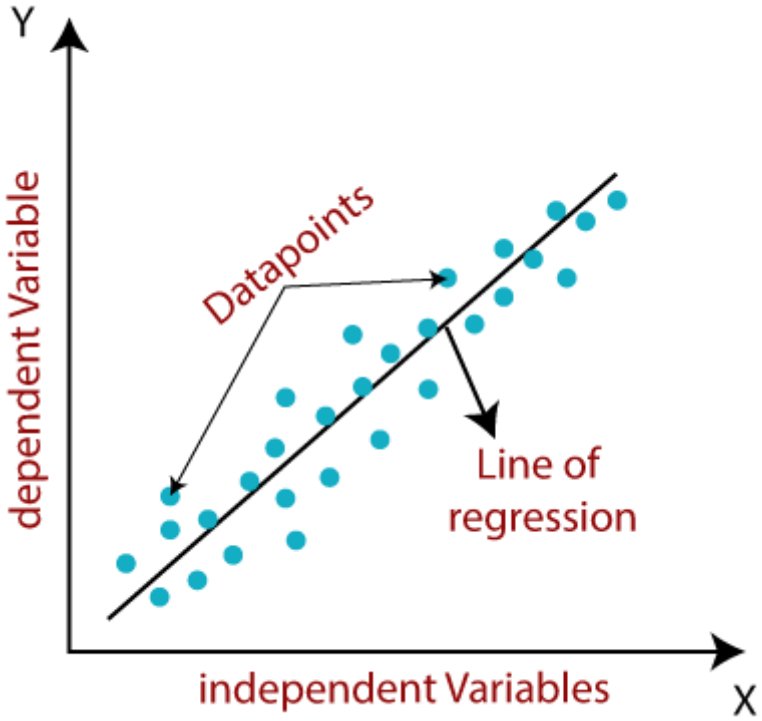
Velocity Corporate Training Center, Pune

Assignment By: Harshal Sunil Kshatriya (Roll no:029)

Topic: Linear Regression Algorithm

Linear Regression

- Linear regression is a statistical model that allows to explain a dependent variable y based on variation in one or multiple independent variables.
- It does this, based on linear relationships between the independent and dependent variables.
- Linear regression makes predictions for continuous/real or numeric variables such as sales, salary, age, product price, etc.



Assumptions of Linear Regression

These are some formal checks while building a Linear Regression model, which ensures to get the best possible result from the given dataset.

1. Linear relationship between the features and target:
   - Linear regression assumes the linear relationship between the dependent and independent variables.
2. Small or no multicollinearity between the features:
   - Multicollinearity means high-correlation between the independent variables.
   - Due to multicollinearity, it may difficult to find the true relationship between the predictors and target variables. Or we can say, it is difficult to determine which predictor variable is affecting the target variable and which is not.

- So, the model assumes either little or no multicollinearity between the features or independent variables.

3. Normal distribution of error terms:
   - Linear regression assumes that the error term should follow the normal distribution pattern.
   - If error terms are not normally distributed, then confidence intervals will become either too wide or too narrow, which may cause difficulties in finding coefficients.
   - It can be checked using the distribution plot or q-q plot.
4. Homoscedasticity Assumption:
   - Homoscedasticity is a situation when the error term is the same for all the values of independent variables.
   - With homoscedasticity, there should be no clear pattern distribution of data in the scatter plot.

Types of Linear Regression

Linear regression can be further divided into two types of the algorithm:

1. Simple Linear Regression:
   - If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.
   - Mathematically, we can represent a simple linear regression as:

$$y = mx + c$$

where,

$$y = \text{dependent variable} \quad ; \quad x = \text{single independent variable}$$

$$m = \text{slope of regression line} \quad ; \quad c = \text{y-intercept}$$

- The values for x and y variables are training datasets for Linear Regression model representation.

