis ultimately depends on the availability of the appropriate data. It is therefore essential that we spend some time discussing the types of data that one may encounter in empirical analysis. Three types of data may be available for empirical analysis

Time Series A Time Series is an ordered sequence of values of a variable observed at equally spaced time intervals. For example, monthly sales figures, daily stock prices, weekly interest rates, etc.

Cross Sectional Cross Sectional data has observations from the same time period, for different types of variables. For example, data on price, mileage, and country of origin for automobiles in India in 2005.

They would typically take the following shape:



TIME SERIES

Month	Sales Figures
Jan-05	$ 10,211.23
Feb-05	$ 11,428.20
Mar-05	$ 16,015.16
Apr-05	$ 11,270.42
May-05	$ 8,800.27
Jun-05	$ 11,211.79
Jul-05	$ 20,579.84
Aug-05	$ 20,845.64
Sep-05	$ 11,749.87
Oct-05	$ 16,999.06
Nov-05	$ 10,205.13
Dec-05	$ 448.78
Jan-06	$ 5,834.65
Feb-06	$ 19,435.55
Mar-06	$ 20,007.73
Apr-06	$ 8,533.44
May-06	$ 10,211.23
Jun-06	$ 11,428.20
Jul-06	$ 16,015.16

CROSS-SECTIONAL DATA

Jan-05 campaign

Response	Income	State
0	$2,731.05	CA
1	$ 128.00	MA
1	$9,213.66	MA
1	$9,514.64	CA
0	$5,424.21	GA
0	$3,324.45	MA
0	$1,564.56	MA
1	$7,681.96	MA
0	$6,624.43	CA
1	$ 598.00	MI
1	$5,094.40	MI
1	$4,235.72	AZ
0	$1,107.95	MI
0	$3,248.80	NY
0	$8,581.17	NY
1	$7,831.56	NC
1	$5,849.05	NC
0	$5,600.70	SC
0	$6,559.48	CA

ANALYTIXLABS

Data Types Contd....

Pooled Data

This type is a combination of both time series and cross-sectional data is pooled data. The following example will help in understanding:

State	Price in 2003	Price in 2004
AL	20.18	20.58
AK	15.31	15.62
AZ	14.23	14.51
AR	20.64	21.05
CA	5.51	5.62
CO	19.76	20.15
CT	0.81	0.82
DE	37.30	38.04
FL	8.41	8.58
GA	44.66	45.55

A special type of pooled data is Panel, Longitudinal or Micropanel data. This is a special type of pooled data in which the same cross-sectional unit (say a family, or a firm) is surveyed over time. For example, in each periodic survey of the government, the same household (or people living at the same address) is interviewed to find out if there has been a change in the housing and financial conditions of that household since the last survey.

By interviewing the same household periodically, the panel data provides very useful information on the dynamics of household behavior.

Types of Forecast

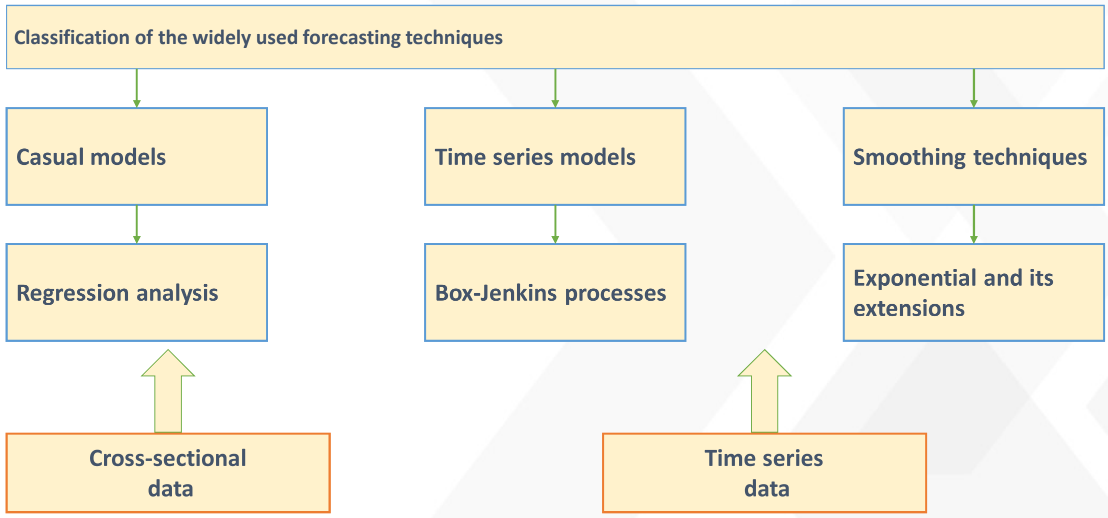
Short-term forecasts
Are needed for the scheduling of personnel, production and transportation. As part of the scheduling process, forecasts of demand are often also required.

