Linear regression is a method used to find the best-fit line that can predict values based on independent variables. In regression, we have a set of records with X and Y values, where X represents the input (independent variable) and Y represents the output (dependent variable). The goal is to learn a function from these records that can predict the value of Y when given an unknown X. In simple terms, linear regression helps to predict a continuous value for Y based on the values of X.

Equation for Linear Regression

The equation for Linear Regression is:

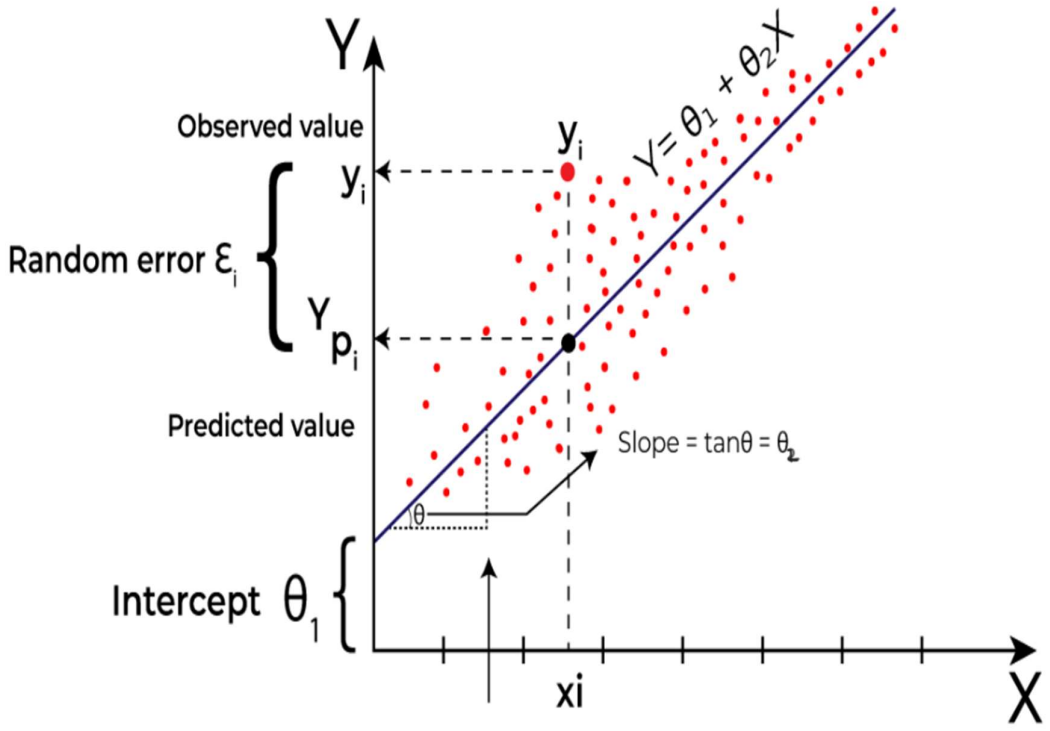
$y\hat{}=\theta 1+\theta 2X$

where,

- $y\hat{}$ is the predicted value
- X is the independent variable
- $\theta 2$ is the slope of the line
- $\theta 1$ is the y-intercept

Best Fit Line

The best-fit line in linear regression is the line that best represents the relationship between the input (independent variable) and output (dependent variable). It’s found by minimizing the difference between the actual data points and the line's predictions. This line helps us make the most accurate predictions, and its slope shows how much the output changes when the input changes by one unit.



Use of Best Fit Line

Prediction: Once the best fit line is established, it can be used to make predictions. Given a value of the independent variable(s), the equation allows us to estimate the corresponding value of the dependent variable. This is particularly useful for forecasting and understanding trends in data.

Understanding Relationships: The slope (β_1) of the best fit line indicates the strength and direction of the relationship between the variables. A positive slope suggests a positive correlation, while a negative slope indicates a negative correlation.

Visual Representation: The best fit line is often plotted on a scatterplot of the data points. It visually represents the linear trend in the data and helps assess how well the model fits the observed data.

Cost Function of Linear Regression

In regression, the difference between the observed value of the dependent variable (*yi*) and the predicted value (*y^*) is known as error or residual. It can be expressed as follows:

$\epsilon i = y\hat{} - yi$

The **cost function** is used to determine the optimal values for the coefficients ($\theta 1$ and $\theta 2$) that result in the best-fit line for the data points. In linear regression, the most commonly used cost function is the **Mean Squared Error (MSE)**, which calculates the

average of the squared differences between the predicted values (*y^*) and the actual observed values (*yi*). This cost function helps to minimize the overall error and find the line that best fits the data.

