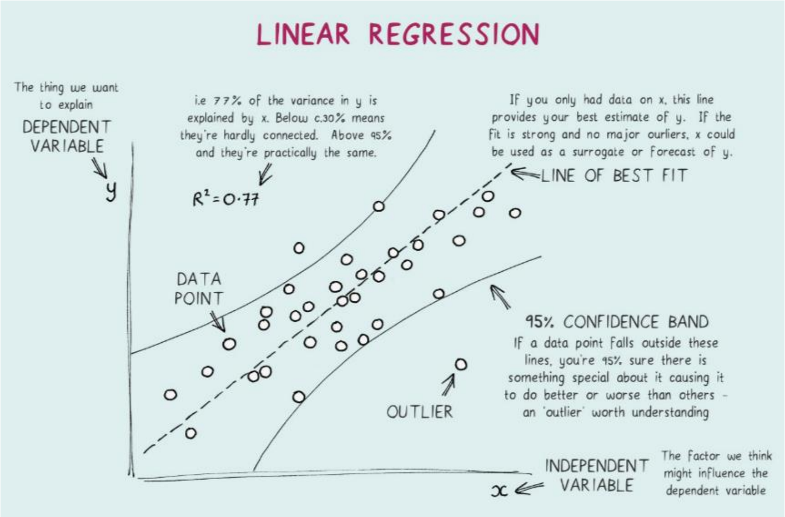
1. **What is Linear Regression?**
   Definition: It is a statistical method that is used for predictive analysis. Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression.
   Linear Regression is a machine learning algorithm based on supervised learning.



Formula:

General Form, $$y = c + b*x$$

Hypothesis Function for Linear Regression,

$$y(x) = \theta_0 + \theta_1 x + \mathcal{E}$$

Once we find the best θ1 and θ2 values, we get the BEST FIT LINE.

Where Θ_0: Intercept
Θ_1: Slope (or Coefficient of x)
X: Independent Variable (or Training Variable)
Y: Dependent Variable (or Target Variable)
Ɛ: Error

Types of Linear Regression

Types	Number of Dependent Variable	Number of Independent Variable
Simple linear regression	1	1
Multiple linear regression	1	2+
Logistic regression	1	2+
Ordinal regression	1	1+
Multinomial regression	1	1+
Discriminant analysis	1	1+

Uses:

(a) Determining The Strength of Predictors, (b) Forecasting an Effect, and (c) Trend Forecasting

a. How to calculate the error in linear regression?

*A high value for the loss means our model performed very poorly. A low value for the loss means our model performed very well.

1. Residual Sum of Squares (RSS):

$$RSS = \sum_{i=1}^{m} (y - \hat{y})^2$$

Here, m = no. of data points
y = actual value
$\hat{y}$ = Predicted value

2. Mean Absolute Error (MAE): MSE is great for ignoring outliers. (Robust to outliners)

$$MAE = \frac{\sum_{i=1}^{m} |y_i - x_i|}{n}$$

Where; n = Total no. of data points
y_i = Prediction
x_i = Actual value

3. Mean Squared Error (MSE): MSE is great for learning outliers.

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

where, n = Total no. of data points
y_i = observed value
$\hat{y}_i$ = Predicted value

OR

$$MSE = \frac{1}{N} \sum_{i=1}^{n} (y_i - (mx_i + b))^2$$

where, N = Total No. of observations (data points)
y_i = Actual value of observation
$(mx_i + b)$ = Predicted value
$\frac{1}{N} \sum_{i=1}^{n}$ = Mean

4. Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \hat{x}_i)^2}{N}}$$

where, n = Total no. of data points
x_i = Actual value
$\hat{x}_i$ = Predicted value

5. Mean Percentage Error (MPE):

MEAN PERCENTAGE ERROR

$$MPE = \frac{\sum_{i=1}^{n} ((y - \hat{y})/y)}{n} \times 100$$

where, n = Total no. of observation
y = Actual value
$\hat{y}$ = Predicted value

6. Mean Absolute Percentage Error (MAPE):

$$MAPE = \frac{\sum_{i=1}^{n} |(y-\hat{y})/y|}{n} \times 100$$

where, n = Total no. of observed value

y = Actual value

$\hat{y}$ = Predicted value.