Medium-term forecasts
Are needed to determine future resource requirements, in order to purchase raw materials, hire personnel, or buy machinery and equipment.

Long-term forecasts
Are used in strategic planning. Such decisions must take account of market opportunities, environmental factors and internal resources.

For Case Studies, please refer:
https://www.otexts.org/fpp/1/5

Statistical Forecasting



Time Series Data - Components

Time Series Analysis – Time Series components

Time Series data allows us to either to do analysis or forecasting.
Time series analysis comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data.

Time series forecasting is the use of a model to predict future values based on previously observed values.

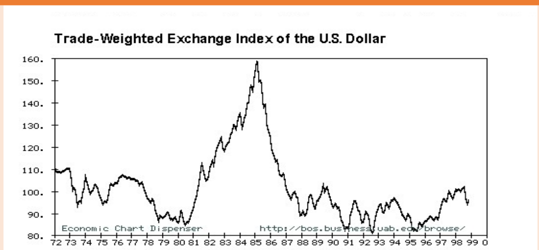
Components: Time Series can be decomposed into 4 components

- **Trend:** Trend is the gradual, long-run (or secular) evolution of the variables that we are seeking to forecast.
- **Seasonal Effects:** Many series display a regular pattern of variability depending on the time of year. This pattern is known as the seasonal effect.
- **Cyclic Component:** Fluctuations around the trend, excluding the irregular component, revealing a succession of phases of expansion and contraction.
- **Irregular Component (White Noise):** The unexplained remaining variability.

Components of a Time Series

Components of a Time series are:

Irregular Components:
These are disturbances or residual variation that remain after all the other behaviors have been accounted for.



Component Analysis

Trend Analysis can be done using linear and non linear with time as an explanatory variable. Also it does not assume the condition of equally spaced time series.

Seasonal Index represents the extent of seasonal influence for a particular segment of the year. A seasonal index is how much the average for that particular period tends to be above (or below) the grand average.

Cyclical Analysis: The cycles present in the data can be studied by removing the trend component. This is done by expressing each actual value in the time series as a percentage of the calculated trend for the same date.

Residual Analysis: Inherent in the collection of data taken over time is some form of random variation. There exist methods for reducing of canceling the effect due to random variation. An often-used technique in industry is "smoothing".

Smoothing

- **Averaging Methods**
- **Exponential Smoothing Methods**

ANALYTIXLABS

Forecasting Techniques- Classification

ANALYTIXLABS

Forecasting Techniques

Forecasting Techniques

- Judgmental (Qualitative)
  - Executive Committee
  - Delphi Technique
  - Survey of Sales Force
  - Survey of Customers
  - Game Theory
  - Judgmental Bootstrapping
- Statistical (Quantitative)
  - Time Series
    - ARIMA
    - Smoothing
    - ARCH/GARCH Models
    - Moving Averages
  - Casual
    - Simple Linear Regression
    - Multiple Linear Regression
    - Neural Networks

ANALYTIXLABS

Time Series Techniques (1/2)

Is there any pattern?

- **Pattern Based**
  - Trend (T)
  - Seasonal index (SI)
  - Combined trend and seasonal index (Comb)
- **Pattern-less**
  - Simple Averages
  - Moving Averages(MA)
  - Exponential smoothing (ES)
  - Naive

ANALYTIXLABS

Time Series Techniques (2/2)

Types of Time Series Techniques

Time Series Models

How many Dependent variables?

=1

- Univariate time series model
- ARIMA(p,d,q), ARCH, GARCH, etc

- Single equation multivariate model
- ARIMAX(p,d,q) with ARCH, GARCH, etc

>1

- Simultaneous equation model or structural equation model
- VAR or VECM
- STATESPACE
- PANEL DATA MODEL,etc

ANALYTIXLABS

Pattern-less Techniques

ANALYTIXLABS

Smoothening Techniques – Moving Averages

A moving average is used to analyze a set of data points by creating a series of averages of different subsets of the full data set.

It is commonly used to smooth out short-term fluctuations and highlight longer-term trends or cycles.

There are 3 types of moving averages:

- Simple Moving Average
- Cumulative Moving Average
- Weighted Moving Average

ANALYTIXLABS

Smoothing Techniques – Moving Averages

Simple Moving Average (SMA): This Involves simply taking a certain number of past periods adding together; then dividing by the number of periods.

Simple Moving Averages (MA) is effective and efficient approach provided the time series is stationary in both mean and variance.

MAt+1 = [Dt + Dt-1 + ... +Dt-n+1] / n

Weighted Moving Average (WMA): A slightly more intricate method for smoothing a raw time series {xt} is to calculate a weighted moving average by first choosing a set of weighting factors (w1,w2…wk) such that summation equal 1.

