**Linear Regression Interview questions**

**1. What are the important assumptions of Linear regression**?

A linear relationship

Restricted Multi-collinearity value

Homoscedasticity

Firstly, there has to be a linear relationship between the dependent and the independent variables. To check this relationship, a scatter plot proves to be useful.

Secondly, there must no or very little multi-collinearity between the independent variables in the dataset. The value needs to be restricted, which depends on the domain requirement.

The third is the homoscedasticity. It is one of the most important assumptions which states that the errors are equally distributed.

**2. What is heteroscedasticity?**

Heteroscedasticity is exactly the opposite of homoscedasticity, which means that the error terms are not equally distributed. To correct this phenomenon, usually, a log function is used.

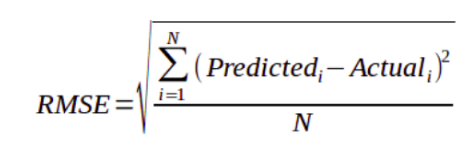
**3. What is the difference between R square and adjusted R square?**

R square and adjusted R square values are used for model validation in case of linear regression. R square indicates the variation of all the independent variables on the dependent variable. i.e. it considers all the independent variable to explain the variation. In the case of Adjusted R squared, it considers only significant variables(P values less than 0.05) to indicate the percentage of variation in the model.

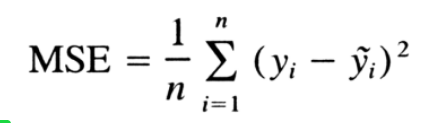
**4. How to find RMSE and MSE?**

RMSE and MSE are the two of the most common measures of accuracy for a linear regression.

RMSE indicates the Root mean square error, which indicated by the formulae:



Where MSE indicates the Mean square error represented by the formulae:



**5. What are the possible ways of improving the accuracy of a linear regression model?**

There could be multiple ways of improving the accuracy of a linear regression, most commonly used ways are as follows:

1. Outlier Treatment:

-Regression is sensitive to outliers, hence it becomes very important to treat the outliers with appropriate values. Replacing the values with mean, median, mode or percentile depending on the distribution can prove to be useful.

**6. How to interpret a Q-Q plot in a Linear regression model?**

A Q-Q plot is used to check the normality of errors. In the above chart mentioned, Majority of the data follows a normal distribution with tails curled. This shows that the errors are mostly normally distributed but some observations may be due to significantly higher/lower values are affecting the normality of errors.

**7. What is the significance of an F-test in a linear model?**

– The use of F-test is to test the goodness of the model. When the model is re-iterated to improve the accuracy with changes, the F-test values prove to be useful in terms of understanding the effect of overall regression.

**8. What are the disadvantages of the linear model?**

– Linear regression is sensitive to outliers which may affect the result.

– Over-fitting

– Under-fitting