Links :

https://medium.com/analytics-vidhya/preparing-for-interview-on-machine-learning-3145caeea06b

https://medium.com/analytics-vidhya/60-interview-questions-on-machine-learning-8afbdac2d22d

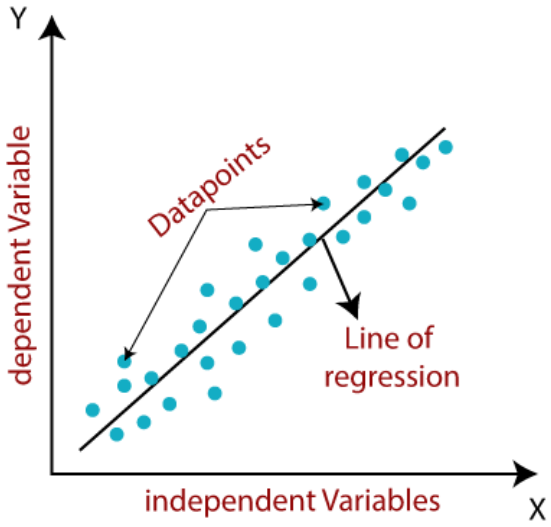
https://www.mygreatlearning.com/blog/machine-learning-interview-questions/

Q.1. What is linear regression?

Linear regression is a method of finding the best straight line fitting to the given data, i.e. finding the best linear relationship between the independent and dependent variables.

One variable, denoted x, is regarded as the predictor, explanatory, or independent variable.The other variable, denoted y, is regarded as the response, outcome, or dependent variable.

In technical terms, linear regression is a machine learning algorithm that finds the best linear-fit relationship on any given data, between independent and dependent variables. It is mostly done by the Sum of Squared Residuals Method.



Q.2. How do you represent a simple linear regression?

Y = b0 +b1 x1 + e

Y — dependent variable

X1 — independent variable

e — Error term = Y — Y(hat)

Q.3. What is multiple linear regression?

In multiple linear regression that are more than one predictor.

Good models require multiple independent variables in order to address the higher complexity of the problem.

$$Y = b0 + b1\ x1 + b2\ x2 + \ldots + bk\ xk + e$$

Q.4. What does linearity mean?

It means a linear relationship. To check if there is linear relationship between x and y the simplest thing to do is plot a scatter plot between x and y.

Q.5. What are the assumptions made in linear regression model?

The important assumptions in linear regression analysis are:

- There should be a linear and additive relationship between dependent (response) variable and independent (predictor) variable(s). A linear relationship suggests that a change in response Y due to one unit change in X is constant, regardless of the value of X. An additive relationship suggests that the effect of X on Y is independent of other variables.
- There should be no correlation between the residual (error) terms.
- The independent variables should not be correlated.
- The error terms must have constant variance. This phenomenon is known as homoskedasticity.
- The error terms must be normally distributed.

Q.6. What if these assumptions get violated ?

To understand the outcomes of violating such assumptions we have to dive into the assumptions.

•Linear and Additive: If we fit a linear model to a non-linear and non-additive data set, the regression algorithm would fail to capture the trend mathematically, thus resulting in an inefficient model. Also, this will result in erroneous predictions on an unseen data set.

•Autocorrelation: Autocorrelation occurs when the residuals are not independent from each other. In other words when the value of y(x+1) is not independent from the value of y(x).The presence of correlation in error terms drastically reduces model's accuracy. This usually occurs in time series models where the next instant is dependent on previous instant. If the error terms are correlated, the estimated standard errors tend to underestimate the true standard error. If this happens, it causes confidence intervals and prediction intervals to be narrower.

•Multicollinearity: This phenomenon exists when the independent variables are found to be

moderately or highly correlated. In a model with correlated variables, it becomes a tough task to figure out the true relationship of predictors with response variable. In other words, it becomes difficult to find out which variable is actually contributing to predict the response variable.Moreover, with presence of correlated predictors, the standard errors tend to increase. And, with large standard errors, the confidence interval becomes wider leading to less precise estimates of slope parameters.