2. Multiple Linear regression:
   - If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.
   - Mathematically, we can represent a simple linear regression as:

$$y = m_1x_1 + m_2x_2 + \cdots \ldots + m_nx_n + c$$

where,

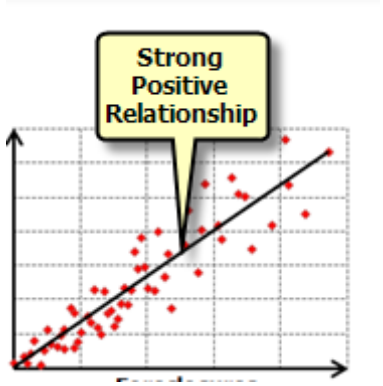
$$y = \text{dependent variable} \quad ; \quad x_1 = 1^{st} \text{ independent variable}$$

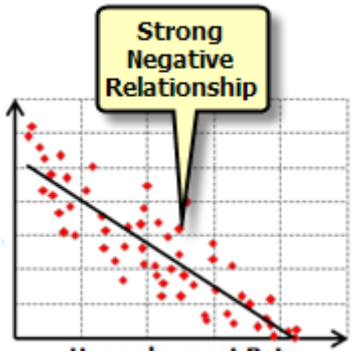
$$x_2 = 2^{nd} \text{ independent variable} \quad ; \quad x_n = n^{th}/\text{last independent variable}$$

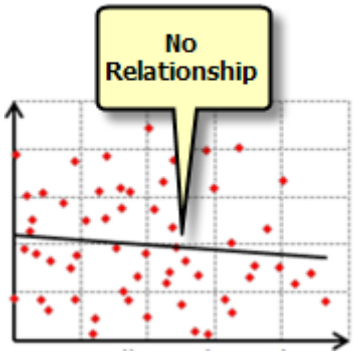
$$m_1, m_2, \ldots \ldots, m_n = \text{slopes} \quad ; \quad c = \text{y-intercept}$$

Regression Line possible orientations

- A linear line showing the relationship between the dependent and independent variables is called a regression line.
- In this session we will see orientations of line.
- A regression line can show two types of relationship:
  1. Positive Linear Relationship:
     - If the dependent variable increases on the Y-axis and independent variable increases on X-axis, then such a relationship is termed as a Positive linear relationship.
  2. Negative Linear Relationship:
     - If the dependent variable increases on the Y-axis and independent variable decreases on X-axis and vice versa, then such a relationship is termed as a Negative linear relationship.
  3. No Relationship:
     - If the best fit line is flat (not sloped) then we can say that there is no relationship among the variables.
     - It means there will be no change in our dependent variable (y) by increasing or decreasing our independent variable (x) value.







Now how do we know what kind of relationship these variables have? Well by using correlation or covariance we can see what type of relationship is there.

- Covariance:

  - It tells us the direction of the relationship between X and Y but it doesn't tell us how positive or negative the relationship is.
  - If the covariance value is negative then we can say that if our independent variable (X) increases then our dependent variable (Y) decreases and vice versa.
- Correlation:
  - It is a statistical measure that tells us the direction of the relationship as well as the strength of the relationship (how much positive the variables are correlated, how much negative the variables are correlated).
  - The range of correlation is between -1<correlation< +1. It will be called perfect correlation if all the points fall on the best fit line.
  - Good Predictors: If R < -0.7 or R > 0.7
  - Bad Predictors >> If -0.3 < R < 0.3
  - Mathematical Formula is

$$Coefficient\ of\ Correlation, R = \frac{covariance(x,y)}{Standerd\ Deviation\ of\ x * Standerd\ Deviation\ of\ y}$$

$$Coefficient\ of\ Correlation, R = \frac{cov(x,y}{\sigma_x * \sigma_y}$$

$$Coefficient\ of\ Correlation, R = \frac{\sum(x_i - x_{mean})\,(y_i - y_{mean})}{\sqrt{\sum(x_i - x_{mean})^2} * \sqrt{\sum(y_i - y_{mean})^2}}$$

Finding the best fit line:

- When working with linear regression, our main goal is to find the best fit line that means the error between predicted values and actual values should be minimized.
- Also the best fit line will have the least error and passes through maximum number of datapoints
- The different values for the coefficient of lines (m, c) gives a different line of regression, so we need to calculate the best values for m and c to find the best fit line, so to calculate this we use cost function.