7. Huber Loss:

   Huber Loss is the combination of MSE and MAE. It is also differentiable at 0. It is basically absolute error, which becomes quadratic when error is small. Huber loss can be helpful in cases, as it curves around the minima which decreases the gradient.

   However, the problem with Huber loss is that we might need to train hyper-parameter delta which is an iterative process.

$$L_\delta(a) = \begin{cases} \frac{1}{2}a^2 & \text{for } |a| \le \delta, \\ \delta(|a| - \frac{1}{2}\delta), & \text{otherwise.} \end{cases}$$

OR

$$L_\delta(y, f(x)) = \begin{cases} \frac{1}{2}(y - f(x))^2 & \text{for } |y - f(x)| \le \delta, \\ \delta \cdot \left(|y - f(x)| - \frac{1}{2}\delta\right), & \text{otherwise.} \end{cases}$$

2. Difference between predicted value and real value?

The actual value is the value that is obtained by observation or by measuring the available data. It is also called the observed value. The predicted value is the value of the variable predicted based on the regression analysis.

The difference between the actual and the predicted value is the residual which is defined as:

$$e = y - \hat{y}$$

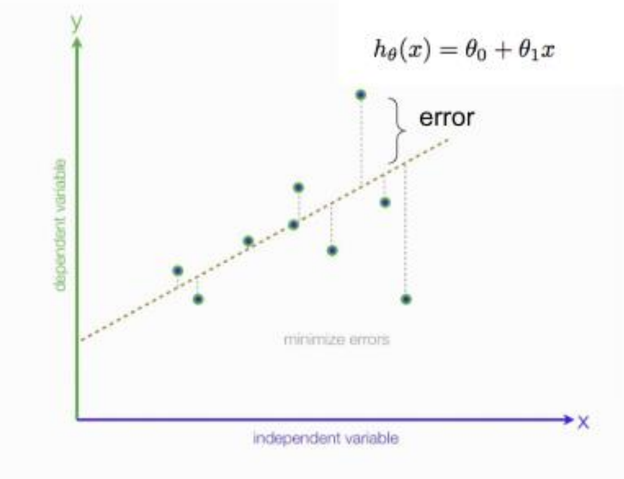
Here, e is the residual, y is the observed or actual value and $\hat{y}$ is the predicted value.

3. Difference between cost and loss function?

Loss Function	**Cost Function**
To calculating loss, we consider only a single data point	To calculating the sum of error for multiple data **(Also known as MSE)**
The loss function is to capture the difference between the actual and predicted values for a single record	Cost functions aggregate the difference for the entire training dataset
For a single training cycle loss is calculated numerous times	But the cost function is only calculated once
There are many different loss functions we can choose from, and each has its advantages and shortcomings. The Most commonly used loss functions are Mean-squared error and Hinge loss.	The cost functions serve two purposes. (a) Its value for the test data estimates our model's performance on unseen objects. That allows us to compare different models and choose the best. (b) We use it to train our models.
Example: For instance, let's say that our model predicts a flat's price (in thousands of dollars) based on the number of rooms, area (m2), floor, and the neighborhood in the city (A or B). Let's suppose	Example: if L is our loss function, then we calculate the cost function by aggregating the loss L over the training, validation, or test

that its prediction for x= [4, 70, 1, A] is USD 110k. If the actual selling price is USD 105k, then the square loss is: L (110, 105) = (110-105)2 = 52 = 25	data $D = \{(x_i, y_i)\}_{i=1}^{n}$. For example, we can compute the cost as the mean loss: $Cost(f,D) = \frac{1}{n}\sum_{i=1}^{n} L(\hat{y}_i, y_i)$ $\left[(\hat{y}_i = f(x_i)) \right.$

Cost Function in single view:



Hypothesis:

$$h_\theta(x) = \theta_0 + \theta_1 x$$

Parameters:

$$\theta_0, \theta_1$$

Cost Function:

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$$

Goal:

$$\underset{\theta_0, \theta_1}{minimize} J(\theta_0, \theta_1)$$

4. **What is Gradient Descent?**

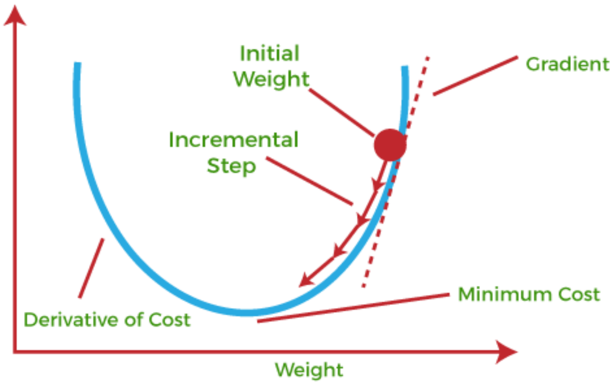
Gradient Descent is an iterative first order optimization algorithm used to find a local minimum/maximum of a given function. This method is commonly used in machine learning and deep learning to minimize a cost/loss function (e.g., in a linear regression).

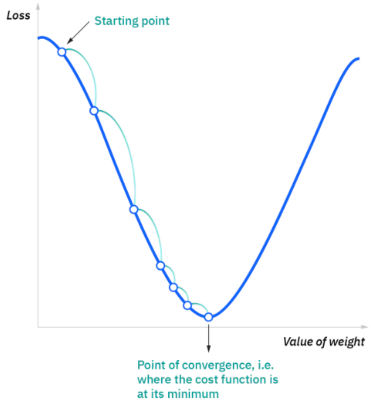
The best way to define the local minimum or local maximum of a function using gradient descent is as follows:

i. If we move towards a negative gradient or away from the gradient of the function at the current point, it will give the local minimum of that function.

ii. Whenever we move towards a positive gradient or towards the gradient of the function at the current point, we will get the local maximum of that function.

This entire procedure is known as Gradient Ascent, which is also known as steepest descent.




Loss
Starting point
Value of weight
Point of convergence, i.e. where the cost function is at its minimum

Function requirements Gradient Descent algorithm does not work for all functions. There are two specific requirements. A function must be;

1. Differentiable: If a function is differentiable, it has a derivative for each point in its domain- not all functions meet these criteria.
2. Convex: For a univariate function, this means that the line segment connecting two function's points lays on or above its curve (it does not cross it). If it does it means that it has a local minimum which is not a global one.

$$Loss = \overbrace{\tfrac{1}{2}((y_i - f(x_i)^2)}^{\text{Quadratic Equation.}} \, , \quad \text{if } |y_i - f(x_i)| < \delta$$

$$\underbrace{\delta |y_i - f(x_i)|}_{\text{Linear Equation}} - \frac{1}{2}\delta^2$$

Gradient: Intuitively it is a slope of a curve at a given point in a specified direction. In the case of **a univariate function**, it is simply the **first derivative at a selected point**. In the case of **a multivariate function**, it is a **vector of derivatives** in each main direction (along variable axes).

5. **Explain how Gradient Decent works in Linear Regression?**

Gradient Descent Algorithm iteratively calculates the next point using gradient at the current position, scales it (by a learning rate) and subtracts obtained value from the current position (makes a step). It subtracts the value because we want to minimise the function (to maximise it would be adding). This process can be written as:

$$p_{n+1} = p_n - \eta \nabla f(p_n)$$

There is an important parameter η which scales the gradient and thus controls the step size. In machine learning, it is called **learning rate** and have a strong influence on performance.

The smaller learning rate the longer GD converges, or may reach maximum iteration before reaching the optimum point

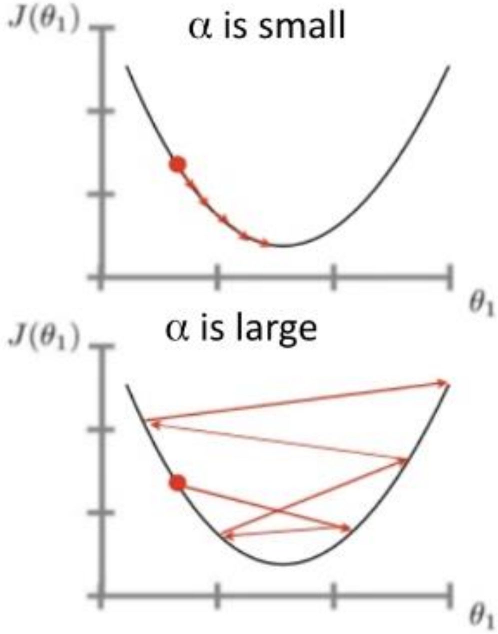
If learning rate is too big the algorithm may not converge to the optimal point (jump around) or even to diverge completely.

