Regression Analysis

Overview:

- Regression Analysis is a technique to find out the relationship between different variables.
- Regression looks closely into how a dependent variable is affected upon varying an independent variable while keeping the other independent variables constant
- Regression analysis is used for prediction and forecasting.
- Regression analysis is a form of predictive modelling technique which investigates the relationship between a **dependent** (target) and **independent variable (s)** (predictor).

Advantages:

- **Can be used to predict the future:** By using the relevant model to a data set, Regression Analysis can accurately predict a lot of useful information like Stock Prices, Medical Conditions and even Sentiments of the public

Advantages:

- **Can be used to predict the future:** By using the relevant model to a data set, Regression Analysis can accurately predict a lot of useful information like Stock Prices, Medical Conditions and even Sentiments of the public
- **Can be used to back major decisions and policies**: Results from regression analysis adds a scientific backing to a decision or policy and makes it even more reliable as it likelihood of success is then high.

Advantages:

- **Can be used to predict the future:** By using the relevant model to a data set, Regression Analysis can accurately predict a lot of useful information like Stock Prices, Medical Conditions and even Sentiments of the public
- **Can be used to back major decisions and policies**: Results from regression analysis adds a scientific backing to a decision or policy and makes it even more reliable as it likelihood of success is then high.
- **Can correct an error in thinking or disabuse:** Sometimes, an anomaly between the prediction of regression analysis and a decision/thinking can help correct the fallacy of the decision.

Advantages:

- **Can be used to predict the future:** By using the relevant model to a data set, Regression Analysis can accurately predict a lot of useful information like Stock Prices, Medical Conditions and even Sentiments of the public
- **Can be used to back major decisions and policies**: Results from regression analysis adds a scientific backing to a decision or policy and makes it even more reliable as it likelihood of success is then high.
- **Can correct an error in thinking or disabuse:** Sometimes, an anomaly between the prediction of regression analysis and a decision/thinking can help correct the fallacy of the decision.
- **Provides a new perspective**: Large data sets realise their potential to provide new dimensions to a study through the application of Regression Analysis.

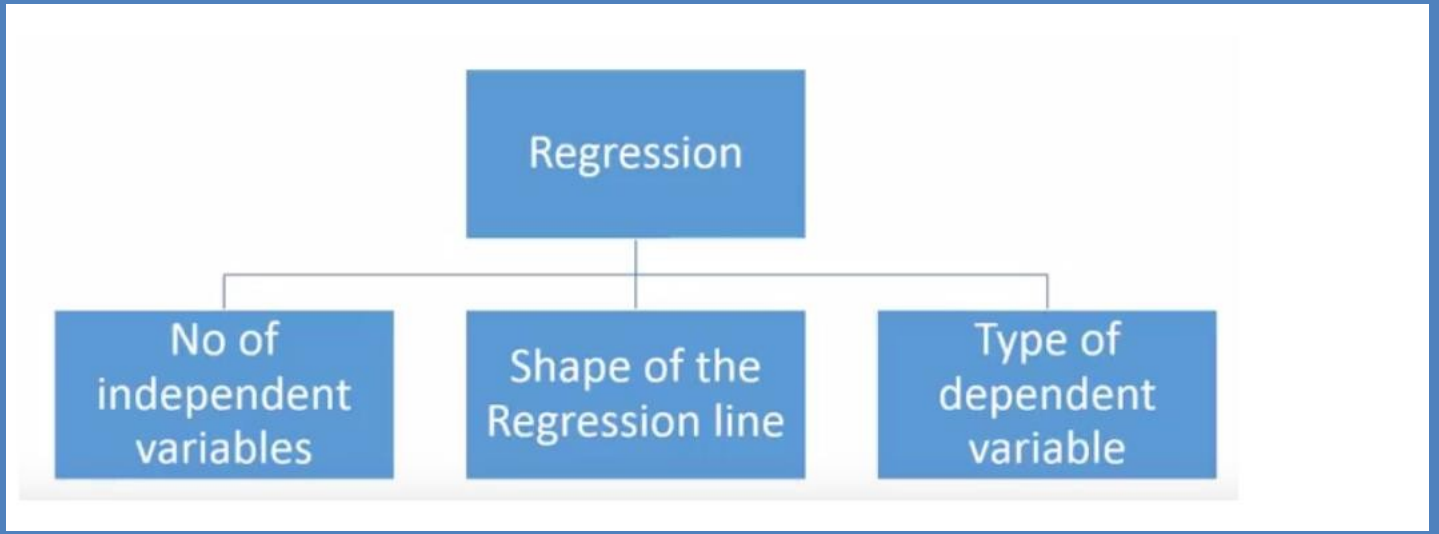
How to select the right regression model:

- Data exploration is an inevitable part of building predictive model. It should be you first step before selecting the right model like identify the relationship and impact of variables
- To compare the goodness of fit for different models, we can analyse different metrics like statistical significance of parameters, R-square, Adjusted r-square, AIC, BIC and error term. Another one is the Mallow's Cp criterion. This essentially checks for possible bias in your model, by comparing the model with all possible submodels (or a careful selection of them).

How to select the right regression model:

- Cross-validation is the best way to evaluate models used for prediction. Here you divide your data set into two group (train and validate). A simple mean squared difference between the observed and predicted values give you a measure for the prediction accuracy.
- If your data set has multiple confounding variables, you should not choose automatic model selection method because you do not want to put these in a model at the same time.
- It'll also depend on your objective. It can occur that a less powerful model is easy to implement as compared to a highly statistically significant model.
- Regression regularization methods(Lasso, Ridge and ElasticNet) works well in case of high dimensionality and multicollinearity among the variables in the data set.

How to select the right regression model:



- Regression is a ***supervised learning technique*** which helps in finding the correlation between variables and enables us to predict the continuous output variable based on the one or more predictor variables. It is mainly used for **prediction, forecasting, time series modeling, and determining the causal-effect relationship between variables**.
- In Regression, we plot a graph between the variables which best fits the given datapoints, using this plot, the machine learning model can make predictions about the data.
- In simple words, ***"Regression shows a line or curve that passes through all the datapoints on target-predictor graph in such a way that the vertical distance between the datapoints and the regression line is minimum."*** The distance between datapoints and line tells whether a model has captured a strong relationship or not.

Some examples of regression can be as:

- Prediction of rain using temperature and other factors
- Determining Market trends
- Prediction of road accidents due to rash driving.

Terminologies Related to the Regression Analysis

- **Dependent Variable:** The main factor in Regression analysis which we want to predict or understand is called the dependent variable. It is also called **target variable**.
- **Independent Variable:** The factors which affect the dependent variables or which are used to predict the values of the dependent variables are called independent variable, also called as a **predictor**.
- **Outliers:** Outlier is an observation which contains either very low value or very high value in comparison to other observed values. An outlier may hamper the result, so it should be avoided.

- **Multicollinearity:** If the independent variables are highly correlated with each other than other variables, then such condition is called Multicollinearity. It should not be present in the dataset, because it creates problem while ranking the most affecting variable.
- **Underfitting and Overfitting:** If our algorithm works well with the training dataset but not well with test dataset, then such problem is called **Overfitting**. And if our algorithm does not perform well even with training dataset, then such problem is called **underfitting**.

- Regression analysis helps in the prediction of a continuous variable. There are various scenarios in the real world where we need some future predictions such as weather condition, sales prediction, marketing trends, etc., for such case we need some technology which can make predictions more accurately. So for such case we need Regression analysis which is a statistical method and used in machine learning and data science. Below are some other reasons for using Regression analysis:
- Regression estimates the relationship between the target and the independent variable.

❖ It is used to find the trends in data.