•Heteroskedasticity: The presence of non-constant variance in the error terms results in heteroskedasticity. Generally, non-constant variance arises in presence of outliers. It looks like that these values get too much weight, thereby disproportionately influences the model's performance. When this phenomenon occurs, the confidence interval tends to be unrealistically wide or narrow.

•Normal Distribution of error terms: If the error terms are non- normally distributed, confidence intervals may become too wide or narrow. Presence of non — normal distribution suggests that there are a few unusual data points which must be studied closely to make a better model.

Q.7 What is the assumption of homoscedasticity?

In simple terms it means the equal variance.There is no relationship between the error term and the predicted Y

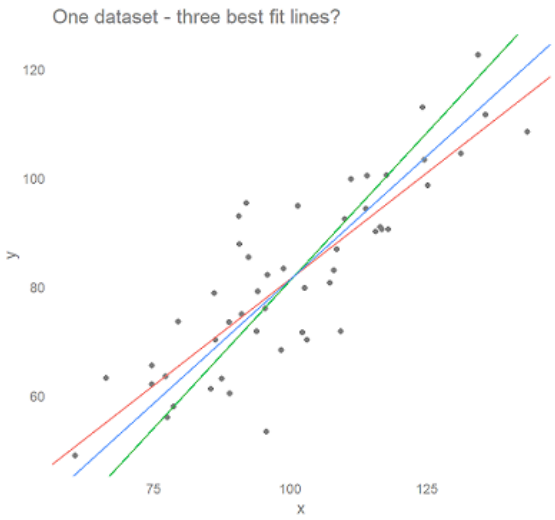
Q.8. What is the assumption of normality?

It means the normal distribution of the error term. The mean of the residuals should be zero. The standard deviation of the residuals should be constant

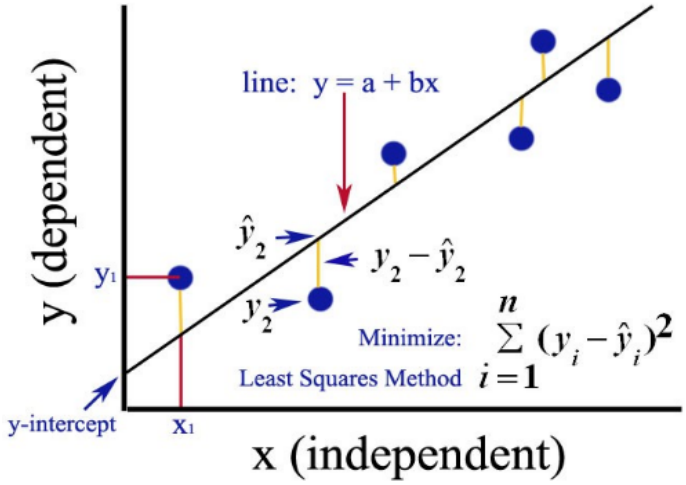
Q.10. What does multicollinearity mean?

When two or more variables have high correlation.If there is perfect multicollinearity then standard error will be infinite. Imperfect multicollinearity means that the correlation is slightly less than 1 or slightly more than -1. However imperfect multicollinearity also causes serious issues in the model

Q.12. How to find the best fit line in a linear regression model?



To find the best fit line for our model we have to make the distance with respect to all the points minimum. We have to find that line which is closest to all the points. In statistics, this vertical distance is called residual.



