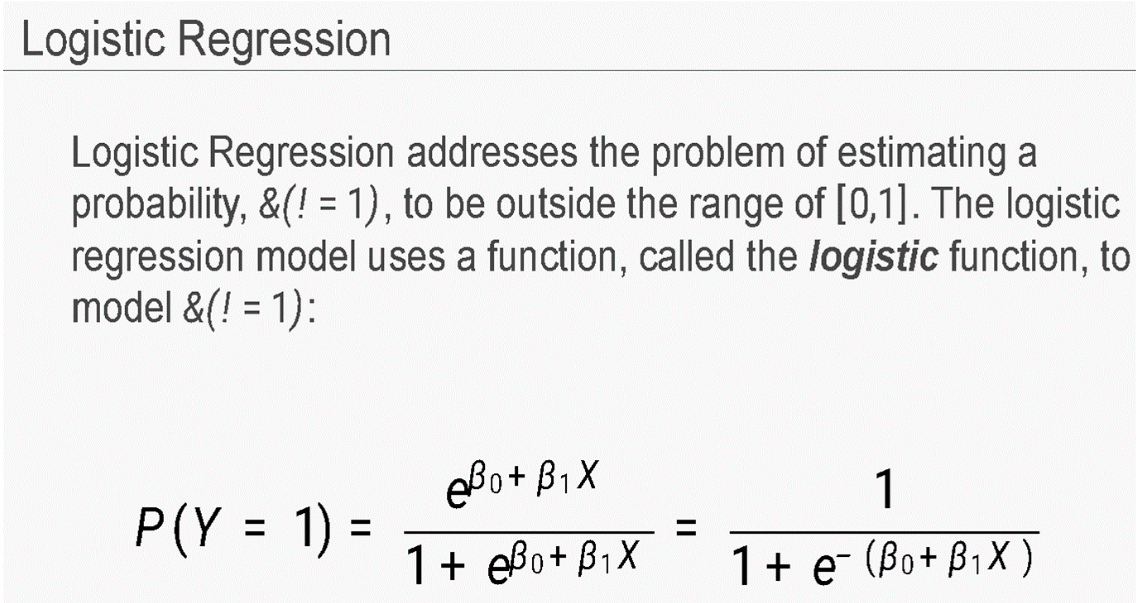
Logistic Regression Definition - Logistic regression is a statistical method that is used for building machine learning models where the dependent variable is dichotomous: i.e. binary. Logistic regression is used to describe data and the relationship between one dependent variable and one or more independent variables. The independent variables can be continuous, nominal, discrete, ordinal, or of interval type



Logistic Regression

Logistic Regression addresses the problem of estimating a probability, &(! = 1), to be outside the range of [0,1]. The logistic regression model uses a function, called the ***logistic*** function, to model &(! = 1):

$$P(Y = 1) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

Outcome variable is binary

Purpose of the analysis is to assess the effects of multiple explanatory variables, which can be numeric and/or categorical, on the outcome variable.

Types of Logistic Regression -

Binomial logistic - the dependent variable is binary in nature. For example, the output can be Success/Failure, 0/1 , True/False, or Yes/No.

Multinomial logistic regression is used when you have one categorical dependent variable with two or more unordered levels (i.e two or more discrete outcomes. For example, let’s imagine that you want to predict what will be the most-used transportation type in the year 2030. The transport type will be the dependent variable, with possible outputs of train, bus, tram, and bike (for example).

Ordinal logistic regression is used when the dependent variable (Y) is ordered (i.e., ordinal). The dependent variable has a meaningful order and more than two categories or levels. Examples of such variables might be t-shirt size (XS/S/M/L/XL), answers on an opinion poll (Agree/Disagree/Neutral), or scores on a test (Poor/Average/Good).