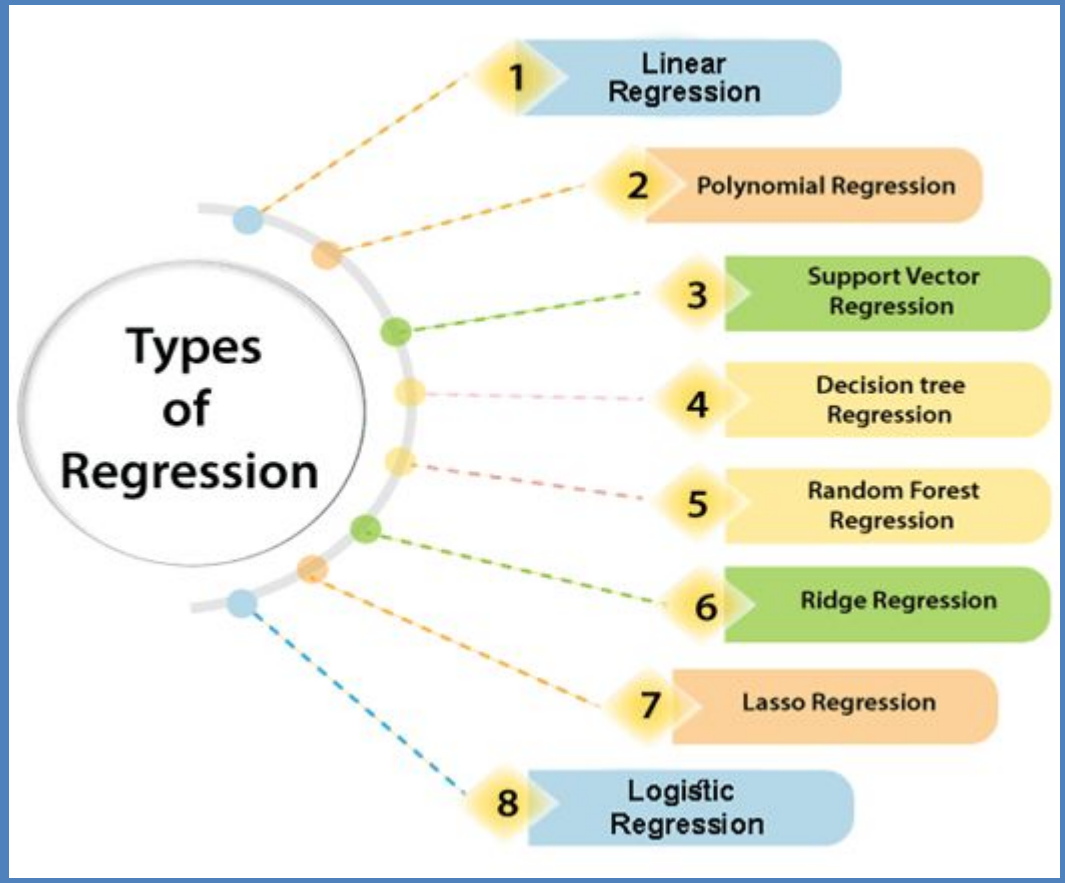
❖ It helps to predict real/continuous values.

❖ By performing the regression, we can confidently determine the **most important factor, the least important factor, and how each factor is affecting the other factors**.

Types of Regression

There are various types of regressions. Each type has its own importance on different scenarios, but at the core, all the regression methods analyze the effect of the independent variable on dependent variables.

- **Linear Regression**
- **Polynomial Regression**
- **Support Vector Regression**
- **Decision Tree Regression**
- **Random Forest Regression**
- **Ridge Regression**
- **Lasso Regression**
- **Logistic Regression**

