LINEAR REGRESSION →

Linear regression is a method or a process in which one dependent or response variable can be predicted using independent variable or more than one independent variables when they have a linear relationship with each other. The relationship can be given by the following equation:

$$y = \beta_0 + \beta_1 x + \text{errors}$$

here y = response variable or dependent variable

x = predictor variable or independent variable.

There can be more than one independent variable as well.

β_0 and β_1 are known as the model parameters. Linear regression is basically prediction using weighted sum of input features, along with the sum of bias which is also known as intercept. The main linear regression equation can be given as follows.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \cdots\cdots \beta_n x_n$$

Note:-

The difference between the actual and predicted for the population size is called as errors; and for the sample is called as the residuals.

While using linear regression, we try to find the model parameters which give us the best fit line and the minimum errors or residuals. We do this by using the gradient descent where we change the value of β_1 & β_0 and then measure the residuals. We again measure the new residual after changing the model parameters. The errors or the residuals are called as the cost function which we try to minimise

$$\text{Cost function} = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2 \quad \text{(or)} \quad \sum_{i=1}^{n} (\epsilon_i)^2$$

This method of obtaining a best fit line is called as the method of least squares.

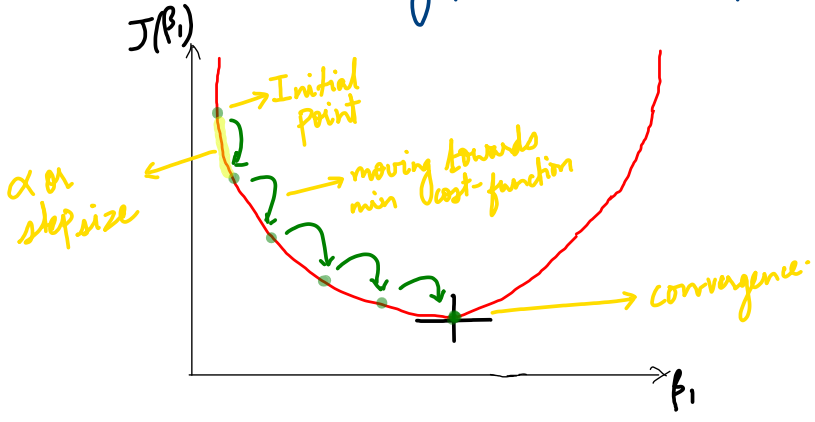
The method of finding the best values of model parameters such that the cost function is minimum is called as the OPTIMIZATION. The least square cost function is given by →

$$\text{Least square cost function} \rightarrow J(\beta_0, \beta_1) = \frac{1}{2n} \left[\sum_{i=1}^{n} \left(Y_i - (\beta_0 + \beta_1 X_i) \right)^2 \right]$$

- $\frac{1}{2}$: constant
- n : no. of observations
- Y_i : actual value
- $\beta_0 + \beta_1 X_i$: predicted value

Now we will look at the graph or visualisation of gradient descent and see how theoritically we can reduce the cost function.

At the very begning we initialise the value of β_0 & β_1. The we check what is the value of cost function at these values, then we change the values of β_0 & β_1 to reduce the cost function When we plot these values on a x-y plane, then the plot of gradient descent looks as shown below.



Now in case of more than one predictors, or independent variables, the gradient descent plot is visualised in 3D. or xyz plane. In that case, we can reduce the value of J by moving towards the lower descent. The direction of the lower descent is given by $-\Delta J$. ΔJ is the partial derivative of the cost function. So, the new values of the model parameters or $\beta_0, \beta_1, \beta_2 \dots \beta_n$ can be calculated as follows →

$$\text{New parameter or } \beta = \text{Old } \beta - \alpha \times \frac{\partial J}{\partial \beta}$$

→ α is the step-size which we are taking while gradient descent and is constant

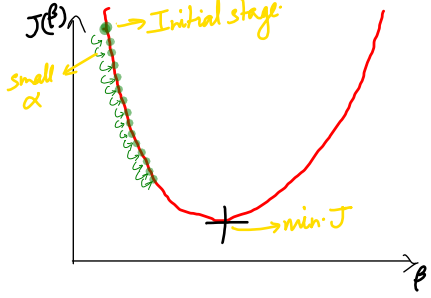
So, new intercept or bias can be given as →

$$\text{New } \beta_0 = \text{Old } \beta_0 - \alpha \frac{1}{n} \left[\sum_{i=1}^{n} (H_0(x_i) - Y_i) \right]$$

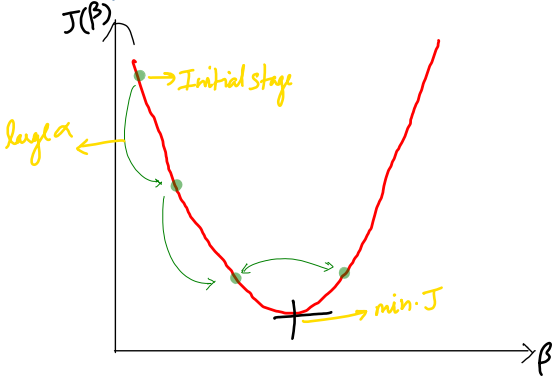
Optimization using gradient descent is repeated up until convergence is achieved.