Cost function

- The different values for coefficient of lines (m, c) gives the different line of regression, and the cost function is used to estimate the values of the coefficient for the best fit line.
- Cost function optimizes the regression coefficients. It measures how a linear regression model is performing.
- We can use the cost function to find the accuracy of the mapping function, which maps the input variable to the output variable. This mapping function is also known as Hypothesis function.
- For Linear Regression, we use the Mean Squared Error (MSE) cost function, which is the average of squared error occurred between the predicted values and actual values.
- It can be written as:

$$Cost\ Function, MSE = \frac{1}{n}\sum_{i=0}^{n}(y_i - y_{pred})^2$$

$$Cost\ Function, MSE = \frac{1}{n}\sum_{i=0}^{n}[y_i - (mx_i + c)]^2$$

$where,$

$n = Total\ number\ of\ observations$

$y_i = Actual\ dependent\ values \quad ; \quad x_i = Actual\ independent\ values$

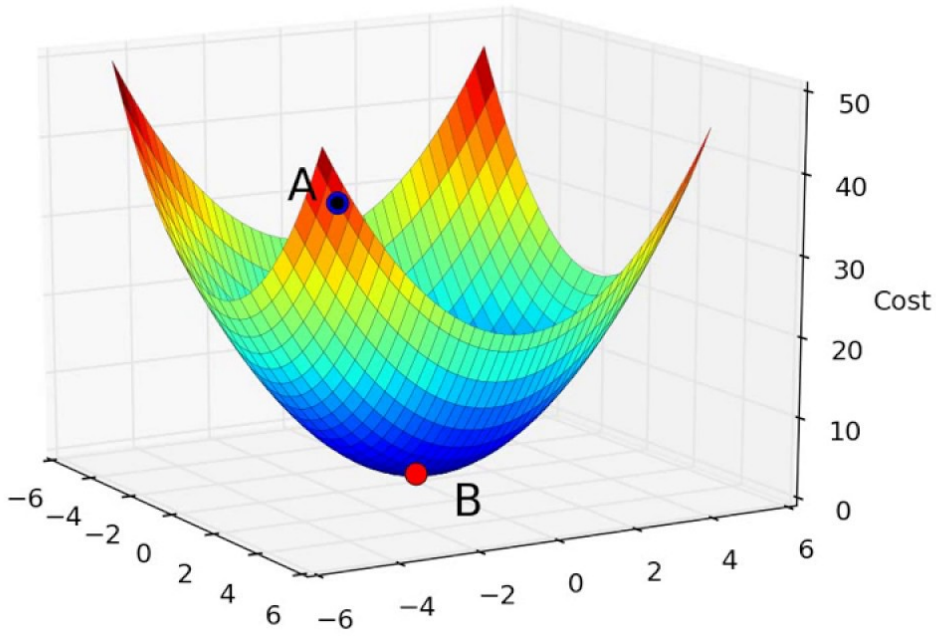
$y_{pred} = mx_i + c = Predicted\ dependent\ values$