Residual is equal to the difference between the observed value and the predicted value. For data points above the line, the residual is positive, and for data points below the line, the residual is negative. So, if were to find out the sum of all the residuals, then due to the negative errors there will be subtractions in the distance and the the value of resultant distance would be less than the actual. So, to eliminate the negative sign we have to square each residual and find out it's sum.

formula:

Residual = Observed - Predicted

$$\text{residual} = y - \hat{y}$$

$$\text{sum of squared residuals} = \Sigma(y - \hat{y})^2$$

This distance is known as Sum of Squared Residuals(SSE) and the method is known as Least Squares Method as we need to find that value of m and b of the linear regression line for which SSE is minimum.

Q.13. Why do we square the error instead of using modulus?

It's true that one could choose to use the absolute error instead of the squared error. In fact, the absolute error is often closer to what we want when making predictions from our model. But, we want to penalize those predicted values which is contributing the maximum error. Moreover looking a little deeper, the squared error is everywhere differentiable, while the absolute error is not (its derivative is undefined at 0). This makes the squared error more amenable to the techniques of mathematical optimization. To optimize the squared error, we can just set its derivative equal to 0 and solve. To optimize the absolute error often requires more complex techniques. Actually we find the Root Mean Squared Error so that the unit of RMSE and the dependent variable are equal.

Q.14. What are techniques adopted to find the slope and the intercept of the linear regression line which best fits the model?

There are mainly two methods:

- Ordinary Least Squares(Statistics domain)
- Gradient Descent(Calculus family)

Q.15. Explain Ordinary Least Squares Regression in brief.

Ordinary least squares (OLS) regression is a statistical method of analysis that estimates the relationship between one or more independent variables and a dependent variable. The method estimates the relationship by minimizing the sum of the squares of the difference between the observed and predicted values of the dependent variable configured as a straight line. OLS regression is used in bivariate model, that is, a model in which there is only one independent variable (X)

predicting a dependent variable (Y). However, the logic of OLS regression can also be used in multivariate model in which there are two or more independent variables.

What are the limitations of OLS?

OLS is computationally too expensive. It performs well with small data. For larger data Gradient Descent is preferred.

Q.17.How to evaluate regression models?

There are five metrics used to evaluate regression models:

- Mean Absolute Error(MAE)
- Mean Squared Error(MSE)
- Root Mean Squared Error(RMSE)
- R-Squared(Coefficient of Determination)
- Adjusted R-Squared

Q.18. Which evaluation technique should you prefer to use for data having a lot of outliers in it?

Mean Absolute Error(MAE) is preferable to use for data having too many outliers in it because MAE is robust to outliers whereas MSE and RMSE are very susceptible to outliers and starts penalizing the outliers by squaring the residuals.

Q.19. What is a residual? How is it computed?

Residual is also called Error. It is the difference between the predicted y value and the actual y value.

Residual = Actual y value – Predicted y value.

It can be positive or negative.

If residuals are always 0, then your model has a Perfect R square i.e. 1.

Q.20. What is TSS, ESS and RSS? and What is the relationship between them?

- TSS stands for Total Sum of Squares. It measures the total variability.

$$TSS = \sum(y - y(mean))2$$

- ESS stands for Explained Sum of Squares. It measures the variability that is explained.

$$ESS = \sum(y(pred) - y(mean))^2$$

- RSS stands for Residual Sum of Squares. It measures the difference between the observed Y and predicted Y.

