1. What is a Linear Regression?

Linear Regression is a statistical method to find out the best fit line for a dataset where the accuracy is more and error should be minimal. It shows linear relation between one dependent variable (y) and one independent feature (x) and predict the value of y using x. It is used to predictive the value of 'y' (label) based on the input 'x' (feature).

2. How we can calculate error in linear regression?

Error is the difference between actual value and predicted value.

There are 4 ways to calculate error in linear regression

i. MAE (Mean Absolute Error)

Mean absolute error is the average of the residuals (difference between actual point and predicted point) of a linear regression.

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

ii. MSE (Mean Square Error)

Mean square error is the average of square of residuals of a linear regression.

As we are squaring the difference in MSE, we can not compare directly MSE to the MAE.

Outliers in the dataset affect more in the MSE compared to MAE, as we are finding the square it grows error quadratically.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$

iii. MAPE (Mean Absolute Percentage Error)

Mean absolute percentage error is the percentage equivalent of MAE.

MAPE is the ratio of the residual over actual.

$$MAPE = \frac{100\%}{n}\sum\left|\frac{y - \hat{y}}{y}\right|$$

Multiplying by 100% converts to percentage

The residual

Each residual is scaled against the actual value

Iv. **MPE (Mean Percentage Error)**

Mean Percentage error is exactly same as MAPE but it lacks absolute value operation. MPE will give us positive and negative errors which help us to see if our model is underestimates (more negative error) or overestimates (more positive error).

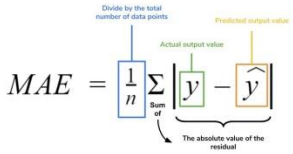
$$MPE = \frac{100\%}{n} \Sigma \left(\frac{y - \widehat{y}}{y} \right)$$

3. Difference between loss and cost function ?

Loss function is to calculate the deviation from actual value to predicted value for a single record. And cost function is the aggregate of the residuals for entire training dataset.

4. Difference between MAE, MSE and RMSE ?

i. MAE (Mean Absolute Error) represents the average of the absolute difference between actual points and predicted points in the dataset in linear regression. It measures the average of the residuals in a dataset.



$$MAE = \frac{1}{n} \Sigma \left| y - \widehat{y} \right|$$

ii. MSE (Mean Standard Error) represents the average of the square difference between actual points and predicted points of the data set in linear regression. It measures the variance of the errors (residuals).

.

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2$$

iii. RMSE (Root Mean Square Error) is the square root of the Mean Squared Error. It measures the standard deviation of the errors (residuals).

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

5. Explain how Gradient Descent works in linear regression ?

Gradient descent is an algorithm that finds the best fit line which will predict the value of y with more accuracy and low error.

While training a model, the model calculates the cost function which measures the Root Mean Squared Error between actual value and predicted value. Gradient Descent tries to minimize the cost function.

To minimize the cost function model needs the best value of θ_1 (intercept) and θ_2 (Slope). Initially model selects random value of θ_1 and θ_2 randomly. But then iteratively it updates the minimize value of θ_1 and θ_2 .By time model will find the best value of θ_1 and θ_2 and tries to find the point where error will be minimum.

$$\frac{\partial}{\partial \mathrm{m}} = \frac{2}{N}\sum_{i=1}^{N} - x_i(y_i - (mx_i + b))$$

$$\frac{\partial}{\partial \mathrm{b}} = \frac{2}{N}\sum_{i=1}^{N} - (y_i - (mx_i + b))$$

6. Explain what the term intercept means ?

Intercept means a line crosses an axis.

Intercept is of two types

i. X-intercept

ii. Y-intercept

i. <u>X-intercept</u>

X-intercept is the line where a line crosses X axis. At this point Y coordinate will be zero.

ii. <u>Y-intercept</u>

Y-intercept is the line where a line crosses Y axis. At this point X coordinate will be zero.

7. Write the assumptions of linear regression ?

The linear regression has five key assumptions

i. Linear relationship
(Linear relationship is a correlation between two variables which describes how much one variable changes as related to change in the other variable)

ii. Multivariate Normality
(Multiple regression assumes that the residuals are normally distributed)

iii. No or little multicollinearity (Multicollinearity is a statistical concept which describes the relationship between two features)

iv. No auto-correlation
(Auto-correlation refers to correlation between the values of independent variables.)

v. Homosedacity
(Homosedacity describes the situation where the error term is same or constant across all the values of independent variable)

8. **How hypothesis testing is used in linear regression ?**

   In Linear regression model when we try to fit the straight line we get the slope and intercept for the line. Whenever we re-run the model by reducing slope and intercept we test if the line is significant or not by checking if the coefficient is significant.

   Steps to perform hypothesis test :

   i. **Formulate a hypothesis**
   ii. **Determine the significance level**
   iii. **Determine the type of test**
   iv. **Calculate the test statistic value and p-values**
   v. **Make decision.**

9. **How would you decide the importance of a variable for the multivariate regression ?**

   Generally the predictor variable with the largest standardized regression coefficient as the most important variable, the predictor variable with the next largest standardized regression coefficient as the next important variable.

10. **Relation between R- squared and Adjusted R-squared ?**

    As per R-squared value all the independent variables affect the result of the model, whereas the adjusted-R squared value defines only the independent variables which actually have an effect on the performance of the model.