In practice this is used to give more weight to recent terms as compared to older terms, whereas in SMA; all terms have equal weight

Tip – Use this method to remove or reduce irregularities (random fluctuations) in time series data

ANALYTIXLABS

Exponential Smoothening

Exponential smoothing is a technique that can be applied to time series data, either to produce smoothed data for presentation, or to make forecasts.
This technique is commonly applied to financial market and economic data.
The simplest form of exponential smoothing is given by the formulae:

$$s_1 = x_0$$
$$s_t = \alpha x_{t-1} + (1-\alpha)s_{t-1} = s_{t-1} + \alpha(x_{t-1} - s_{t-1}), t > 1$$

where α is the smoothing factor, $\{x_t\}$ is the raw data sequence and $\{s_t\}$ is the sequence obtained after applying the exponential smoothing algorithm.

The term exponential is used as due to back substitution we get a geometric progression which is the discrete version of the exponential function.

There are various forms of exponential smoothing:
- Simple Exponential Smoothing
- Double Exponential Smoothing
- Triple Exponential Smoothing

While simple exponential smoothing requires stationary condition, the double-exponential smoothing can capture linear trends, and triple-exponential smoothing can handle almost all other business time series (parabolic etc)

ANALYTIXLABS

Pattern Based Technique - Decomposition Method

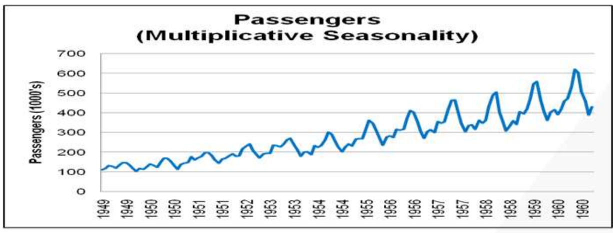
ANALYTIXLABS

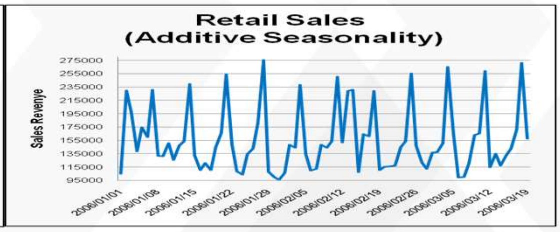
Seasonality

Seasonality: Many time series data follow recurring seasonal patterns. For example sales may peak around Christmas year after year. Movie ticket sales may increase noticeably on weekends. Thus, it may be useful to smooth the seasonal component independently with an extra parameter

Seasonality can be Additive or Multiplicative in nature

Detection: Detection of seasonality can involve plotting and visually inspecting the series, by method of indexing and also by analyzing the autocorrelogram





- Fig 1 (Passengers) displays a multiplicative seasonality with a exponentially rising trend
- Fig 2 (Retail Sales) displays an additive seasonality with constant mean (no linear trend)

ANALYTIXLABS

Decomposition Models

- There are two types of Decomposition Models following the classical decomposition of a Time series into trend, seasonal, cyclical and irregular components
  - Additive:
    - In the Additive model, the observed time series (Y_t) is considered to be the sum of four independent components: the trend T_t the seasonal S_t, the cyclical C_t and the irregular I_t

  $Y_t = T_t + S_t + C_t + I_t$
  - Multiplicative:
    - In the Multiplicative model, the observed time series (Y_t) is considered to be the product of four independent components: the trend T_t the seasonal S_t, the cyclical C_t and the irregular I_t

  $Y_t = T_t \times S_t \times C_t \times I_t$
- It is difficult to estimate cyclical component because of the following reasons:
  - Paucity of data in real life case studies
  - Period of oscillation is not stable, even if the data is there
    - Hence, for short term forecasting we do not include the Cyclical component. But long term forecasts should include adjustments for Cyclical influences

ANALYTIXLABS

Multiplicative Decomposition Method (1/3)

Trend Forecasting:

Calculating Trend using Regression

Example: Calculate regression model for demand (Y) versus time period (x)

- b_0
- b_1 (the trend, or "average change per quarter")
- R_2
- s_{yx} (similar to MAD or MAPE)

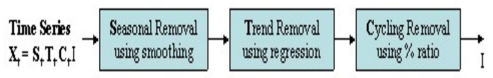
Seasonal Index Forecasting:

What's an index?

- Ratio
- SI = ratio between actual and average demand

Example:

- SI for quarter demand is 1.20
  - What's that mean?
  - Use it to forecast demand for next fall
- So, where did the 1.20 come from?