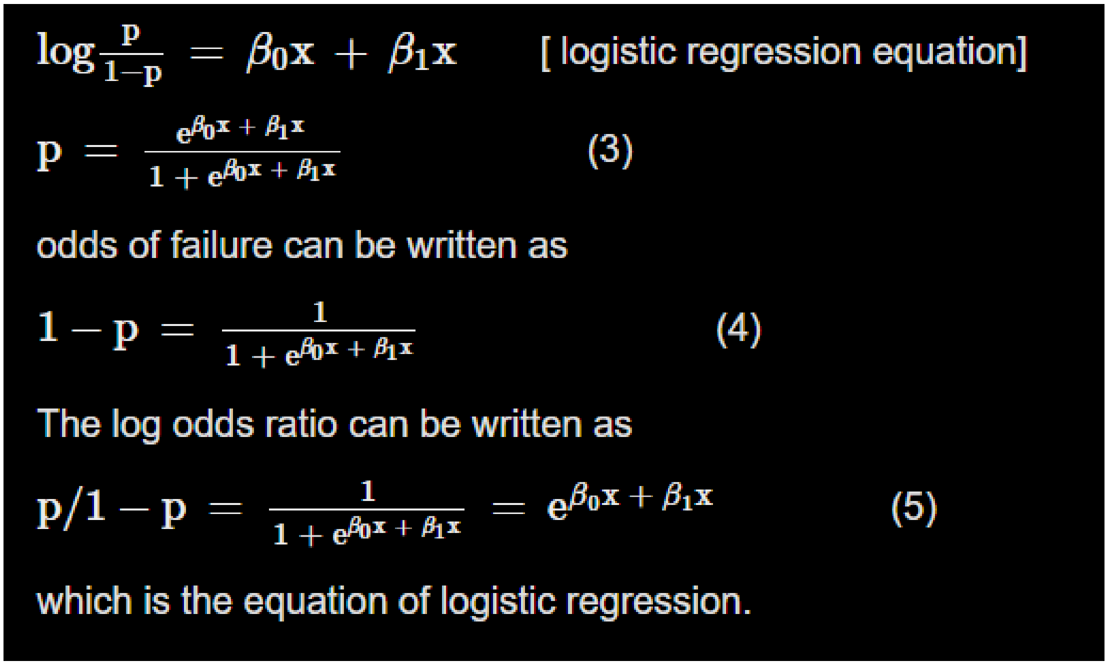
Concept of Probability -

The probability of the outcome is measured by the odds of occurrence of an event.
If P is the probability of an event, then (1-P) is the probability of it not occurring.

Odds of success = P / 1-P

Odds - odds are the chances of success divided by the chances of failure. It is represented in the form of a ratio.
Odds\ Ratio = p/1-p
where,
p -> success odds
1-p -> failure odds

Log Odds -
Log odds play an important role in logistic regression as it converts the LR model from probability based to a likelihood based model. Both probability and log odds have their own set of properties, however log odds makes interpreting the output easier. Thus, using log odds is slightly more advantageous over probability.
Before getting into the details of logistic regression, let us briefly understand what odds are.
Odds : Simply put, odds are the chances of success divided by the chances of failure. It is represented in the form of a ratio. (As shown in equation given below)



$$\log\frac{p}{1-p} = \beta_0 x + \beta_1 x \qquad \text{[logistic regression equation]}$$

$$p = \frac{e^{\beta_0 x + \beta_1 x}}{1 + e^{\beta_0 x + \beta_1 x}} \qquad (3)$$

odds of failure can be written as

$$1 - p = \frac{1}{1 + e^{\beta_0 x + \beta_1 x}} \qquad (4)$$

The log odds ratio can be written as

$$p/1 - p = \frac{1}{1 + e^{\beta_0 x + \beta_1 x}} = e^{\beta_0 x + \beta_1 x} \qquad (5)$$

which is the equation of logistic regression.

Why logistic regression is a type of regression -
Logistic regression is a supervised learning algorithm used to predict a dependent categorical target variable. In essence, if you have a large set of data that you want to categorise, logistic regression may be able to help.

For example, if you were given a dog and an orange and you wanted to find out whether each of these items was an animal or not, the desired result would be for the dog to end up classified as an animal, and for the orange to be categorised as not an animal. Animal is your target; it is dependent on your data in order to be able to classify the item correctly. In this example, there are only two possible answers (binary logistic regression), animal or not an animal. However, it is also possible to set up your logistic regression with more than two possible categories (multinomial logistic regression).