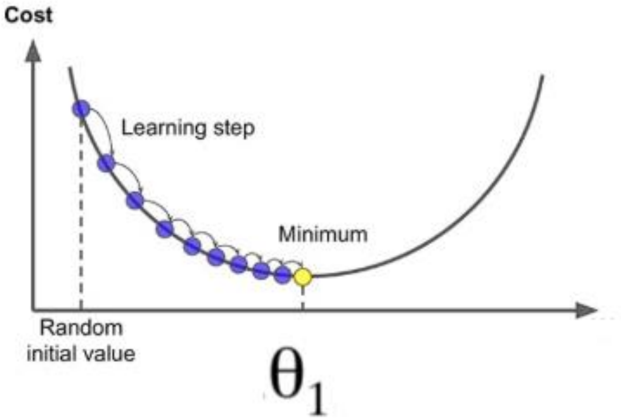
In summary, Gradient Descent method's steps are:

a. choose a starting point (initialisation)
b. calculate gradient at this point
c. make a scaled step in the opposite direction to the gradient (objective: minimise)
d. repeat points 2 and 3 until one of the criteria is met:
e. maximum number of iterations reached
f. step size is smaller than the tolerance (due to scaling or a small gradient).

Learning Rate: Learning Rate is the size of the steps that are taken to reach the minimum. This is typically a small value, and it is evaluated and updated based on the behaviour of the cost function. High learning rates result in larger steps but risks overshooting the minimum.

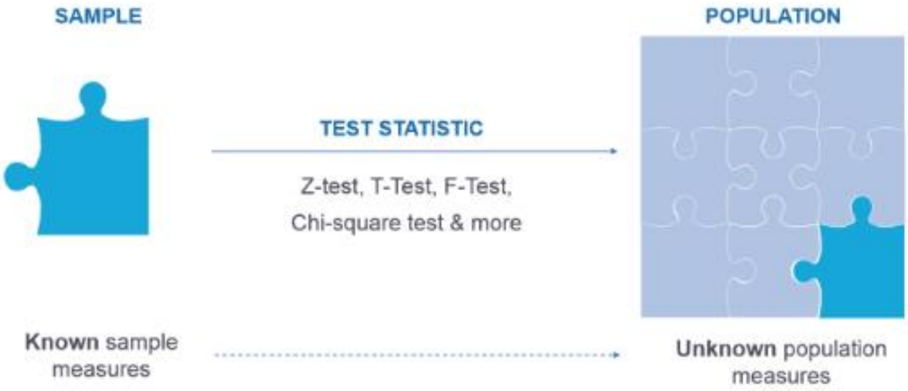
```
repeat until convergence {
    $\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$
        (for $j = 1$ and $j = 0$)
}
```



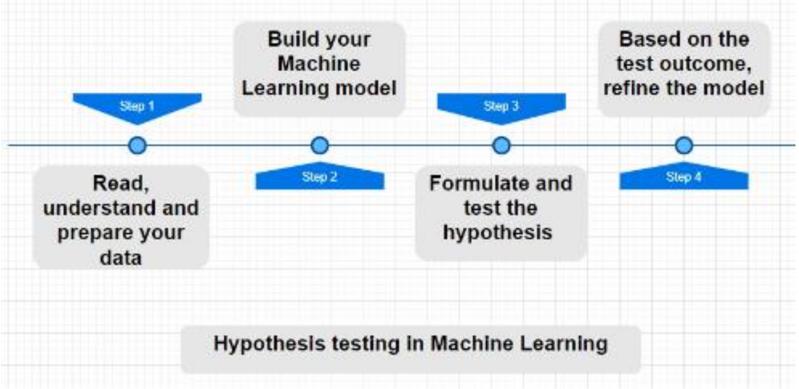


6. **What is the Intercept Term? (y=mx+c)**
Intercept means a line crosses an axis.
Intercept is of two types
X-Intercept: It is the line where a line crosses X-axis. At this point Y coordinates will be zero.
Y-Intercept: It is the line where a line crosses Y-axis. At this point X coordinates will be zero.