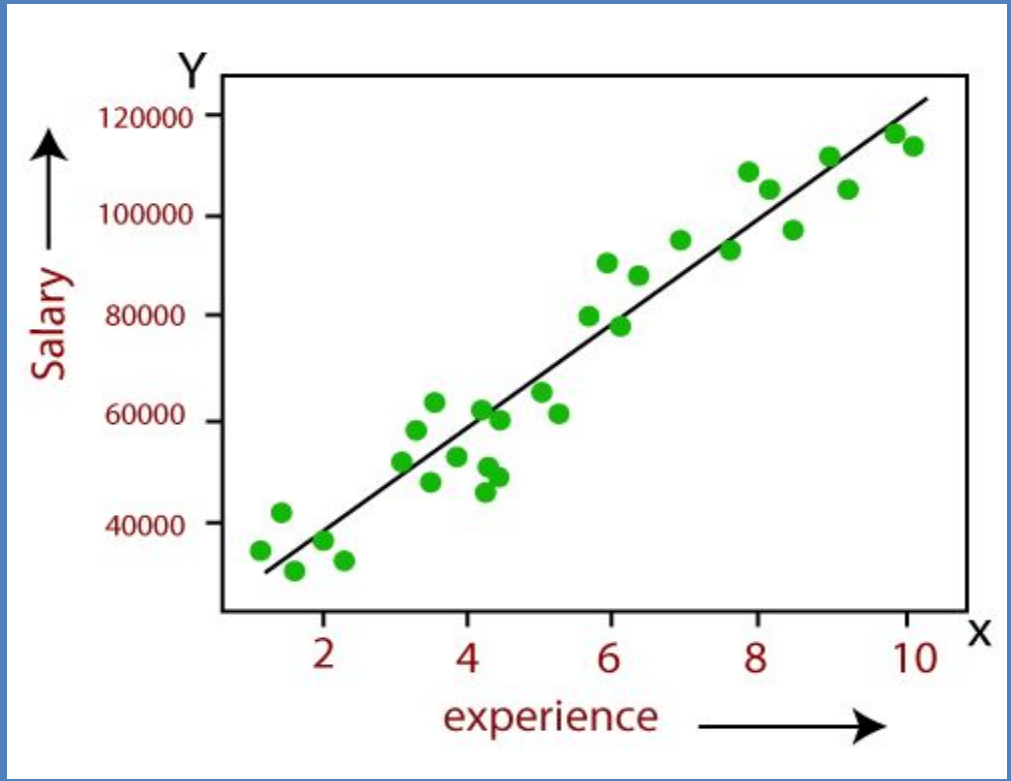

Types of Regression
1 Linear Regression
2 Polynomial Regression
3 Support Vector Regression
4 Decision tree Regression
5 Random Forest Regression
6 Ridge Regression
7 Lasso Regression
8 Logistic Regression

Overview:

- Linear regression is the most simple regression analysis technique. It is the most commonly regression analysis mechanism in predictive analysis
- In this technique, **the dependent variable is continuous**, **independent variable(s) can be continuous or discrete**, and nature of regression line is linear.
- Linear Regression establishes a relationship between **dependent variable (Y)** and one or more **independent variables (X)** using a **best fit straight line** (also known as regression line).
  - It is represented by an equation **Y=a+b*X + e**, where a is intercept, b is slope of the line and e is error term. This equation can be used to predict the value of target variable based on given predictor variable(s).

Linear Regression

Linear regression is a statistical regression method which is used for predictive analysis.
It is one of the very simple and easy algorithms which works on regression and shows the relationship between the continuous variables.
It is used for solving the regression problem in machine learning.
Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), hence called linear regression.
If there is only one input variable (x), then such linear regression is called **simple linear regression**. And if there is more than one input variable, then such linear regression is called **multiple linear regression**.
The relationship between variables in the linear regression model can be explained using the image. Here we are predicting the salary of an employee on the basis of **the year of experience**.



the mathematical equation for Linear regression:

Y= aX+b

Here, Y = dependent variables (target variables),

X= Independent variables (predictor variables),

a and b are the linear coefficients

Some popular applications of linear regression are:

- Analyzing trends and sales estimates
- Salary forecasting
- Real estate prediction
- Arriving at ETAs in traffic

Important notes:

- There must be **linear relationship** between independent and dependent variables
- Multiple regression suffers from **multicollinearity, autocorrelation, heteroskedasticity**.
- Linear Regression is very sensitive to **Outliers**. It can terribly affect the regression line and eventually the forecasted values.
- Multicollinearity can increase the variance of the coefficient estimates and make the estimates very sensitive to minor changes in the model. The result is that the coefficient estimates are unstable
- In case of multiple independent variables, we can go with **forward selection**, **backward elimination** and **step wise approach** for selection of most significant independent variables.
- The difference between simple linear regression and multiple linear regression is that, multiple linear regression has (>1) independent variables, whereas simple linear regression has only 1 independent variable

Task:

Suppose there is a marketing company A, who does various advertisement every year and get sales on that. The below list shows the advertisement made by the company in the last 5 years and the corresponding sales:

Advertisement	Sales
$90	$1000
$120	$1300
$150	$1800
$100	$1200
$130	$1380
$200	??

Sales:-

Logistic Regression

Overview:

- In Logistic Regression, the dependent variable is binary that is it has two values. It can have values like True/False or 0/1 or Yes/No
- This model is used to determine the chance whether a dichotomous outcome depends on one or more free (independent) variables
- Logistic regression is used to find the probability of event=Success and event=Failure. We should use logistic regression when the dependent variable is binary (0/ 1, True/ False, Yes/ No) in nature. Here the value of Y ranges from 0 to 1 and it can represented by following equation.