The Three Signals Decomposition and Its Reversal Processes For Forecasting

ANALYTIXLABS

Multiplicative Decomposition Method (2/3)

Combined Forecasting Approach:

✓ Start by calculating seasonal indices

✓ Then, deseasonalize the demand

- Divide actual demand values by their SI values **(y' = y / SI)**
- Results in transformed data (new time series)
- Seasonal effect removed (recall TS components)

✓ Calculate trend model for deseasonalized data

✓ Use resulting trend model with SI

ANALYTIXLABS

Classical Multiplicative Decomposition(3/3)

Quarter	Year	Period	Data **Y=TxCxSxR**	MA4	CMA4x2 (= TxC)	SR (=SxRx100)	S	Seasonally Adjusted Data (=Y/S = TxC)
1st	1985	1st-1985	4109,0				**101,8**	4036,7
2nd	1985	2nd-1985	3874,0	3942,8			**97,5**	3974,1
3rd	1985	3rd-1985	3842,0	3967,3	3955,0	97,1	**98,5**	3899,0
4th	1985	4th-1985	3946,0	3961,3	3964,3	99,5	**102,2**	3861,4
1st	1986	1st-1986	4207,0	4008,3	3984,8	105,6	**101,8**	4132,9
2nd	1986	2nd-1986	3850,0	4086,8	4047,5	95,1	**97,5**	3949,5
3rd	1986	3rd-1986	4030,0	4083,3	4085,0	98,7	**98,5**	4089,8
4th	1986	4th-1986	4260,0	4133,5	4108,4	103,7	**102,2**	4168,7
1st	1987	1st-1987	4193,0	4157,5	4145,5	101,1	**101,8**	4119,2
2nd	1987	2nd-1987	4051,0	4203,8	4180,6	96,9	**97,5**	4155,7
3rd	1987	3rd-1987	4126,0	4241,5	4222,6	97,7	**98,5**	4187,2
4th	1987	4th-1987	4445,0	4308,5	4275,0	104,0	**102,2**	4349,7
1st	1988	1st-1988	4344,0	4419,8	4364,1	99,5	**101,8**	4267,5
2nd	1988	2nd-1988	4319,0	4452,5	4436,1	97,4	**97,5**	4430,6
3rd	1988	3rd-1988	4571,0	4541,3	4496,9	101,6	**98,5**	4638,8
4th	1988	4th-1988	4576,0	4615,0	4578,1	100,0	**102,2**	4477,9
1st	1989	1st-1989	4699,0	4625,5	4620,3	101,7	**101,8**	4616,3
2nd	1989	2nd-1989	4614,0	4666,0	4645,8	99,3	**97,5**	4733,2
3rd	1989	3rd-1989	4613,0				**98,5**	4681,5
4th	1989	4th-1989	4738,0				**102,2**	4636,5

Models for Time series Analysis

Models for Time Series Analysis

Models for time series data can have many forms and represent different stochastic processes.

When modeling variations in the level of a process, three broad classes of practical importance are the autoregressive (AR) models, the integrated (I) models, and the moving average (MA) models. These three classes depend linearly on previous data points.

Combinations of these ideas produce autoregressive moving average (ARMA) and autoregressive integrated moving average (ARIMA) models.

There are more variations to the above mentioned models which will not be a part of the discussion.

ANALYTIXLABS

Stationary

Before understanding the models for Time Series, the concept of stationarity needs to grasped. Within stationarity there are two important ideas

Strict stationarity

Second order stationarity

A sequence is strongly or strictly stationary if the sequence $\{x_t\}$ has the same distribution for all sets of time points, i.e., its joint probability distribution does not change when shifted in time or space.

$$F_X(x_{t_1+\tau}, \ldots, x_{t_k+\tau}) = F_X(x_{t_1}, \ldots, x_{t_k}).$$

where F(.) is the cumulative distribution function.

A sequence $\{x_t\}$ is weakly or second order stationary if the mean and the second order moments are time invariant, i.e., independent of time.

$$E(x_1) = E(x_2) = \ldots = E(x_t) = \mu$$

$$V(x_1) = V(x_2) = \ldots = V(x_t) = \sigma^2$$

$$Cov(x_t, x_{t-k}) = Cov(x_{t+l}, x_{t-k+l}) = \gamma_k$$

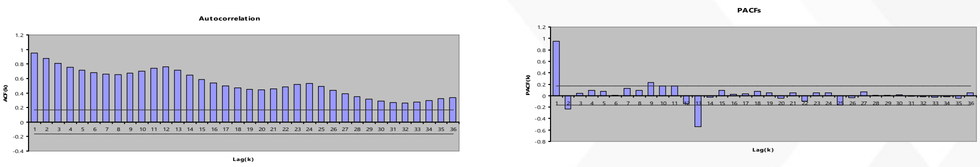
ANALYTIXLABS

Stationary

In Practice, we call a series stationary if the mean and variance is constant over-time.