7. **How is hypothesis testing used in linear regression?**
Any data science project starts with exploring the data. When we perform an analysis on a sample through exploratory data analysis and inferential statistics, we get information about the sample.
Hypothesis testing is done to confirm our observation about the population using sample data, within the desired error level.



How to perform Hypothesis Testing: To trust your model and make predictions, we utilize hypothesis testing. When we will use sample data to train our model, we make assumptions about our population. By performing hypothesis testing, we validate these assumptions for a desired significance level.


Step 1
Read, understand and prepare your data
Step 2
Build your Machine Learning model
Step 3
Formulate and test the hypothesis
Step 4
Based on the test outcome, refine the model
Hypothesis testing in Machine Learning

Steps to perform hypothesis test are as follows:

1. Formulate a Hypothesis
2. Determine the significance level
3. Determine the type of test
4. Calculate the Test Statistic values and the p values
5. Make Decision

1. **Formulating a Hypothesis:**
   The null hypothesis represented as H_0 is the initial claim that is based on the prevailing belief about the population.
   The alternate hypothesis represented as H_1 is the challenge to the null hypothesis. It is the claim which we would like to prove as True

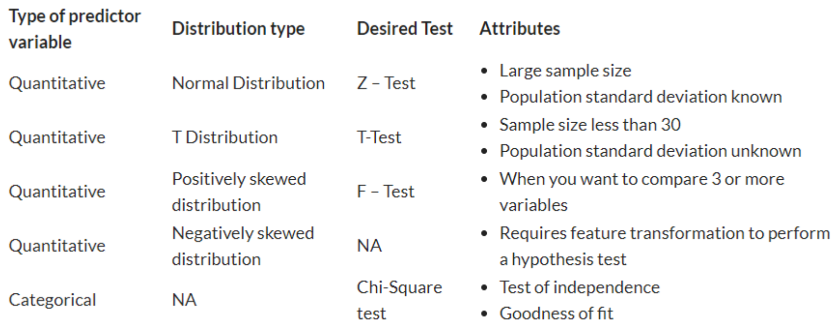
$$\textbf{Null Hypothesis } (\boldsymbol{H_0})\textbf{:}\ \beta_1 = 0$$

$$\textbf{Alternate Hypothesis } (\boldsymbol{H_A})\textbf{:}\ \beta_1 \ \neq 0$$

2. **Determine the significance level:**
   we set the Significance level at 10%, 5%, or 1%.
   The significance level is the proportion of the sample mean lying in critical regions.

3. **Determine the type of test:**

We choose the type of test statistic based on the predictor variable – quantitative or categorical.



Type of predictor variable	Distribution type	Desired Test	Attributes
Quantitative	Normal Distribution	Z – Test	• Large sample size • Population standard deviation known
Quantitative	T Distribution	T-Test	• Sample size less than 30 • Population standard deviation unknown
Quantitative	Positively skewed distribution	F – Test	• When you want to compare 3 or more variables
Quantitative	Negatively skewed distribution	NA	• Requires feature transformation to perform a hypothesis test
Categorical	NA	Chi-Square test	• Test of independence • Goodness of fit

4. **Calculate the Test Statistic values and the p values:**
   Now Calculate the **p-value** from the cumulative probability for the test.

5. **Make Decision:**
   p-value < **0.05**, we can reject the null hypothesis.
   p-value>**0.05**, we fail to reject the null hypothesis.