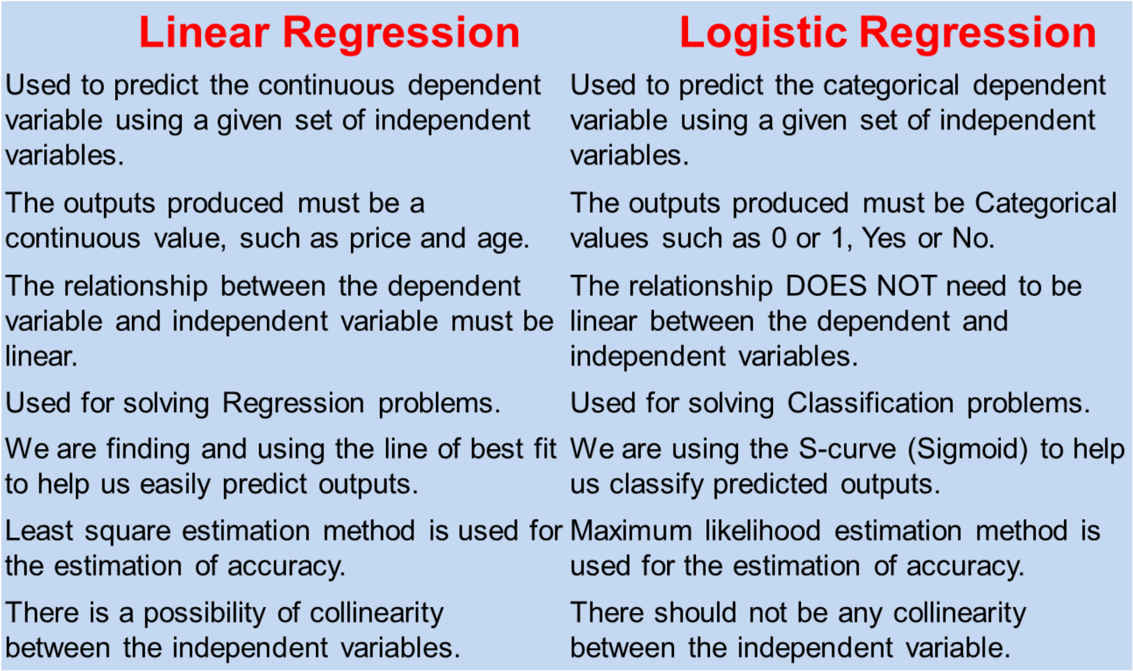
SLR vs Logistic

One of the most important differences between logistic regression and linear regression is in how we compare models. Remember for linear regression we looked at how the adjusted R2 changed. If there was a significant increase when we added another variable (or interaction) then we thought the model had improved.

For logistic regression there are a variety of ways of looking at model improvement. The best way of comparing models is to use something called the likelihood-ratio test.

When we were using OLS(Ordinary Least Squares) regression, we were trying to minimise the sum of squares, for logistic regression we are trying to maximise something called the likelihood function (normally called L).

To see whether our model has improved by adding a variable (or interaction, or squared term), we can compare the maximum of the likelihood function for each model (just like we compared the R2 before for OLS regressions).



Linear Regression	Logistic Regression
Used to predict the continuous dependent variable using a given set of independent variables.	Used to predict the categorical dependent variable using a given set of independent variables.
The outputs produced must be a continuous value, such as price and age.	The outputs produced must be Categorical values such as 0 or 1, Yes or No.
The relationship between the dependent variable and independent variable must be linear.	The relationship DOES NOT need to be linear between the dependent and independent variables.
Used for solving Regression problems.	Used for solving Classification problems.
We are finding and using the line of best fit to help us easily predict outputs.	We are using the S-curve (Sigmoid) to help us classify predicted outputs.
Least square estimation method is used for the estimation of accuracy.	Maximum likelihood estimation method is used for the estimation of accuracy.
There is a possibility of collinearity between the independent variables.	There should not be any collinearity between the independent variable.

Linear regression and logistic regression are both types of regression analysis used for modelling the relationship between one or more predictor variables and an outcome variable. There are some similarities between these two methods, such as:

Both methods involve estimating model coefficients: In linear regression, the coefficients represent the slope and intercept of the linear relationship between the predictor variables and the outcome variable, whereas in logistic regression, the coefficients represent the log odds of the outcome variable.

Both methods use the maximum likelihood estimation: The objective of both methods is to maximize the likelihood of the observed data given the model parameters.

Both methods can be used to make predictions: Once the coefficients are estimated, both methods can be used to make predictions on new data based on the values of the predictor variables.

Both methods can handle multiple predictor variables: Both methods can handle multiple predictor variables and can model complex relationships between the predictor variables and the outcome variable.

However, there are also some key differences between linear and logistic regression. The most significant difference is that linear regression is used for modelling continuous outcome variables, whereas logistic regression is used for modelling binary or categorical outcome variables. Additionally, the form of the relationship between the predictor variables and the outcome variable is different in these two methods: linear regression models a linear relationship, whereas logistic regression models a non-linear relationship between the log odds of the outcome variable and the predictor variables.

MLE - The maximum likelihood function for logistic regression is used to estimate the parameters of the logistic regression model, given a set of observed data. The goal is to find the values of the parameters that maximise the likelihood of the observed data.
Maximum Likelihood Estimation involves treating the problem as an optimization or search problem, where we seek a set of parameters that results in the best fit for the joint probability of the data sample (X).

First, it involves defining a parameter called theta that defines both the choice of the probability density function and the parameters of that distribution. It may be a vector of numerical values whose values change smoothly and map to different probability distributions and their parameters.

In Maximum Likelihood Estimation, we wish to maximise the probability of observing the data from the joint probability distribution given a specific probability distribution and its parameters, stated formally as:

P(X | theta)
In practice, it is often more convenient to maximise the log-likelihood function, Maximising the log-likelihood function is equivalent to maximising the likelihood function, since the logarithm is a monotonic function. Moreover, the log-likelihood function is often easier to work with, since it involves sums rather than products, and the logarithm simplifies the expression of the logistic function.

