1) **What is Linear Regression ?**

Ans) It is the algorithm to find a best fit line to a given dataset .This algorithm uses the equation "y = m*x + c" to find the line.

This algorithm treats the dependent feature and independent feature to have a linear relationship.

Here,

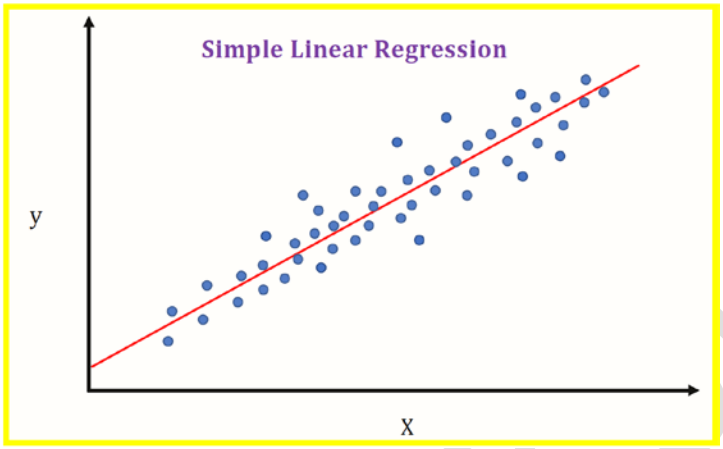
y is dependent variable.

x is independent variable.

m is the slope (change in value of y per unit change in x)

c is the intercept (The point at which line intersects the y axis)

The optimal line is found out by altering the values of “m” and “c”.



2) **How can we calculate error in Linear Regression ?**

Ans) In simple words, It is the method of calculation of

Error = True values - Predicted values (For independent Feature)

Ideally the value of error should be zero, but, practically this is not accepted because it can lead to over-fitting (“TOO GOOD IS BAD” in Machine Learning).

Mathematically,

J(c,m) = ((y(pred)- y(real))^2)/m (J → Error Function) ----------(1)

J is function of c (intercept) and m (slope) as there can be error in both the values.

m is number of observations

Equation (1) is the Mean Squared Error.

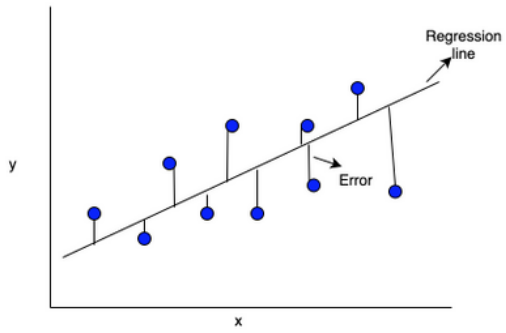
Other methods for finding errors are Mean Absolute Error(MAE) , Root Mean Squared Error(RMSE).

3) **What is error ?**

Ans) Mathematically,

Error = $y_{pred} - y_{true}$

The error arises because the Linear model does not correctly predict all the true outcomes.(This indeed is necessary as to avoid over fitting).



4) Difference between loss and cost function ?

Ans) Loss is error in prediction of individual observation.

$(y_{pred} - y_{true})^2$ = Loss

Cost function is sum of losses for all the observations. (m is number of observations)

Sum$((y_{pred} - y_{true})^2 /m)$ (for all m observations)

5) Mean Absolute Error ,Mean Square error ,Root Mean Squared Error.

Ans) $|(y_{pred} - y_{true})|/m$ Mean Absolute Error (here all the values obtained will be taken in modulus)

$(y_{pred} - y_{true})^2/m$ Mean Square Error (square of –ve number is also positive)

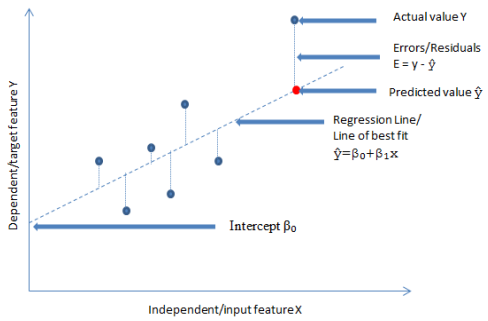
$((y_{pred} - y_{true})^2/m)^{1/2}$ Root Mean Squared Error

6) Explain what intercept term mean ?

Ans) For

y = mx + c

If we take x = 0 then y = c (i.e the point at which the line intersects the y axis)



7) Write all assumptions for linear regressions.

Ans)

a) There should be linear relationship between dependent and independent features.

b) There should be no multi-co linearity between the dependent features.

c) The data is normally distributed i.e For any fixed value of X ,Y should be normally distributed.(This can be obtained by plotting boxplot of residuals.)

d) The variance of residual is the same for any value of X.(Homoscedasticity) .The residuals should be randomly distributed around the horizontal axis.

8) How is Hypothesis testing used in Linear Regression ?

Ans) It is used for testing if the error has been reduced or not i.e(When we plot the residuals obtained by say MSE method, then optimal choice is to select the minima of this obtained parabola*). In real scenarios this is not a proper practice as it can lead to over fitting.

Thus,we define a 95% Confidence Interval for finding this minima point. This point thus is treated as optimal point where the value of error is minimum.

*As the equation is squared the graph is parabolic.

9) How do you decide the importance of variable for a multivariate regression ?

Ans) The main check is for R^2.

If, on addition of a variable to the dataset the value of R^2 increases beyond a certain pre-decided Threshold then it means that variable is important. This also shows, how strongly the variable is related to the output. This is just a statistical test and it does not account for significance of the variable from the domain perspective.

For eg. Probability of Rainfall is dependent on Moisture Content in the Atmosphere but Moisture Content in turn is dependent of Temperature .Thus statistically both are relevant to the model but from Domain perspective

perspective Temperature is more Important.

9) What is R^2 vs Adjusted R^2 ?

Ans) R stands for Residuals which means the error which in turn is $(y_{pred} - y_{true})$

R^2 = Variance of model/Total Variance

Variance of model → Measure of how much the observed values differ from the average / predicted value.

Total Variance → Sum of residuals (Obtained by MSE ,MAE or RMSE method)

The value of variance has range $0 < R^2 < 1$ (where 1 means perfect fit and 0 means "perfect misfit")

This can be interpreted as the relation between the independent variable plotted on X axis and dependent variable plotted on y axis .

If R^2 = 1 the Dependent and Independent variable are perfectly related and vice versa. Practically, R^2 =1 is not possible.

Thus R^2 numerically depicts the strength of co-relation between the Dependent and Independent variable.

However, R^2 will always increase as we add more independent variables, but, by doing this one might end up adding totally irrelevant variables to the equation. Thus to overcome this drawback we adapt to Adjusted R^2.

$$Adjusted\ R^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

Where

R^2 Sample R-Squared

N Total Sample Size

p Number of independent variable

Intuitively, it is the measure of increase in R^2 for every new independent variable added. From formula it is evident that every new addition to the independent variable penalizes the whole term(term which is subtracted from 1) .Thus , it is a simple trade off between gain in R^2 vs penalty for including the new variable.

If the gain is more we add the variable or else we discard it.

Eg Price of Pizza is dependent on cost of Raw Materials but if the age of person preparing it is added as an independent variable then R^2 will increase but this does not make sense practically. Thus to avoid such situation we use Adjusted R^2 because as we add more variables the denominator term will increase and overall value of term decreases .Also gain in R^2 is insignificant as compared to the loss born by adding it.

Thus we remove discard this variable.

Note : Open to suggestions and corrections.

THANKS TO SUDHANSHU SIR, KRISH SIR, SUNNY SIR and INEURON team.