8. **What are the Hypothesis Tests?**

   a. **Z-statistic – Z Test:**
   Z-statistic is used when the sample follows a normal distribution. It is calculated based on the population parameters like mean and standard deviation.
   One sample Z test is used when we want to compare a sample mean with a population mean
   Two sample Z test is used when we want to compare the mean of two samples

   b. **T-statistic – T-Test:**
   T-statistic is used when the sample follows a T distribution and population parameters are unknown. T distribution is similar to a normal distribution, it is shorter than normal distribution and has a flatter tail.
   If the sample size is less than 30 and population parameters are not known, we use T distribution. Here also, we can use one Sample T-test and a two-sample T-test.

   c. **F-statistic – F test:**
   For samples involving three or more groups, we prefer the F Test. Performing T-test on multiple groups increases the chances of Type-1 error. ANOVA is used in such cases.
   Analysis of variance (ANOVA) can determine whether the means of three or more groups are different. ANOVA uses F-tests to statistically test the equality of means.
   F-statistic is used when the data is positively skewed and follows an F distribution. F distributions are always positive and skewed right.
   F = Variation between the sample means/variation within the samples
   For negatively skewed data we would need to perform feature transformation

   d. **Chi-Square Test:**
   For categorical variables, we would be performing a chi-Square test.
   Following are the two types of chi-squared tests:
   Chi-squared test of independence – We use the Chi-Square test to determine whether there is a significant relationship between two categorical variables.
   Chi-squared Goodness of fit helps us determine if the sample data correctly represents the population.

9. R-Squared Vs Adjusted R Squared?

R Squared is used to determine the strength of correlation between the predictors and the target. In simple terms it lets us know how good a regression model is when compared to the average.

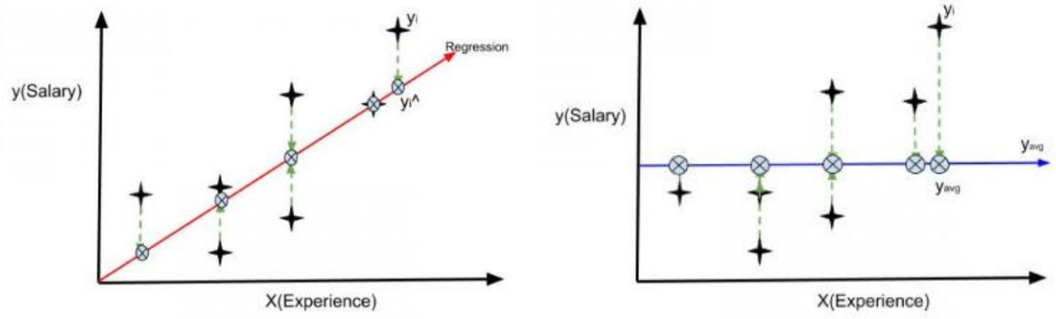
$$R^2 = 1 - \frac{SSR}{SST}$$

$$= 1 - \frac{\Sigma (y_i - \hat{y}_i)^2}{\Sigma (y_i - \bar{y})^2}$$

Where;

SSR: Sum of Squares of Residuals, the sum of the squares of the difference between the actual observed value (y) and the predicted value (y^).

SST: Total Sum of Squares, the sum of the squares of the difference between the actual observed value (y) and the average of the observed y value (y avg)



R Squared will help us determine the best fit for a model. The closer R Squared is to one the better the regression is.

Adjusted R Squared: For a multiple regression model, R-squared increases or remains the same as we add new predictors to the model, even if the newly added predictors are independent of the target variable and don't add any value to the predicting power of the model. Adjusted R-squared eliminates this drawback of R-squared. It only increases if the newly added predictor improves the model's predicting power. Adding independent and irrelevant predictors to a regression model results in a decrease in the adjusted R-squared.

$$R^2_{adj} = 1 - \frac{(1 - R^2)(n - 1)}{n - p - 1}$$

$where:$

$R^2 = R - squared$

$n \quad = number\ of\ samples/rows\ in\ the\ data\ set$

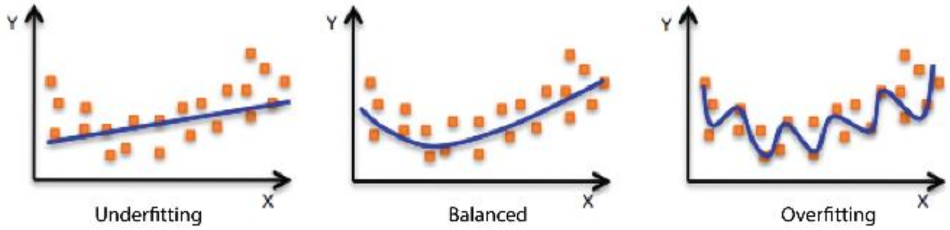
$p \quad = number\ of\ predictors/features$

Overfitting: Is a high R-squared good?