Logistic Response function - The logistic response function, also known as the logistic function or the sigmoid function, is a mathematical function used in logistic regression to model the relationship between a binary response variable and one or more predictor variables.

The logistic function is defined as:

f(x) = 1 / (1 + e^(-x))

where x is the input to the function. The output of the function, f(x), is a value between 0 and 1, which can be interpreted as the probability that the response variable takes the value 1, given the value of the predictor variable(s).

The logistic function has an S-shaped curve, which starts at 0 when x approaches negative infinity, increases smoothly towards 0.5 as x approaches 0, and then levels off towards 1 as x approaches positive infinity. This curve is useful for modelling binary outcomes that are influenced by one or more predictor variables.

In logistic regression, the logistic response function is used to transform the linear combination of the predictor variables and their coefficients into a predicted probability of the binary response variable taking the value 1. The coefficients are estimated using maximum likelihood estimation, and the predicted probabilities can be used to make predictions about new observations or to calculate measures of model performance such as accuracy or area under the receiver operating characteristic curve.

Generalised linear models
A generalised linear model (GLM) is a statistical framework that extends the linear regression model to handle non-normal distributions and non-continuous response variables. GLMs are used to model the relationship between a response variable and one or more predictor variables.
In a GLM, the response variable is assumed to follow a probability distribution from the exponential family of distributions, such as the normal, binomial, Poisson, or

gamma distribution. The relationship between the predictor variables and the response variable is modelled using a linear predictor function, which is a linear combination of the predictor variables and their associated regression coefficients.

The linear predictor function is then transformed using a link function, which maps the linear predictor to the expected value of the response variable. The link function is typically chosen to match the properties of the response variable and the distribution it follows.
q12
For example, in logistic regression, the response variable is binary (0 or 1) and the link function is the logistic function, which maps the linear predictor to the probability of the response variable taking the value 1. In Poisson regression, the response variable is a count variable and the link function is the logarithmic function, which maps the linear predictor to the expected value of the response variable.

GLMs are a flexible framework that can handle a wide range of response variables and predictor variables, including categorical and continuous variables, and can account for overdispersion and heteroscedasticity. They can also incorporate interaction terms and higher order terms to capture more complex relationships between the response variable and the predictor variables.

GLMs are commonly used in many fields such as biology, engineering, social sciences, and economics, to model a wide range of phenomena, including disease incidence, financial returns, and consumer behaviour.

TIME SERIES ANALYSIS

A **time series** is a collection of observations or data points taken at regular time intervals over a period of time. The time series data can be used to analyse trends, patterns, and seasonal fluctuations over time. Time series analysis is a statistical technique that is used to analyse and model the behaviour of time series data.

applications of time series analysis include:

Economic forecasting: Time series analysis is used to analyse and forecast economic indicators such as GDP, inflation, and unemployment rates.

Financial analysis: Time series analysis is used to analyse financial market trends and forecast stock prices, interest rates, and exchange rates.

Weather forecasting: Time series analysis is used to analyse and forecast weather patterns and climate change.

Sales forecasting: Time series analysis is used to forecast sales of products or services based on historical sales data.

Quality control: Time series analysis is used to monitor and control manufacturing processes, detect anomalies, and predict equipment failures.

Epidemiology: Time series analysis is used to monitor and predict the spread of diseases and outbreaks.

Traffic analysis: Time series analysis is used to analyse and forecast traffic patterns, which is important for city planning and transportation management.

Energy demand forecasting: Time series analysis is used to forecast energy demand based on historical usage patterns, which is important for energy companies and utilities.

The four main components of a time series are:

Trend: This component represents the long-term pattern or direction of the time series. It shows whether the series is increasing, decreasing, or staying relatively constant over time. Various fluctuations but for a longer period of time.

Seasonality: This component represents the regular, repeating patterns that occur in the time series over shorter periods of time, such as days, weeks, or months. Seasonality is often seen in data related to weather, holidays, or other recurring events.

Cyclical: This component represents the longer-term, non-repeating fluctuations in the time series that can last for several years or even decades. Cyclical patterns can be caused by factors such as changes in the business cycle, demographic shifts, or changes in government policies.cycles are often irregular both in height of peak and duration.

Random: This component represents the unpredictable, random fluctuations in the time series that cannot be explained by the other three components. Random variation can be caused by factors such as measurement error, natural disasters, or unexpected events.

Decomposition

Decomposition is a common technique in time series analysis that separates a time series into its different components: trend, seasonality, cyclical, and random. The process of decomposition involves breaking down the original time series into these four components, each of which can be analysed and modelled separately.
The decomposition process typically involves the following steps:

Detrending: The trend component is isolated and removed from the time series. This is typically done using a moving average or polynomial regression to smooth out the data and identify the long-term trend.