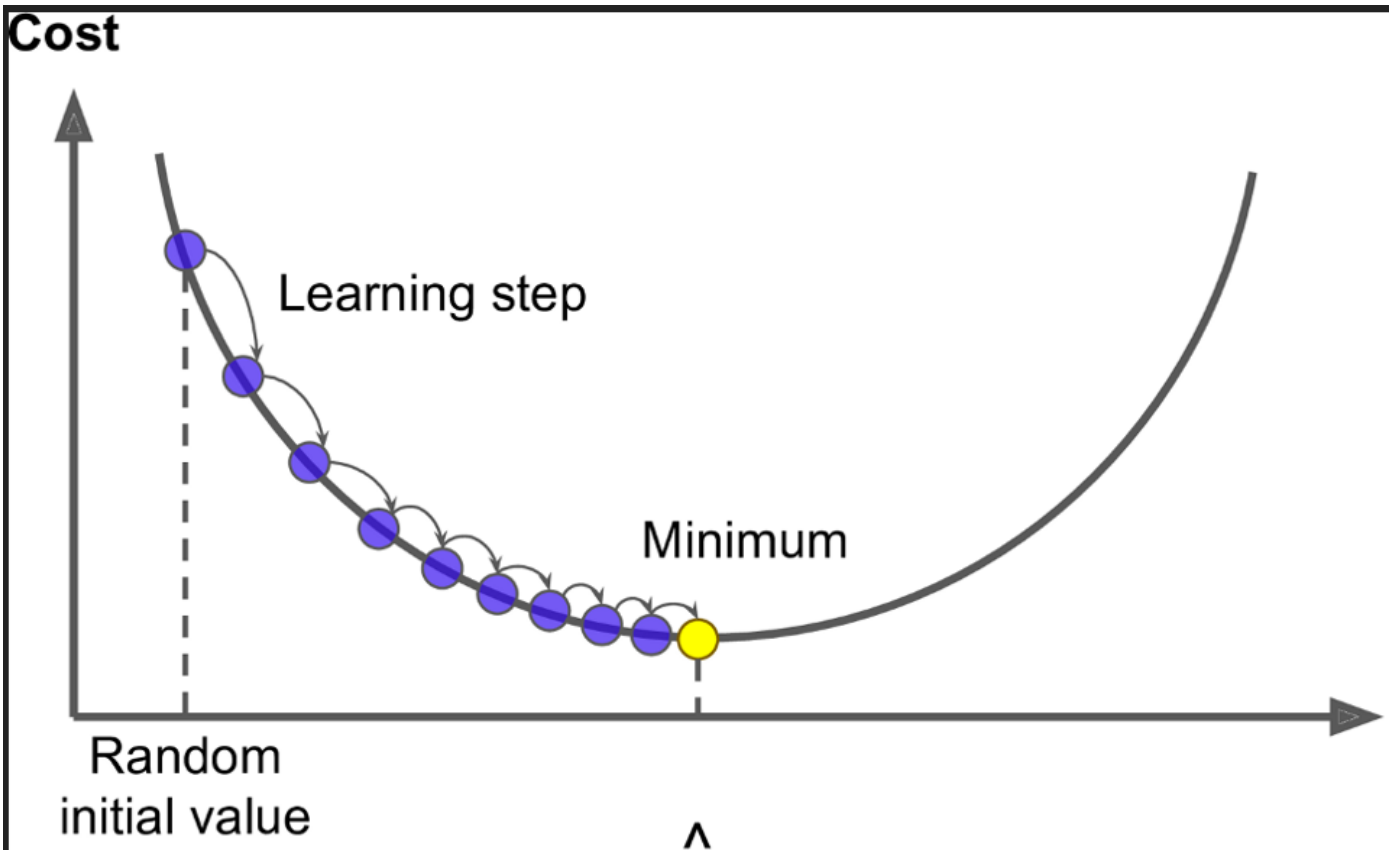
Residuals:

- The distance between the actual value and predicted values is called residual.
- If the observed points are far from the regression line, then the residual will be high, and so cost function will high.
- If the scatter points are close to the regression line, then the residual will be small and hence the cost function.
- if Residulas are:
  - Positive >> Then data points are above the regression line
  - Negative >> Then data points are below the regression line
  - Zero >> Then data points are on the regression line

Gradient Descent:

- Gradient descent is used to minimize the MSE by calculating the gradient of the cost function.
- A regression model uses gradient descent to update the coefficients of the line by reducing the cost function.
- It is done by a random selection of values of coefficient and then iteratively update the values to reach the minimum cost function.
- Function requirements:
  1. differentiable
  2. convex





- Gradient Descent method's steps are:

1. choose a starting point (initialisation)
2. calculate gradient at this point
3. make a scaled step in the opposite direction to the gradient (objective: minimise)
4. repeat points 2 and 3 until one of the criteria is met:
    - maximum number of iterations reached
    - step size is smaller than the tolerance.