Detection:

- By analyzing the autocorrelogram where Non-stationary series => ACF tapers gradually



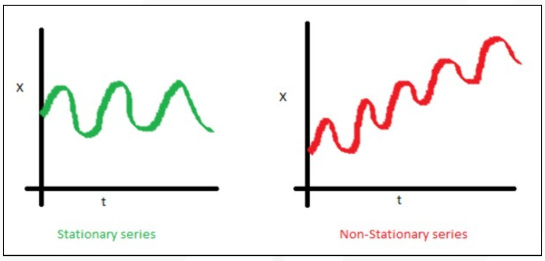
Shows Nonstationarity (doesn't hit 0) & Seasonality with a lag of 12

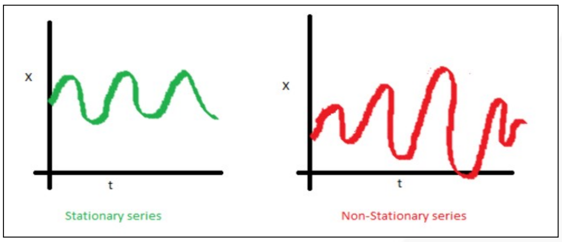
- We can test the series is stationary or not By Performing the Augmented Dickey-Fuller (ADF) Test. This is also called unit root test.

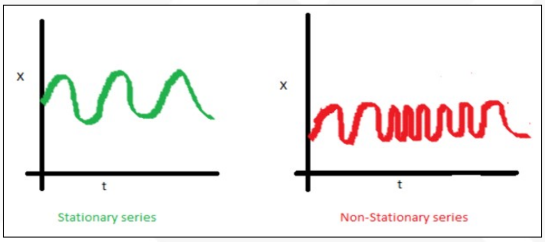
ANALYTIXLABS

Stationary Vs. Non-stationary Series

1. The mean of the series should not be a function of time rather should be a constant
2. The variance of the series should not a be a function of time. This property is known as homoscedasticity
3. The covariance of the i th term and the (i + m) th term should not be a function of time







ANALYTIXLABS

Stationarity

Why do we need to take care of Stationarity?

- The reason I took up this section first was that until unless your time series is stationary, you cannot build a time series model.
- In cases where the stationary criterion are violated, the first requisite becomes to stationarize the time series and then try stochastic models to predict this time series.
- There are multiple ways of bringing this stationarity. Some of them are Detrending, Differencing etc.

How to make a series stationary --

- Differencing - Involves taking difference between successive values
- Log Transformations - makes nonconstant variance constant & removes exponential trends

ANALYTIXLABS

ACF

Auto correlation function: Autocorelate means the correlation's between time series and the same time series lag. It occurs when residual error terms from observations of the same variable at different times are correlated. ACF describes the strength of the relationships between different points in the series.

<table>
<tr><th colspan="5">Autocorrelations</th></tr>
<tr><th>Lag</th><th>Covariance</th><th>Correlation</th><th>-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1</th><th>Std Error</th></tr>
<tr><td>0</td><td>1074393</td><td>1.00000</td><td>| |********************|</td><td>0</td></tr>
<tr><td>1</td><td>-299119</td><td>-.27841</td><td>| ******| . |</td><td>0.087370</td></tr>
<tr><td>2</td><td>-203014</td><td>-.18896</td><td>| ****| . |</td><td>0.093899</td></tr>
<tr><td>3</td><td>-204074</td><td>-.18994</td><td>| ****| . |</td><td>0.096758</td></tr>
<tr><td>4</td><td>182077</td><td>0.16947</td><td>| . |***. |</td><td>0.099563</td></tr>
<tr><td>5</td><td>20177.398</td><td>0.01878</td><td>| . | . |</td><td>0.101742</td></tr>
<tr><td>6</td><td>9435.171</td><td>0.00878</td><td>| . | . |</td><td>0.101768</td></tr>
<tr><td>7</td><td>-102060</td><td>-.09499</td><td>| . **| . |</td><td>0.101774</td></tr>
<tr><td>8</td><td>23076.431</td><td>0.02148</td><td>| . | . |</td><td>0.102448</td></tr>
<tr><td>9</td><td>195254</td><td>0.18173</td><td>| . |**** |</td><td>0.102483</td></tr>
<tr><td>10</td><td>-41248.469</td><td>-.03839</td><td>| . *| . |</td><td>0.104914</td></tr>
<tr><td>11</td><td>-192020</td><td>-.17872</td><td>| ****| . |</td><td>0.105021</td></tr>
<tr><td>12</td><td>14018.991</td><td>0.01305</td><td>| . | . |</td><td>0.107318</td></tr>
</table>

While examining correlograms one should keep in mind that autocorrelations for consecutive lags are formally dependent.

Consider the following example. If the first element is closely related to the second, and the second to the third, then the first element must also be somewhat related to the third one, etc.

This implies that the pattern of serial dependencies can change considerably after removing the first order auto correlation (i.e., after differencing the series with a lag of 1).

ANALYTIXLABS

PACF

Partial Auto Correlation Function: Partial autocorrelations are also correlation coefficients between the basic time series and the same time series lag and we will eliminate the influence of the members between

Partial Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	-0.27841	| ******| . |
2	-0.28886	| ******| . |
3	-0.39995	| ********| . |
4	-0.16005	| ***| . |
5	-0.17300	| ***| . |
6	-0.12588	| ***| . |
7	-0.17279	| ***| . |
8	-0.15830	| ***| . |
9	0.10484	| . |**. |
10	0.08140	| . |**. |
11	-0.05537	| . *| . |
12	0.01630	| . | . |

Another useful method to examine serial dependencies is to examine the partial autocorrelation function (PACF) - an extension of autocorrelation, where the dependence on the intermediate elements (those within the lag) is removed.