Seasonality extraction: The seasonality component is extracted from the detrended time series. This is typically done by analysing the periodic patterns that repeat over shorter time intervals, such as daily, weekly, or monthly cycles.

Cyclical component identification: The cyclical component is identified by analysing longer-term fluctuations in the time series that do not repeat at regular intervals. This can be done using spectral analysis or other advanced techniques.

Residual analysis: The random component, or residual, is identified by subtracting the trend, seasonality, and cyclical components from the original time series. The residual component represents the unpredictable, random fluctuations in the data that cannot be explained by the other components.

Additive time series

In an additive time series, the values of the time series are modelled as a sum of the different components: trend, seasonality, cyclical, and random. That is, the time series values at a given time are equal to the sum of the trend, seasonality, cyclical, and random components at that time. an additive model is more appropriate when the magnitude of the seasonal and trend components are relatively constant over time The additive model is expressed as:

$Y(t) = T(t) + S(t) + C(t) + e(t)$

where Y(t) is the observed value at time t, T(t) is the trend component, S(t) is the seasonal component, C(t) is the cyclical component, and e(t) is the random component.

Multiplicative time series

In a multiplicative time series, the values of the time series are modelled as a product of the different components: trend, seasonality, cyclical, and random. That is, the time series values at a given time are equal to the product of the trend, seasonality, cyclical, and random components at that time.while a multiplicative model is more appropriate when the magnitude of the seasonal and trend components vary with time. The multiplicative model is expressed as:

$Y(t) = T(t) \times S(t) \times C(t) \times e(t)$

where Y(t) is the observed value at time t, T(t) is the trend component, S(t) is the seasonal component, C(t) is the cyclical component, and e(t) is the random component.

Autocorrelation

ACF stands for Autocorrelation Function, which is a **statistical tool** used in time series analysis to **measure** the **correlation** between a **time series** and its **lagged values.** In time series analysis, autocorrelation refers to the **correlation between a time series and its**

own past values. The ACF function **calculates** the **correlation coefficient** between a time series and its lagged values, where the **lag** is the time between the **observation** and its corresponding **lagged value.** The ACF is a plot of the correlation coefficient as a function of the lag. The correlation coefficient ranges from **-1 to 1,** where a value of 1 indicates a **perfect positive correlation** (i.e., the series and its lagged value are identical), **0** indicates **no correlation**, and **-1** indicates a **perfect negative correlation.**

Interpreting the ACF plot is important in time series analysis, as it can provide insights into the underlying patterns in the data. For example:

If the ACF plot shows a **significant correlation at lag 1** (i.e., the first lag), this may indicate the presence of a **strong trend** in the **data.**

If the ACF plot shows a significant **correlation at regular lags** (e.g., lags 12, 24, 36, etc.), this may indicate the **presence of seasonality** in the data.

If the ACF plot shows a **gradual decline in correlation** as the lag increases, this may indicate that the series is stationary and that there is no long-term trend or seasonality.

If the ACF plot shows a sharp decline in correlation at a specific lag, this may indicate the presence of an autoregressive (AR) or moving average (MA) component in the data.

Overall, the ACF plot is a useful tool for understanding the underlying patterns in a time series and identifying potential models that can be used for forecasting and analysis.

PACF

PACF stands for **Partial Autocorrelation Function**, which is another statistical tool used in time series analysis to measure the correlation between a time series and its lagged values, **but with the effects of intermediate lags removed.**

In time series analysis, partial autocorrelation refers to the **correlation between a time series and its own past values, after removing the effects of intermediate lags**. The PACF function calculates the correlation coefficient between a time series and its lagged values, after removing the effects of intermediate lags.

The PACF is a plot of the correlation coefficient as a function of the lag, similar to the ACF plot. However, unlike the ACF plot, which includes the **effects of all intermediate lags, the PACF plot only includes the direct effect of each lag**.