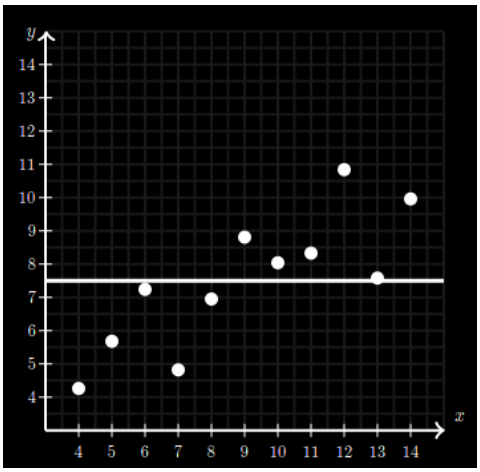
$$RSS = \sum(y - y(pred))^2$$

$$TSS = ESS + RSS$$

Total variability = Explained variability + Unexplained variability

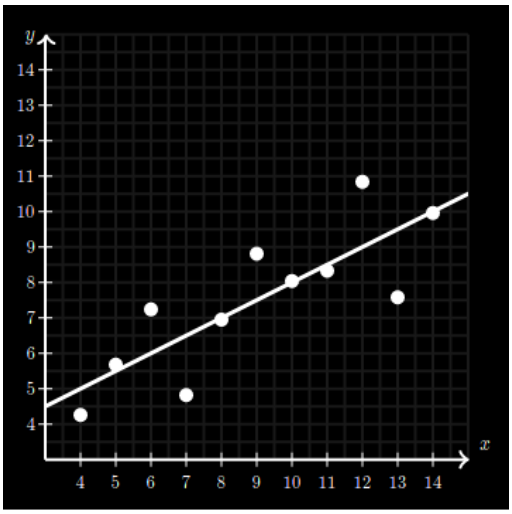
Q.21. What's the intuition behind R-Squared?

We use linear regression to predict y given some value of x. But suppose that we had to predict a y value without a corresponding x value.Without using regression on the x variable, our most reasonable estimate would be to simply predict the average of the y values.



However, this line will not fit the data very well(as we can see in the figure above). One way to measure the fit of the line is to calculate the sum of the squared residuals — this gives us an overall sense of how much prediction error a given model has.

Now, if we predict the same data with regression we will see that the least-squares regression line will seem to fit the data pretty well (as shown in the figure below).



We will find that using least-squares regression, the sum of the squared residuals has been considerably reduced.

So using least-squares regression eliminated a considerable amount of prediction error. R-squared tells us what percent of the prediction error in the y variable is eliminated when we use least-squares regression on the x variable.

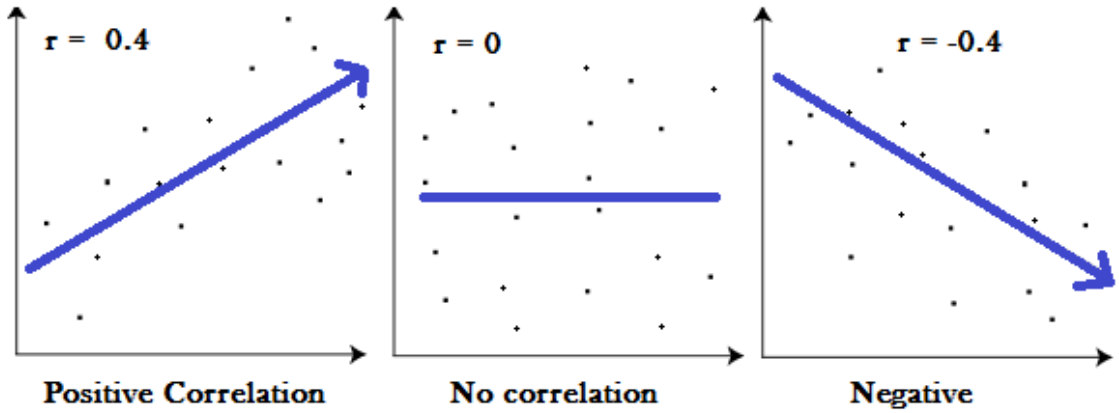
As a result, R^2 is also called the coefficient of determination. Many formal definitions say that R^2 tells us what percent of the variability in the y variable is accounted for by the regression on the x variable. The value of R^2 varies from 0 to 1.

Q.26. What is the Coefficient of Correlation: Definition, Formula and Easy Steps

Correlation coefficients are used in statistics to measure how strong a relationship is between two variables. There are several types of correlation coefficient, but the most popular is Pearson's. Pearson's correlation (also called Pearson's R) is a correlation coefficient commonly used in linear regression. If you're starting out in statistics, you'll probably learn about Pearson's R first. In fact, when anyone refers to the correlation coefficient, they are usually talking about Pearson's.