In other words the partial autocorrelation is similar to autocorrelation, except that when calculating it, the (auto) correlations with all the elements within the lag are partialled out.

If a lag of 1 is specified (i.e., there are no intermediate elements within the lag), then the partial autocorrelation is equivalent to auto correlation. In a sense, the partial autocorrelation provides a "cleaner" picture of serial dependencies for individual lags (not confounded by other serial dependencies).

ANALYTIXLABS

Augment Dickey Puller Test(Unit-Root Test)

An augmented Dickey–Fuller test (ADF) is a test for a unit root in a time series sample. It is an augmented version of the Dickey–Fuller test for a larger and more complicated set of time series models.

The augmented Dickey–Fuller (ADF) statistic, used in the test, is a negative number. The more negative it is, the stronger the rejection of the hypothesis that there is a unit root at some level of confidence

Augmented Dickey-Fuller Unit Root Tests

Type	Lags	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
Zero Mean	0	-112.675	0.0001	-9.61	<.0001		
	1	-157.129	0.0001	-8.82	<.0001		
	2	-208.873	0.0001	-7.63	<.0001		
Single Mean	0	-113.484	0.0001	-9.63	<.0001	46.38	0.0010
	1	-160.181	0.0001	-8.86	<.0001	39.25	0.0010
	2	-220.278	0.0001	-7.71	<.0001	29.73	0.0010
Trend	0	-113.617	0.0001	-9.60	<.0001	46.10	0.0010
	1	-160.347	0.0001	-8.83	<.0001	38.98	0.0010
	2	-220.789	0.0001	-7.69	<.0001	29.55	0.0010

H_0 rejected. Series is stationary

ANALYTIXLABS

ARIMA (1/2)

ARIMA stands for "**Autoregressive Integrated Moving Average**", and is a combination of the following two types of processes:

- **Autoregressive (AR) Process**: Future values depend on previous value of the data:

$$Y_t = c + a_1 Y_{t-1} + e_t \text{ , } -1<a_1<1$$

AR(p) process, having the following mathematical form:

$$Y_t = c + a_1 Y_{t-1} + a_2 Y_{t-2} + a_3 Y_{t-3} + \dots + a_p Y_{t-p} + e_t$$

- **Moving Average (MA) Process**: Future values depend on previous values of the errors

$$Y_t = c + e_t - b_1 e_{t-1} \text{ , } -1<b_1<1$$

MA(q) process, having the following mathematical form:

$$Y_t = c + e_t - b_1 e_{t-1} + b_2 e_{t-2} + b_3 e_{t-3} + \dots + b_q e_{t-q}$$

- **Autoregressive Moving Average (ARMA) models:**

AR models can be effectively coupled with MA models to form a general and useful class of time series models. These can only be used when the data are stationary

ANALYTIXLABS

ARIMA (2/2)

- **Autoregressive Integrated Moving Average (ARIMA) models:**

This class of models can be extended to non-stationary series by allowing differencing of the data series.

– $Y_t = d + a_1 Y_{t-1} + a_2 Y_{t-2} + a_3 Y_{t-3} + \dots + a_p Y_{t-p} + e_t - b_1 e_{t-1} + b_2 e_{t-2} + b_3 e_{t-3} + \dots + b_q e_{t-q}$

- **Different configurations of ARIMA**

There is a huge variety of ARIMA models. The general non-seasonal model is known as ARIMA (p,d,q):

– **AR: p = order of the autoregressive part**
– **I: d = degree of first differencing involved**
– **MA: q = order of the moving average**

Note that if p,d, and q are equal to zero, the model can also be written in a shorthand notation by dropping the unused parts of the model. For example, an ARIMA(2,0,0) can be written as AR(2) because there is no differencing (I) and no moving average (MA) part.

ANALYTIXLABS

ARIMA - Three Stages

A R I M A

Identification

Estimation

Forecasting

Identification	Here we need to check – (a) Stationarity/ Non-Stationarity (b) Seasonality (c) Order of AR and MA processes (d) White Noise
Estimation	To specify an ARIMA model to fit to the variable specified in the previous IDENTIFY statement and to estimate the parameters of the specified model
Forecasting	Forecast the future values using the model

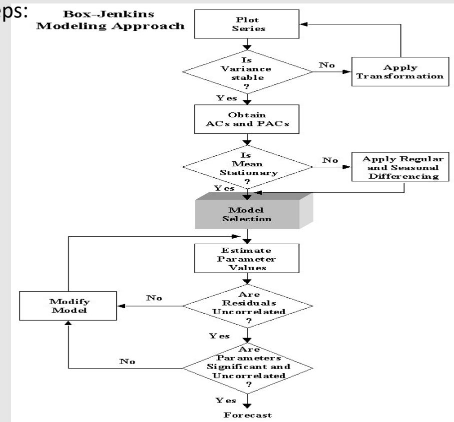
ANALYTIXLABS

Box Jenkins Procedure

Now the million $ question is

How does one know whether it follows a purely AR process or a purely MA process or an ARMA process or an ARIMA process, in which case we must know p, d, and q.