If the training set's R-squared is higher and the R-squared of the validation set is much lower, it indicates overfitting. If the same high R-squared translates to the validation set as well, then we can say that the model is a good fit.

Underfitting: Is a low R-squared bad?

This completely depends on the type of the problem being solved. In some problems which are hard to model, even an R-squared of 0.5 may be considered a good one. There is no rule of thumb to confirm the R-squared to be good or bad. However, a very low R-squared indicates underfitting and adding additional relevant features or using a complex model might help.



Degree of Freedom:

It is the minimum number of independent coordinates that can specify the position of the system completely. In this context, we can define it as the minimum number of data points or observations required to generate a valid regression model.

$$degree\ of\ freedom = n - k - 1$$

Where, K: The number of independent variables
N: The number of observations

Relationship between DOF & Adjusted R^2:

$$Adj.R^2 = 1 - (1 - R^2)\frac{n-1}{n-k-1}$$

10. What is Convergence Algorithm?

An iterative algorithm is said to converge when as the iterations proceed the output gets closer and closer to a specific value. In some circumstances, an algorithm will diverge; its output will undergo larger and larger oscillations, never approaching a useful result.

A model converges when additional training will not improve the model.

11. What are the assumptions made in linear regression?

The Linear Regression has five key assumptions;

a. Linear Relationship (Linearity): Linear Relationship is a correlation between two variables which describes how much one variable change as related to change in the other variable.
b. Homoscedasticity: The variance of residual is the same for any value of X.
c. Independence: Observations are independent of each other.
d. Normality: For any fixed value of X, Y is normally distributed.

12. What is Ridge Regression?

Ridge regression is a method of estimating the coefficients of multiple-regression models in scenarios where the independent variables are highly correlated.

This method performs L2 regularization

It is particularly useful to mitigate the problem of multicollinearity in linear regression, which commonly occurs in models with large numbers of parameters.

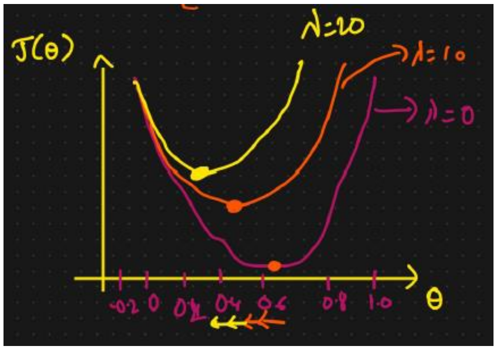
In general, the method provides improved efficiency in parameter estimation problems in exchange for a tolerable amount of bias.

$$\text{Cost Function} = \frac{1}{m}\sum_{i=1}^{m}\left(h_\theta(x)^i - y^i\right)^2 + \underbrace{\boxed{\lambda \sum_{i=1}^{m} (\text{Slope})^2}}_{\text{HYPERPARAMETER}}$$

If Lambda increase Θ will decrease and vice-versa.

Lambda is penalty term.

- It shrinks the parameters. Therefore, it is used to prevent multicollinearity
- It reduces the model complexity by coefficient shrinkage

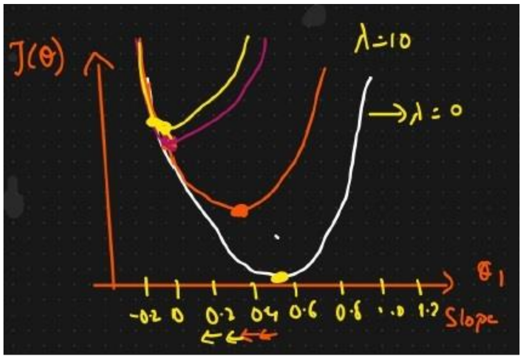


13. What is Ridge Regression?

Lasso (least absolute shrinkage and selection operator) is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the resulting statistical model.