Correlation coefficient formulas are used to find how strong a relationship is between data. The formulas return a value between -1 and 1, where:

- 1 indicates a strong positive relationship.
- -1 indicates a strong negative relationship.
- A result of zero indicates no relationship at all.



- A correlation coefficient of 1 means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other. For example, shoe sizes go up in (almost) perfect correlation with foot length.
- A correlation coefficient of -1 means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other. For example, the amount of gas in a tank decreases in (almost) perfect correlation with speed.
- Zero means that for every increase, there isn't a positive or negative increase. The two just aren't related.

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$

The absolute value of the correlation coefficient gives us the relationship strength. The larger the number, the stronger the relationship. For example, |-.75| = .75, which has a stronger relationship than .65.

Q.27. What is the primary difference between R square and adjusted R square?

In linear regression, you use both these values for model validation. However, there is a clear distinction between the two. R square accounts for the variation of all independent variables on the dependent variable. In other words, it considers each independent variable for explaining the variation. In the case of Adjusted R square, it accounts for the significant variables alone for indicating the percentage of variation in the model. By significant, we refer to the P values less than 0.05.

Q.30. How do you interpret a Q-Q plot in a linear regression model?

As the name suggests, the Q-Q plot is a graphical plotting of the quantiles of two distributions with respect to each other. In other words, you plot quantiles against quantiles.

Whenever you interpret a Q-Q plot, you should concentrate on the 'y = x' line. You also call it the 45-degree line in statistics. It entails that each of your distributions has the same quantiles. In case

you witness a deviation from this line, one of the distributions could be skewed when compared to the other.

Q. 31. What are the advantages and disadvantages of Linear Regression?

- Simple implementation : Linear Regression is a very simple algorithm that can be implemented very easily to give satisfactory results.Furthermore, these models can be trained easily and efficiently even on systems with relatively low computational power when compared to other complex algorithms.Linear regression has a considerably lower time complexity when compared to some of the other machine learning algorithms.The mathematical equations of Linear regression are also fairly easy to understand and interpret.Hence Linear regression is very easy to master.

- Performance on linearly seperable datasets : Linear regression fits linearly seperable datasets almost perfectly and is often used to find the nature of the relationship between variables.

- Overfitting can be reduced by regularization : Overfitting is a situation that arises when a machine learning model fits a dataset very closely and hence captures the noisy data as well.This negatively impacts the performance of model and reduces its accuracy on the test set. Regularization is a technique that can be easily implemented and is capable of effectively reducing the complexity of a function so as to reduce the risk of overfitting.

Disadvantages of Linear Regression

- Prone to underfitting : Underfitting : A sitiuation that arises when a machine learning model fails to capture the data properly.This typically occurs when the hypothesis function cannot fit the data well. Since linear regression assumes a linear relationship between the input and output varaibles, it fails to fit complex datasets properly. In most real life scenarios the relationship between the variables of the dataset isn't linear and hence a straight line doesn't fit the data properly. In such situations a more complex function can capture the data more effectively.Because of this most linear regression models have low accuracy.

- Sensitive to outliers : Outliers of a data set are anomalies or extreme values that deviate from the other data points of the distribution.Data outliers can damage the performance of a machine learning model drastically and can often lead to models with low accuracy.

- Linear Regression assumes that the data is independent : Very often the inputs aren't independent of each other and hence any multicollinearity must be removed before applying linear regression.

OR

Advantages :

1. Easy To implement.

2. performs exceptionally well for linear data.

3. Easy to find out accuracy of model (MSE, R^2, RMSE)

Disadvantages :

1. Lot of processing is required

2. observations are independent to each other. So result accuracy may affect.

3. Missing values are not allowed. It impacts the result

4. Impact of Outliers