The Box Jenkins (BJ) methodology comes in handy in answering the preceding question. The method consists of the following steps:



The steps are

- Plot the time series data to check for trend and seasonality
- Check for stationarity
- Plot ACF, and PACF
- Make the data stationary through transformations (if needed)
- Plot the correlation functions and check again
- Identify the time series process AR, MA, or ARIMA
- Identify the lag
- Fit the ARIMA model
- Check the residuals
- Check the significance for each of the explanatory variables
- Drop the insignificant variables
- Get the final model

ANALYTIXLABS

Box Jenkins Procedure – Summary Rules for Identifying ARIMA

Identifying the order of differencing and the constant

- ✓ If the series has positive autocorrelations out to a high number of lags, then it probably needs a higher order of differencing.
- ✓ If the lag-1 autocorrelation is zero or negative, or the autocorrelations are all small and pattern less, then the series does not need a higher order of differencing. If the lag-1 autocorrelation is -0.5 or more negative, the series may be over differenced.
- ✓ The optimal order of differencing is often the order of differencing at which the standard deviation is lowest.
- ✓ A model with no orders of differencing assumes that the original series is stationary (among other things, mean-reverting). A model with one order of differencing assumes that the original series has a constant average trend. A model with two orders of total differencing assumes that the original series has a time-varying trend.
- ✓ A model with no orders of differencing normally includes a constant term (which represents the mean of the series). A model with two orders of total differencing normally does not include a constant term. In a model with one order of total differencing, a constant term should be included if the series has a non-zero average trend.

ANALYTIXLABS

Box Jenkins Procedure – Summary Rules for Identifying ARIMA

Identifying the number of AR & MA terms

- ✓ If the partial autocorrelation function (PACF) of the differenced series displays a sharp cutoff and/or the lag-1 autocorrelation is positive--i.e., if the series appears slightly "under differenced"--then consider adding one or more AR terms to the model. The lag beyond which the PACF cuts off is the indicated number of AR terms.
- ✓ If the autocorrelation function (ACF) of the differenced series displays a sharp cutoff and/or the lag-1 autocorrelation is negative--i.e., if the series appears slightly "over differenced"--then consider adding an MA term to the model. The lag beyond which the ACF cuts off is the indicated number of MA terms.
- ✓ It is possible for an AR term and an MA term to cancel each other's effects, so if a mixed AR-MA model seems to fit the data, also try a model with one fewer AR term and one fewer MA term--particularly if the parameter estimates in the original model require more than 10 iterations to converge.
- ✓ If there is a unit root in the AR part of the model--i.e., if the sum of the AR coefficients is almost exactly 1--you should reduce the number of AR terms by one and increase the order of differencing by one.
- ✓ If there is a unit root in the MA part of the model--i.e., if the sum of the MA coefficients is almost exactly 1--you should reduce the number of MA terms by one and reduce the order of differencing by one.
- ✓ If the long-term forecasts appear erratic or unstable, there may be a unit root in the AR or MA coefficients.

ANALYTIXLABS

Problems with ARIMA model

- ✓Most authors recommend at least 50 observations.
- ✓They presume weak stationarity, equal-spaced intervals of observations.
- ✓One should have some experience or priori knowledge of the models.
- ✓In ARIMA modeling a time series can be reduced to stationarity by differencing, but Non-stationarity can arise in a variety of ways which is tough to recognize.
- ✓Use of the correlogram and the partial autocorrelation functions for model selection
- ✓In ARIMA MODELS, We assume dependencies between the quantities Yt-1, Yt-2,… If this process contains the seasonal fluctuations, we can expect also the dependence seasons Yt-s, Yt-2s… where s is the length of the period. This process is called SARIMA.

ANALYTIXLABS

Volatility

- ✓Start Volatility refers to the spread of all likely outcomes of an uncertain variable.
- ✓ Statistically, volatility is often measured as the sample standard deviation.
- ✓ Volatility is related to, but not exactly the same as, risk.
- ✓ Risk is associated with undesirable outcome, whereas volatility as a measure strictly for uncertainty could be due to a positive outcome.

ANALYTIXLABS

Auto Regressive Conditional Heteroscedastic Model(ARCH)

- ✓ This model was developed by R. F. Engle in 1982 (Awarded Nobel Prize in Economic Science in 2003) to model varying conditional variance or volatility of a time-series.

