Linear Regression

* Supervised Learning - has dependent variable
* Regression Problem - Dependent Variable is continuous
* Linear model e.g. a model that assumes a linear relationship between the input variables (x) and the single output variable (y)
* Output variable (y) can be calculated from a linear combination of input variables (x)
* Simple method

SIMPLE LINEAR REGRESSION - Single input variable

MULTIPLE LINEAR REGRESSION - Multiple Input Variables

ORDINARY LEAST SQUARES - Most common technique to prepare/train linear regression equation from the data

COEFFICIENT - One scale factor assigned to each input value or column

INTERCEPT (bias coefficient) - additional coefficient - gives the line an additional degree of freedom

One input - Equation of a line -

y = Output

B0 = Intercept

B1 = Coefficient

x = Input

More than one input - plane or hyperplane

Cost Function - squared error =

# 4 Techniques to prepare a Linear Regression Model

1. Simple Linear Regression
2. Ordinary Least Squares
3. Gradient Descent
4. Regularization
5. Lasso Regression
6. Ridge Regression

#### More than one input - Ordinary Least Squares or Gradient Descent

## Simple Linear Regression

* Single input
* Use statistics to estimate the coefficients
* Calculate statistical properties from the data - mean, standard deviation, correlation, covariance

## Ordinary Least Squares

Minimize the sum of squared residuals

## Gradient Descent

Optimize values of coefficients by iteratively minimizing the error of the model (cost function) on training data

This process is repeated until

* a minimum sum squared error is achieved or
* no further improvement is possible.

LEARNING RATE (alpha) - scale factor - parameter that determines the size of improvement step to take on each iteration of the procedure.

USED WHEN the dataset is very large.

## Regularization

Extensions of the training of the linear model called Regularization Methods.

* minimizes the sum of squared error of model on training data
* reduce the complexity of the model

USED WHEN we have collinearity in input values and OLS will overfit the training data.

## Metric used to evaluate the performance of the model

R-SQUARED

* Ratio of the explained variance and total variance
* Tells how close the points are to the regression line

SSE - sum of squared error - How much the target value varies around the regression line (predicted value)

SSR - regression sum of squares - Gives information about how far estimated regression line is from the horizontal ‘no relationship’ line (average of actual output)

SST - total sum of squares - Tells how much the data point move around the mean

* r2 always lies between 0 and 1
* r2 = 0 - unable to capture all variance in the dependent variable
* r2 = 1 - able to capture all variance in the dependent variable
* DRAWBACK - as we increase the number of points, r2 value increases
* The R squared value never decreases with the addition of a new feature. It either increases or remains the same. So, R squared cannot tell about variable importance.

## Evaluation Metrics

RMSE (Root Mean Square Error)

* Better model will have lower RMSE

# Lasso Regression

COST FUNCTION:

Lasso Regression - leads to zero coefficients - which means some features are completely neglected

Therefore, Lasso Regression not only reduces over-fitting but can also help in feature selection.