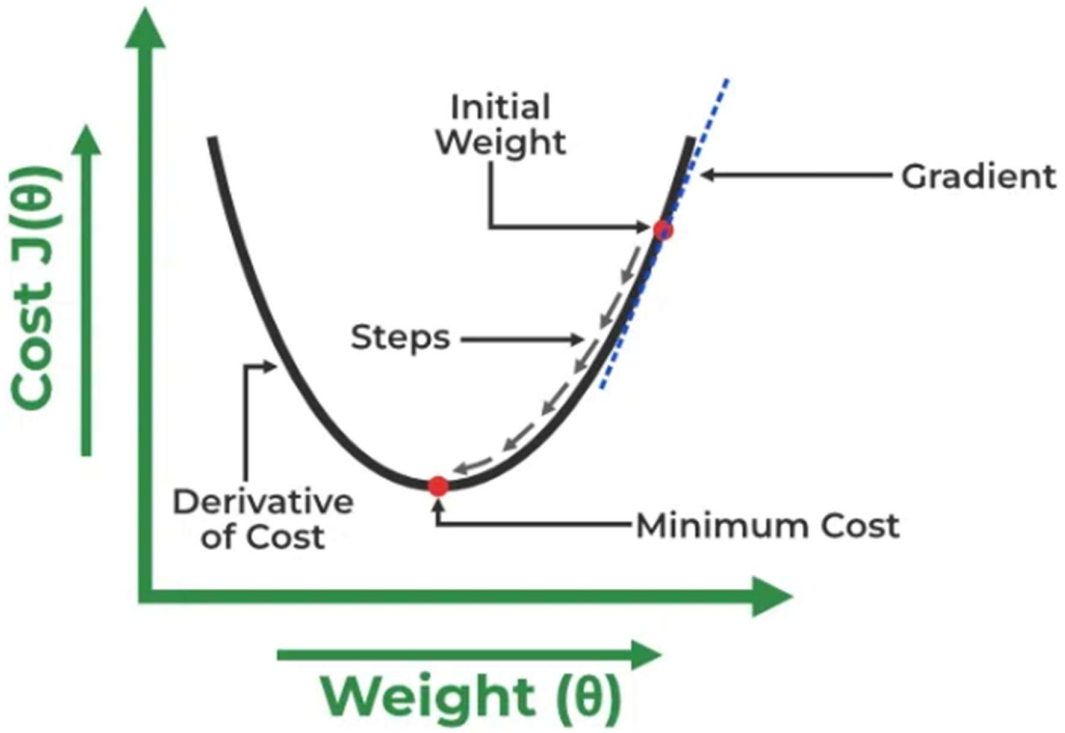
$$minimize \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$ Mean Squared Error

Gradient Descent

Gradient descent is a method used to find the best values for the parameters (like the slope and intercept in linear regression) by minimizing the error or cost. It helps us adjust the parameters so that the model's predictions are as close as possible to the actual data.

Here's a simpler explanation:

- **Start with a Guess:** Begin with random values for the parameters (like slope and intercept).
- **Calculate the Error:** The algorithm checks how far off the predictions are from the actual values. This error is called the "cost."
- **Adjust the Parameters:** To reduce the error, the algorithm changes the parameters slightly. It makes adjustments based on the "gradient" (the direction of the steepest change in error). This is done in small steps.
- **Repeat:** The algorithm keeps adjusting the parameters until the error is as small as possible, or until further changes don't help much.



Gradient Descent

Adjusting Parameters Using Gradient Descent

Let's differentiate the cost function (J) with respect to $\theta 1$:

$$
\begin{aligned}
J'_{\theta_1} &= \frac{\partial J(\theta_1, \theta_2)}{\partial \theta_1} \\
&= \frac{\partial}{\partial \theta_1}\left[\frac{1}{n}\left(\sum_{i=1}^{n}(\hat{y}_i - y_i)^2\right)\right] \\
&= \frac{1}{n}\left[\sum_{i=1}^{n} 2(\hat{y}_i - y_i)\left(\frac{\partial}{\partial \theta_1}(\hat{y}_i - y_i)\right)\right] \\
&= \frac{1}{n}\left[\sum_{i=1}^{n} 2(\hat{y}_i - y_i)\left(\frac{\partial}{\partial \theta_1}(\theta_1 + \theta_2 x_i - y_i)\right)\right] \\
&= \frac{1}{n}\left[\sum_{i=1}^{n} 2(\hat{y}_i - y_i)\,(1 + 0 - 0)\right] \\
&= \frac{1}{n}\left[\sum_{i=1}^{n} (\hat{y}_i - y_i)\,(2)\right] \\
&= \frac{2}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)
\end{aligned}
$$

Differentiating Cost Function

Let's differentiate the cost function (J) with respect to $\theta 2$:

$$
\begin{aligned}
J'_{\theta_2} &= \frac{\partial J(\theta_1, \theta_2)}{\partial \theta_2} \\
&= \frac{\partial}{\partial \theta_2} \left[\frac{1}{n} \left(\sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \right) \right] \\
&= \frac{1}{n} \left[\sum_{i=1}^{n} 2(\hat{y}_i - y_i) \left(\frac{\partial}{\partial \theta_2} (\hat{y}_i - y_i) \right) \right] \\
&= \frac{1}{n} \left[\sum_{i=1}^{n} 2(\hat{y}_i - y_i) \left(\frac{\partial}{\partial \theta_2} (\theta_1 + \theta_2 x_i - y_i) \right) \right] \\
&= \frac{1}{n} \left[\sum_{i=1}^{n} 2(\hat{y}_i - y_i) (0 + x_i - 0) \right] \\
&= \frac{1}{n} \left[\sum_{i=1}^{n} (\hat{y}_i - y_i) (2x_i) \right] \\
&= \frac{2}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i) \cdot x_i
\end{aligned}
$$

Differentiating Cost Function

Now for the new best fit line the new $\theta 1$ and $\theta 2$ are:

$$
\begin{aligned}
\theta_1 &= \theta_1 - \alpha \left(J'_{\theta_1} \right) \\
&= \theta_1 - \alpha \left(\frac{2}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i) \right) \\
\theta_2 &= \theta_2 - \alpha \left(J'_{\theta_2} \right) \\
&= \theta_2 - \alpha \left(\frac{2}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i) \cdot x_i \right)
\end{aligned}
$$

New Coefficients for the Best-fit Line

where, α is the **Learning rate** (which we will learn more about later).