An ARCH model is defined as

$$X_t = \varepsilon_t \sigma_t \ , \ \sigma_t^2 = \alpha_0 + \alpha_1 X_{t-1}^2 + \ldots + \alpha_q X_{t-q}^2 ; t = 1, 2, \ldots, n$$

where $\varepsilon t \sim$ IID(0,1) and $\alpha i \geq 0$ are parameters.

- ✓ The model assumes that positive and negative shocks have the same effects on volatility because it depends on the square of the previous shocks.

ANALYTIXLABS

Generalized Auto Regressive Conditional Heteroscedastic Model

- ✓ Bollerslev (1986) proposed Generalized ARCH (GARCH) model, in which conditional variance is also a linear function of its own lags and has the form.

An GARCH model is defined as

$$X_t = \varepsilon_t \sigma_t \ , \ \sigma_t^2 = \alpha_0 + \alpha_1 X_{t-1}^2 + \ldots + \alpha_p X_{t-p}^2 + \beta_1 \sigma_{t-1}^2 + \ldots + \beta_q \sigma_{t-q}^2 ; t = 1, 2, \ldots, n$$

where $\varepsilon t \sim$ IID(0,1) and αi & $\beta i \geq 0$ are parameters

- ✓ In literature also different types of GARCH model is present like EGARCH, TGARCH, FIGARCH etc.

ANALYTIXLABS

Vector Auto Regression

- ✓It is a multiple time-series generalization of the AR model.
- ✓A VAR requires less restrictive assumptions than other multi-variable methods.
- ✓The VAR(2) model with three time series sequences is:

$$\begin{pmatrix} y_{1t} \\ y_{2t} \\ y_{3t} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} + \begin{pmatrix} \phi_{11}^{(1)} & \phi_{12}^{(1)} & \phi_{13}^{(1)} \\ \phi_{21}^{(1)} & \phi_{22}^{(1)} & \phi_{23}^{(1)} \\ \phi_{31}^{(1)} & \phi_{32}^{(1)} & \phi_{33}^{(1)} \end{pmatrix} \begin{pmatrix} y_{1,t-1} \\ y_{2,t-1} \\ y_{3,t-1} \end{pmatrix} + \begin{pmatrix} \phi_{11}^{(2)} & \phi_{12}^{(2)} & \phi_{13}^{(2)} \\ \phi_{21}^{(2)} & \phi_{22}^{(2)} & \phi_{23}^{(2)} \\ \phi_{31}^{(2)} & \phi_{32}^{(2)} & \phi_{33}^{(2)} \end{pmatrix} \begin{pmatrix} y_{1,t-2} \\ y_{2,t-2} \\ y_{3,t-2} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \\ \varepsilon_{3t} \end{pmatrix}$$

ANALYTIXLABS

Forecasting Accuracy

ANALYTIXLABS

Forecast Accuracy (1/2)

In most Forecasting situations, accuracy is treated as the overriding criterion for selecting a forecasting method.

In many instances, **accuracy** refers to **goodness of fit**, which refers to **how well the forecasting model is able to reproduce the data that is already known**.

The following tree summarizes the techniques that help us in measuring Forecasting Accuracy

How do we measure Forecast Accuracy??

- **Standard Statistical Measures**
- **Out-of-sample accuracy**
- **Naïve Forecast**
- **Theil's U Statistic**
- **ACF of Forecast Error**

Forecast Accuracy (2/2)

Standard Statistical Measures

If Y_t is the actual observation for time period t, and F_t is the forecast for the same time period, then the error is defined as:

$$e_t = Y_t - F_t$$

If there are observations and forecasts for n time periods, then there will be n error terms, and the following standard statistical measures can be defined:

Mean Error (ME) $= \dfrac{\sum_{t=1}^{n} e_t}{n}$

Mean Absolute Error (MAE) $= \dfrac{\sum_{t=1}^{n} |e_t|}{n}$

Mean Squared Error (MSE) $= \dfrac{\sum_{t=1}^{n} e_t^2}{n}$

Percentage Error (PE_t) $= \dfrac{(Y_t - F_t)}{Y_t} \times 100$

Mean Percentage Error (MPE) $= \dfrac{\sum_{t=1}^{n} PE_t}{n}$

Mean Absolute Percentage Error (MAPE) $= \dfrac{\sum_{t=1}^{n} |PE_t|}{n}$

Q&A

Contact us

Visit us on: http://www.analytixlabs.in/

For course registration, please visit: http://www.analytixlabs.co.in/course-registration/

For more information, please contact us: http://www.analytixlabs.co.in/contact-us/

Or email: info@analytixlabs.co.in

Call us we would love to speak with you: (+91) 88021-73069

Join us on:

Twitter - http://twitter.com/#!/AnalytixLabs

Facebook - http://www.facebook.com/analytixlabs

LinkedIn - http://www.linkedin.com/in/analytixlabs

Blog - http://www.analytixlabs.co.in/category/blog/