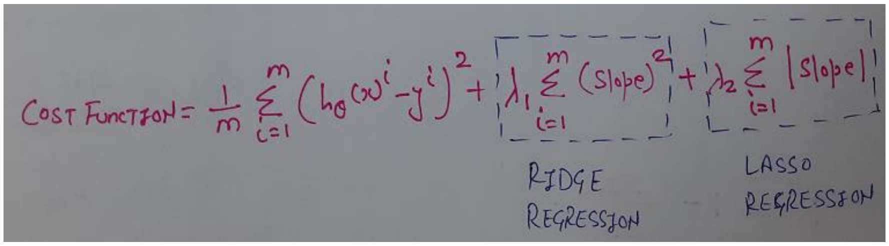
It reduces the features to and helps in the feature selection.

$$\text{COST FUNCTION} = \frac{1}{m}\sum_{i=1}^{m}\left(h_\theta(x)^i - y^i\right)^2 + \lambda \sum_{i=1}^{m} |\text{Slope}|$$



14. What is Elastic Net Regression?

Elastic net linear regression uses the penalties from both the lasso and ridge techniques to regularize regression models. The technique combines both the lasso and ridge regression methods by learning from their shortcomings to improve the regularization of statistical models.

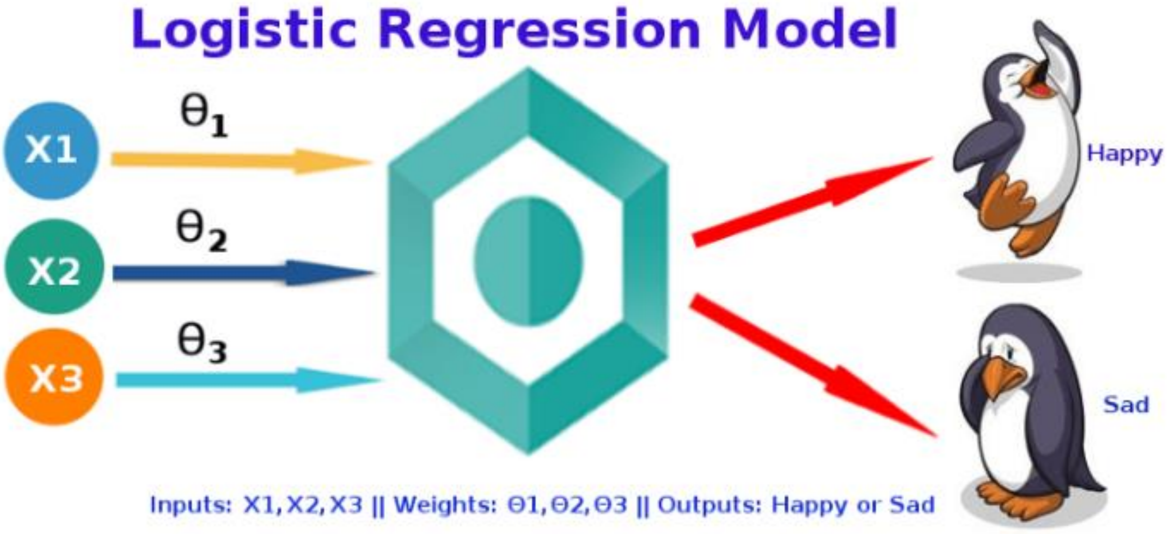


$$\text{COST FUNCTION} = \frac{1}{m}\sum_{i=1}^{m}\left(h_\theta(x)^i - y^i\right)^2 + \underbrace{\boxed{\lambda_1 \sum_{i=1}^{m}(\text{slope})^2}}_{\text{RIDGE REGRESSION}} + \underbrace{\boxed{\lambda_2 \sum_{i=1}^{m}|\text{slope}|}}_{\text{LASSO REGRESSION}}$$

15. What is Logistic Regression?

Logistic Regression is used when the dependent variable(target) is categorical.
This type of statistical model is often used for classification and predictive analytics.
For example,

- To predict whether an email is spam (1) or (0)
- Whether the tumor is malignant (1) or not (0)


Logistic Regression Model
X1
X2
X3
θ1
θ2
θ3
Happy
Sad
Inputs: X1,X2,X3 || Weights: Θ1,Θ2,Θ3 || Outputs: Happy or Sad