Now, important parameter while working with the gradient descent is the step size or α. We have to choose the value of α very carefully.

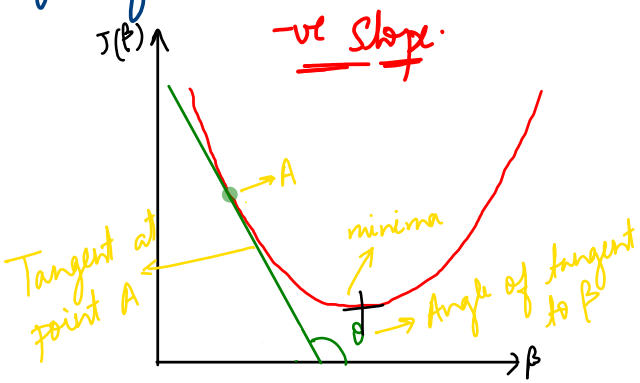
If the value of α is very small, then it may happen that we doesn't reach the minimum value of J at all as can be seen from the plot below.

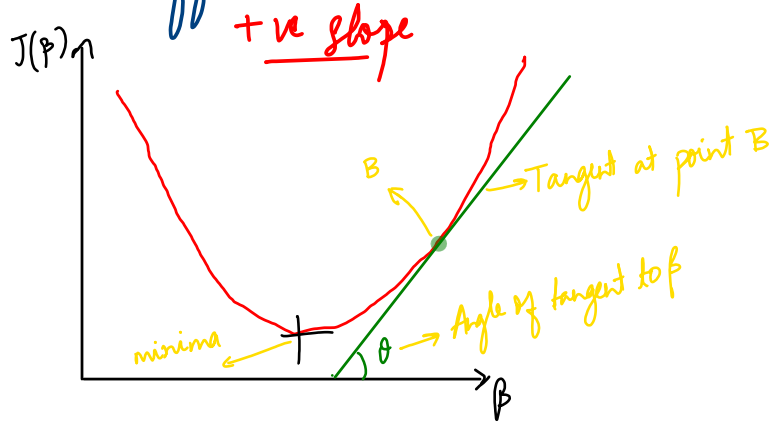


Now, if we take a very large value of α, then it is possible that we may never reach the cost function minima. We will just keep shooting between the values of J.



Also, it is important to know that do we need to increase or decrease the values of model parameters or coefficients in order to reach the lower cost function. For that we have a way to find out whether to increase or decrease the coefficients →





If the slope of the tangent at the initial stage or point is negative, then we increase the value of coefficients or model parameters. If the slope of the tangent is positive, then in that case we decrease the value of the coefficients.

Hence we can say that cost function is inversly proportional to the slope of tangent of coefficients So it can be represented as

$$\text{New } \beta = \text{Old } \beta - \text{slope of tangent} \times \text{constant} \quad \text{(constant → This is } \alpha\text{)}$$

So, we get the equation →

$$\beta_{new} = \beta_{old} - \alpha * \Delta J$$

Assumptions Of Linear Regression →

→ First of all linear regression always assumes that there is a linear relationship between response and the predictors. We can check the relationship between the response and predictors using a scatter plot.

→ Second assumption is that there should not be any correlation between the residuals If there is correlation between the residuals, then it is called as Autocorrelation. This generally happens in case of time-series analysis when the prediction of next time period is dependent on previous time period.

→ There should be a constant variation between the residuals or the error should have constant variance. When there is common variance, then it is known as homoskedasticity. When the variation or variance is not common among the residuals, then it is known as heteroskedasticity. We can check the variance by plotting a residual plot. It is a scatter plot between residuals and predicted values.

Note → It is always usefull after building the model to look at the coefficients and intercept to see how they affect the predicted values.

→ There should not be correlation between the predictors or independent variables. If there is correlation between the predictors then it is called as colinearity. We can find out the colinearity by looking at the pairplots or correlation matrix.

But if a predictor or independent variable can be explained using two or more predictors then it is called as multicolinearity.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

and $x_3 = a x_1 + b x_2$

In the above case, multicolinearity is present as x_3 can be explained using x_1 & x_2. This cannot be identified using the scatter plot or correlation matrix. To identify it, we have the Variance Inflation Factor or VIF. If value of VIF is higher than the set cutoff, we remove that variable. VIF can be calculated as shown:

V_1 | V_2 V_3 V_4 V_5 V_6 V_7 V_8

V_1 → y

V_2 V_3 V_4 V_5 V_6 V_7 V_8 → x

$$R_2 = 1 - \frac{MSE(\text{baseline})}{MSE(\text{model})}$$

$$VIF = \frac{1}{1 - R^2}$$

First we take the predictor for which we are calculating the VIF as response and other variables as predictors. After that we calculate the value of VIF and check if it is greater or smaller than our cutoff. Generally we take the cutoff for VIF as 5 but we can change it as well. If the value is greater than cutoff, we remove that predictor.

→ The residuals should be normally distributed. We can check the normal distribution of the plots using Q-Q plot or distribution plots.