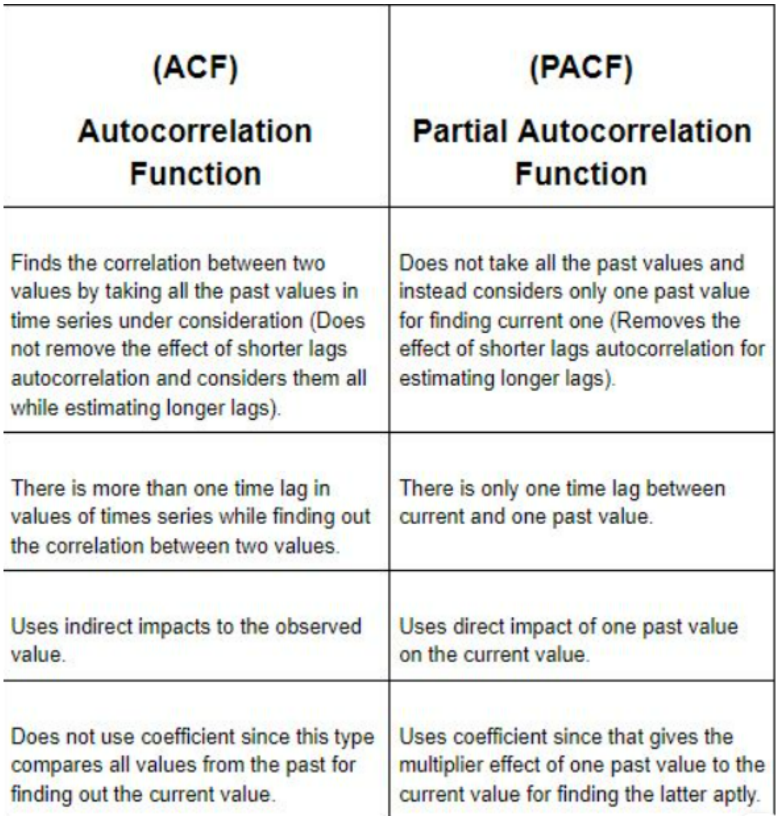
Interpreting the PACF plot is important in time series analysis, as it can provide additional insights into the underlying patterns in the data, beyond what can be seen in the ACF plot. For example:

If the PACF plot shows a significant correlation at lag 1 (i.e., the first lag), this may indicate the presence of a strong autoregressive (AR) component in the data.

If the PACF plot shows a significant correlation at regular lags (e.g., lags 12, 24, 36, etc.), this may indicate the presence of seasonality in the data.

If the PACF plot shows a sharp decline in correlation after a few lags, this may indicate that an AR model of low order (i.e., few lagged terms) is appropriate for modelling the data.

Overall, the PACF plot is a useful tool for understanding the underlying patterns in a time series and identifying potential models that can be used for forecasting and analysis, especially when the ACF plot is not sufficient for identifying the appropriate model.



(ACF) Autocorrelation Function	(PACF) Partial Autocorrelation Function
Finds the correlation between two values by taking all the past values in time series under consideration (Does not remove the effect of shorter lags autocorrelation and considers them all while estimating longer lags).	Does not take all the past values and instead considers only one past value for finding current one (Removes the effect of shorter lags autocorrelation for estimating longer lags).
There is more than one time lag in values of times series while finding out the correlation between two values.	There is only one time lag between current and one past value.
Uses indirect impacts to the observed value.	Uses direct impact of one past value on the current value.
Does not use coefficient since this type compares all values from the past for finding out the current value.	Uses coefficient since that gives the multiplier effect of one past value to the current value for finding the latter aptly.

Evaluation of Time series methods

The evaluation of time series methods involves assessing the performance of various forecasting models to determine which one provides the best fit to the data and produces the most accurate forecasts. There are several commonly used methods for evaluating time series methods, including:

Mean Absolute Error (MAE): This is a measure of the average absolute difference between the actual and forecasted values. It provides a measure of how accurate the forecasts are on average, and lower values indicate better performance.

Root Mean Square Error (RMSE): This is similar to MAE, but it takes into account the squared errors between the actual and forecasted values. RMSE is a popular metric for comparing different forecasting methods, as it emphasises larger errors more than MAE.

Mean Absolute Percentage Error (MAPE): This is a relative measure of forecast accuracy, calculated as the average absolute percentage difference between the actual and forecasted values. MAPE is useful for comparing the accuracy of different forecasting methods across different datasets and scales.

Symmetric Mean Absolute Percentage Error (SMAPE): This is another relative measure of forecast accuracy that takes into account the average of the actual and forecasted values. SMAPE is useful when the actual and forecasted values are close to zero, as it prevents division by zero.

Theil's U-Statistic: This is a measure of the ratio of the RMSE of a given forecasting method to the RMSE of a naive forecasting method (e.g., using the previous value as the forecast). A value less than 1 indicates better performance than the naive method.

Forecast Error Variance Decomposition (FEVD): This is a method for decomposing the variance of the forecast errors into components due to different sources of uncertainty, such as model error, measurement error, and external shocks. FEVD can help identify which sources of uncertainty are most important for a given dataset and forecasting method.

ARIMA model

ARIMA (Autoregressive Integrated Moving Average) models are a class of statistical models used for time series forecasting. ARIMA models capture the autocorrelation, trend, and seasonality in a time series by combining autoregressive (AR), differencing (I), and moving average (MA) components.

The AR component of an ARIMA model captures the relationship between the current value of the time series and its past values, where the degree of relationship is specified by the order of the AR component (denoted by p). Specifically, an AR(p) model predicts the current value of the time series as a linear combination of its past p values, with the coefficients of the linear combination estimated from the data.