Model Performance:

- The Goodness of fit determines how the line of regression fits the set of observations.
- The process of finding the best model out of various models is called optimization. It can be achieved by below method:

1. Residual:
   - Residual = actual value — predicted value
   - If your residual plots look normal, go ahead, and evaluate your model with following various metrics.

$$Residual = y_i - y_{pred}$$

2. Sum of Squared Error(SSE)
   - Sum squared diff between actual values(observed values) and predicted values

$$SSE = \sum_{i=0}^{n} (y_i - y_{pred})^2$$

3. Sum of Squares due to Regression(SSR):
   - Sum of squared difference between predicted and mean of dependent variable

$$SSR = \sum_{i=0}^{n} (y_{pred} - y_{mean})^2$$

4. Total Error(SST):
   - Sum of Squared difference between actual values and mean of dependent variable

$$SST = \sum_{i=0}^{n} (y_i - y_{mean})^2$$

$$SST = SSE + SSR$$

5. Mean Squared Error:
   - The most common metric for regression tasks is MSE.
   - It has a convex shape.
   - It is the average of the squared difference between the predicted and actual value.
   - Since it is differentiable and has a convex shape, it is easier to optimize.

$$MSE = \frac{1}{n} \sum_{i=0}^{n} (y_i - y_{pred})^2$$

6. Root Mean Squared Error (RMSE):
   - This is the square root of the average of the squared difference of the predicted and actual value.
   - R-squared error is better than RMSE. This is because R-squared is a relative measure while RMSE is an absolute measure of fit (highly dependent on the variables — not a normalized measure).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=0}^{n}(y_i - y_{pred})^2}$$

7. Mean Absolute Error (MAE)
   - This is simply the average of the absolute difference between the target value and the value predicted by the model.
   - Not preferred in cases where outliers are prominent.

$$MAE = \frac{1}{n}\sum_{i=0}^{n}\left|y_i - y_{pred}\right|$$

8. R-squared method:
   - R-squared is a statistical method that determines the goodness of fit.
   - It measures the strength of the relationship between the dependent and independent variables on a scale of 0-100%.
   - The high value of R-square determines the less difference between the predicted values and actual values and hence represents a good model.
   - It is also called a coefficient of determination, or coefficient of multiple determination for multiple regression.
   - It can be calculated from the below formula:

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SST - SSE}{SST} = \frac{Explained\ variation}{Total\ variation}$$

$$R^2 = 1 \quad ; \quad when\ SSE = 0$$

$$R^2 = 0 \quad ; \quad when\ SSE = SST$$

$$R^2 < 0 \quad ; \quad when\ SSE > SST$$

9. Adjusted R-squared:

   - Every time we add a new input variable, there will be an increase in the R-squared.
   - So, it is not a good approach to use the R-squared as a deciding quantity as to whether we should add a new input variable or not. Hence, one more quantity is known as “Adjusted R squared” is used
   - It is a modified version of R squared. It is more useful when we add irrelevant variables to our model, which means if we add variables that do not affect the target variable then the adjusted R-squared value will decrease and R squared value will increase.
   - It is always lower than the R-squared
   - Usually, the value of R squared and adjusted R Squared is somewhat the same but if you see a large difference then you need to check out your independent variables again and see if there is any relationship between the target variable and the independent variable.

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

$where,$

$R^2 = R - squared \quad ; \quad n = number\ of\ values \quad ; \quad p = number\ of\ predictors$

Applications of Linear Regression Algorithm

Linear regression is widely used in biological, behavioral and social sciences to describe possible relationships between variables. It ranks as one of the most important tools used in these disciplines. Some of the domians are:

1. Finance
2. Environmental science
3. Machine learning
4. Trend line
5. Epidemiology

Advatages of Linear Regression Algorithm

1. Simple implementation
2. Performance on linearly seperable datasets
3. Overfitting can be reduced by regularization

Disadvatages of Linear Regression Algorithm

1. Linear Regression Is Limited to Linear Relationships
2. Linear Regression Only Looks at the Mean of the Dependent Variable
3. Linear Regression Is Sensitive to Outliers
4. Data Must Be Independent
5. Prone to underfitting