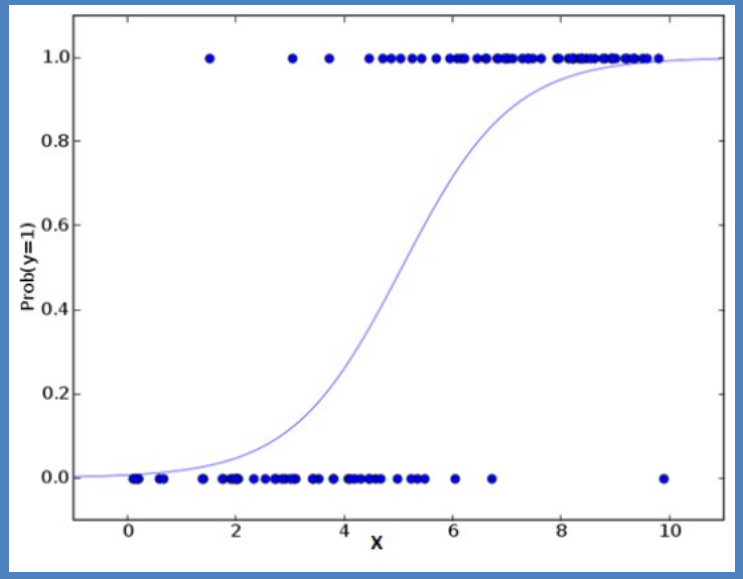
Differences between linear and logistic regressions:

- In Logistic Regression, Conditional Distribution y|x is not a Gaussian distribution but a Bernoulli distribution.
- In Logistic Regression, the predicted outcomes are probabilities determined through logistic function and they are circumscribed between 0 and 1.

Understanding the concept

```
odds= p/ (1-p) = probability of event occurrence / probability of not event occurrence
ln(odds) = ln(p/(1-p))
logit(p) = ln(p/(1-p)) = b0+b1X1+b2X2+b3X3....+bkXk
```

p is the probability of presence of the characteristic of interest.



Understanding the concept

```
odds= p/ (1-p) = probability of event occurrence / probability of not event occurrence
ln(odds) = ln(p/(1-p))
logit(p) = ln(p/(1-p)) = b0+b1X1+b2X2+b3X3....+bkXk
```

p is the probability of presence of the characteristic of interest.

- Since we are working here with a binomial distribution (dependent variable), we need to choose a link function which is best suited for this distribution. And, it is **logit** function. In the equation above, the parameters are chosen to maximize the likelihood of observing the sample values rather than minimizing the sum of squared errors (like in ordinary regression).

Important notes:

- It is widely used for **classification problems**
- Logistic regression doesn’t require linear relationship between dependent and independent variables. It can handle various types of relationships because it applies a non-linear log transformation to the predicted odds ratio
- To avoid over fitting and under fitting, we should include all significant variables. A good approach to ensure this practice is to use a step wise method to estimate the logistic regression

Important notes:

- It requires **large sample sizes** because maximum likelihood estimates are less powerful at low sample sizes than ordinary least square
- The independent variables should not be correlated with each other i.e. **no multi collinearity**. However, we have the options to include interaction effects of categorical variables in the analysis and in the model.
- If the values of dependent variable is ordinal, then it is called as **Ordinal logistic regression**
- If dependent variable is multi class then it is known as **Multinomial Logistic regression**.

Logistic Regression cheatsheet

intuition and maths

logistic regression is a model developed to describe a process influenced from one or more explanatory variables and having a possible outcome enclosed within an upper and a lower bound. It is based on the following formula:

$$f(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$$

in case of a binary response variable is possible to interpret its prediction as the probability of one of the two outcomes showing up

kind of data

as X:

- categorical variables
- continuous variables

as Y:

- boolean/binary variables
- continuos variables included with and upper and a lower bound (opportunely rescaled on the 0-1 range)

how to fit in R

g*lm(y ~ ., family = 'binomial')* fits y against all variables

assumptions

1.absence of autocorrelation in model residuals
2.absence of multicollinearity between variables
3.linear relationship between explanatory variables and log odds

how to test them in R

1.DurbinWatsonTest(glm_object): 0 positive correlation (ko), 2 absence of positive correlation (passed) 4 negative correlation (ko) - *applicable to time series data only*
2.vif(): <=10 passed, >10 ko
3.fitting alternative model showing cubic and quadratic explanatory variables: if this shows being statistically significant the assumption is broken


Thank
You