The I component of an ARIMA model captures the trend or non-stationarity in the time series by differencing the series over time. Specifically, a differencing order (denoted by d) specifies the number of times the time series needs to be differenced to make it stationary (i.e., the mean and variance of the series are constant over time). For example, if the time series exhibits a linear trend, a first-order difference can remove the trend.

The MA component of an ARIMA model captures the relationship between the current value of the time series and its past forecast errors, where the degree of relationship is specified by the order of the MA component (denoted by q). Specifically, an MA(q) model predicts the current value of the time series as a linear combination of its past q forecast errors, with the coefficients of the linear combination estimated from the data.

The order of the ARIMA model is denoted as (p, d, q), where p is the order of the AR component, d is the order of differencing, and q is the order of the MA component. ARIMA models are often chosen based on the Akaike Information Criterion (AIC), which measures the tradeoff between the goodness-of-fit of the model and the complexity of the model.

ARIMA models can be extended to capture seasonal patterns in the time series by adding seasonal AR, differencing, and MA components. The order of the seasonal ARIMA model is denoted as (p, d, q)x(P, D, Q)s, where P, D, and Q denote the order of the seasonal AR, differencing, and MA components, respectively, and s denotes the number of time periods in a season.

p,d,q in arima models

In ARIMA models, p, d, and q are parameters that determine the order of the AR (Autoregressive), I (Integrated or Differencing), and MA (Moving Average) components, respectively.

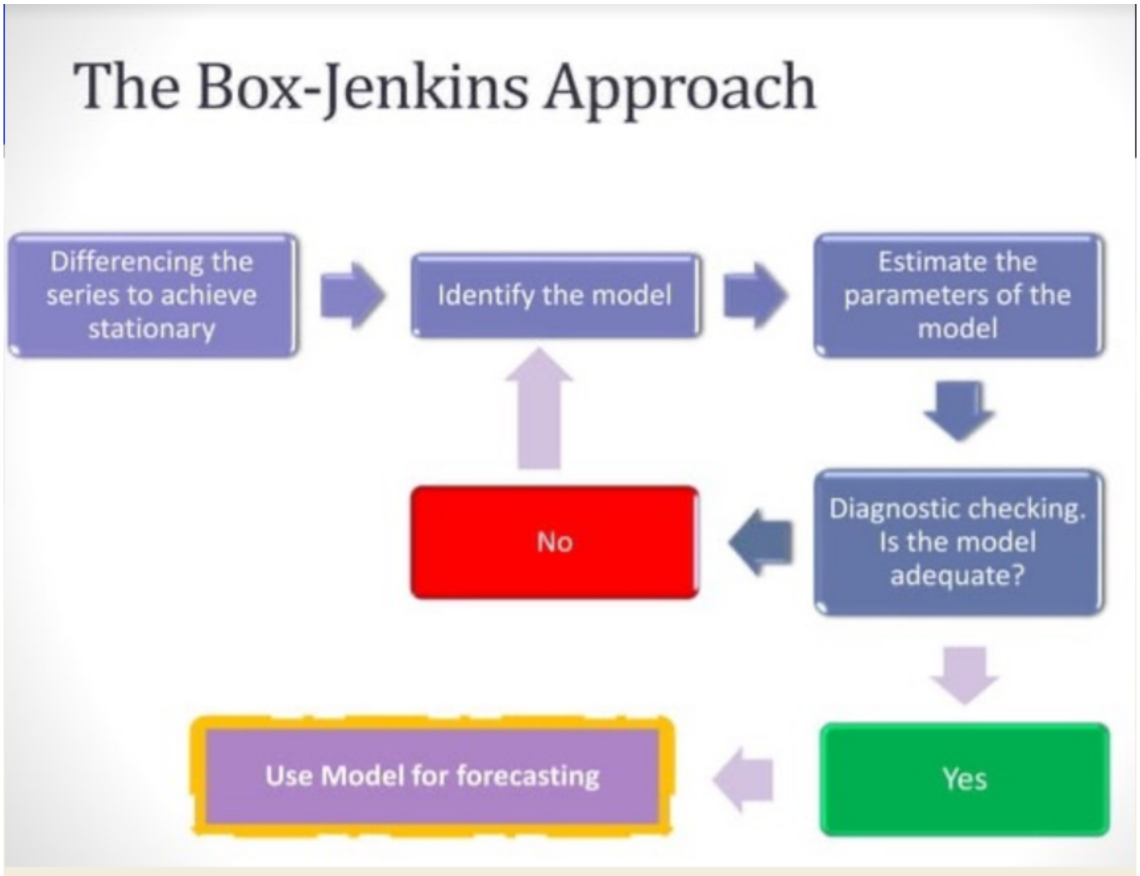
The Autoregressive (AR) component of an ARIMA model captures the relationship between the current value of the time series and its past values. The order of the AR component (p) specifies the number of past values to include in the model. For example, an AR(1) model uses only the lagged value of the time series to make the current prediction, while an AR(2) model uses the two most recent lagged values.

The Integrated (I) or Differencing component of an ARIMA model captures the trend or non-stationarity in the time series. The order of differencing (d) specifies the number of times the time series needs to be differenced to make it stationary, i.e., to make the mean and variance of the series constant over time. For example, if the time series exhibits a linear trend, a first-order difference can remove the trend.

The Moving Average (MA) component of an ARIMA model captures the relationship between the current value of the time series and its past forecast errors. The order of the MA component (q) specifies the number of past forecast errors to include in the model. For example, an MA(1) model uses only the lagged forecast error of the time series to make the current prediction, while an MA(2) model uses the two most recent lagged forecast errors.

The order of an ARIMA model is denoted as (p, d, q), where p is the order of the AR component, d is the order of differencing, and q is the order of the MA component. The choice of the order of an ARIMA model is typically based on visual inspection of the time series, autocorrelation function (ACF) and partial autocorrelation function (PACF) plots, and statistical measures such as the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC).

Jenkins methodology



The Box-Jenkins Approach

Differencing the series to achieve stationary

Identify the model

Estimate the parameters of the model

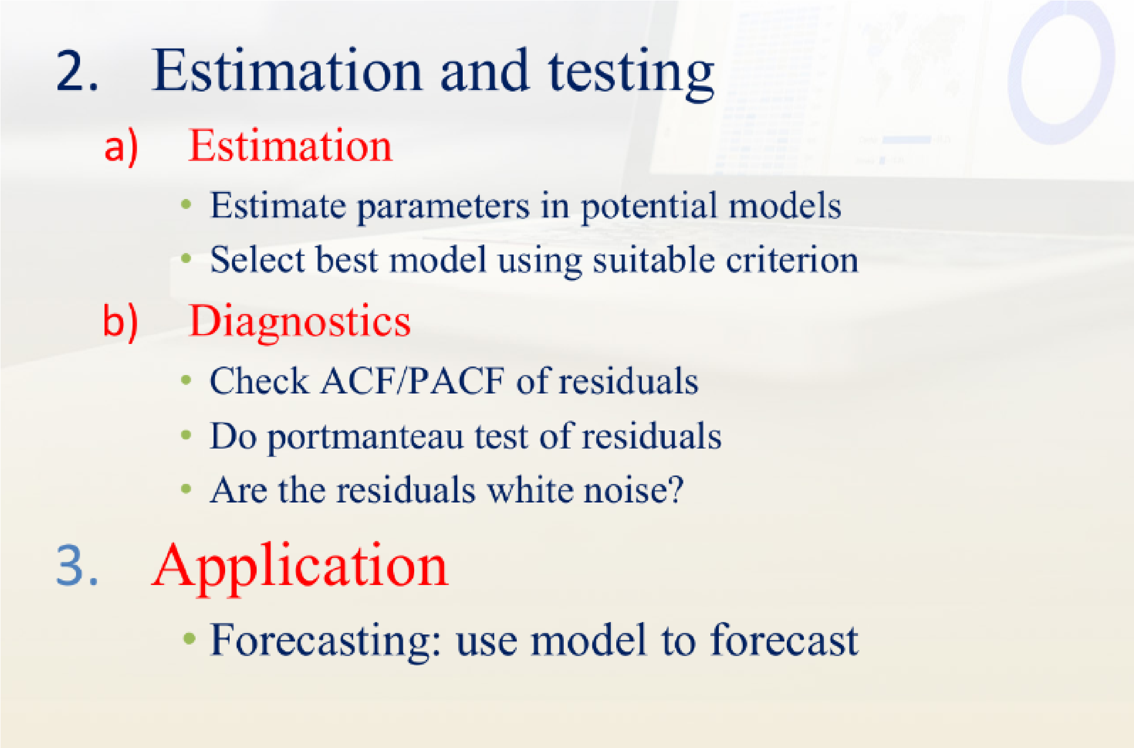
Diagnostic checking. Is the model adequate?

No

Yes

Use Model for forecasting

1. Identification
   a) Data preparation
      – Transform data to stabilize variance
      – Differencing data to obtain stationary series
   b) Model selection
      – Examine data, ACF and PACF to identify potential models



2. Estimation and testing

a) Estimation

- Estimate parameters in potential models
- Select best model using suitable criterion

b) Diagnostics

- Check ACF/PACF of residuals
- Do portmanteau test of residuals
- Are the residuals white noise?

3. Application

- Forecasting: use model to forecast