1. Explain the linear regression algorithm in detail.

Linear regression is a parametric machine learning algorithm where each parameter has a linear relationship with the dependent and respective independent variable.

The hypothesis equation of linear regression is as below:

$$y = theta0 + theta1*X1 + theta2*X2 + theta3*X3 + \ldots\ldots$$

where y denotes the dependent variable; X1, X2, X3... denotes the independent variables and theta1, theta2, theta3 … denotes the linear parameters and theta0 is the intercept

The objective of linear regression algorithm is to estimate the values of theta(s) from a known set of dependent and independent variables called the training set.

The estimation of parameter is obtained by minimizing a loss (or error) function defined on the training data and prediction based on the parameters.

The one of the loss function used is the least square or OLS, defined as

$$E = \Sigma (y\_train\_i - y\_pred\_i)$$

Where y_train_i: the i th training data for dependent variable

Y_pred_i: the data predicted from the i th training data for independent variable and the parameters.

The loss functions are minimized with the help of gradient descent algorithm.

2. What are the assumptions of linear regression regarding residuals?

The following assumptions are made regarding the residuals in linear regression:

a. Normality assumption : It is assumed that the residuals are random and are normally distributed.
b. Zero Mean : The residuals are assumed to have zero mean
c. Constant Variance : It is assumed that residuals terms have same variance
d. Independent Assumption : All error terms are independent of each other

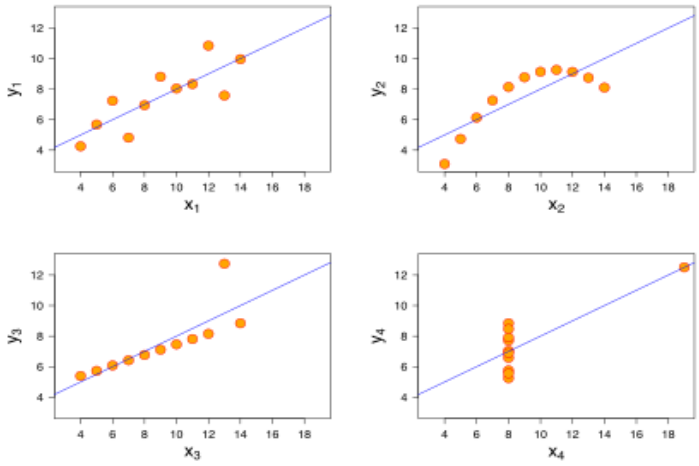
3. What is the coefficient of correlation and the coefficient of determination?

Coefficient of Correlation measures the strength and direction of relative movement between two variables. The values range between –1 and 1, where positive value denotes that the variables move in same direction and negative value denotes they move in opposite direction. There are several types of correlation coefficients like pearson's r and spearman's r.

Coefficient of Determination measures how much the variability of one variable can be explained by the variability in another variable. This variable ranges between 0 and 1. A coefficient of determination of 0.2 between x and y suggests that 20% of variability of x can be explained by y.

4. Explain the Anscombe's quartet in detail.

Anscombe's quartet consists of four datasets having exactly same descriptive statistics but very different distributions. Each dataset appears to be very different when plotted on a graph. These datasets are used to describe the effect of outliers on the statistical properties.



- The first data is a simple linear relationship.
- The second data is not distributed normally; while a relationship between the two variables is obvious, it is not linear.
- The third dataset (bottom left), the distribution is linear, but should have a different regression line. The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph shows an example when a single high leverage point is enough to produce a high correlation coefficient, even when all other points indicate no significant relation

5. What is Pearson's R?

Pearson's R is a coefficient of correlation that measures the strength and direction of relationship between two variables that are linearly associated. It ranges between –1 and 1, with 1 denoting perfect relationship in same direction , -1 denoting perfect relationship in opposite direction and 0 denoting no relation.

The formula for pearson's r for two variables x and y is covariance of x and y divided by variance of x times variance of y.

6. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling the operation performed on the numerical features of a dataset to bring all of them to a close range and thus removing high variance in magnitude between the various features.

Feature scaling is performed mostly in machine learning algorithms that compute distances between data (like ordinary least squares used in linear regression). As higher magnitude features will automatically impact the final distance more than smaller magnitude features, it is important to scale them, so that each has equal footing.

Normalized scaling scales the data between the range of 0 and 1, by subtracting the minimum value and dividing by the maximum value. This changes the inherent distribution of the features and makes them clustered.

Standardizing scales the data while maintaining the inherent distribution of the data by shifting the mean to zero and scaling the variance to 1. This is achieved by subtracting the mean from each point and dividing by the standard deviation.

7. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

VIF becomes infinite for those variables which are completely explained by combination of some other variable , thus making the R2 value 1 and thus taking the vif to infinite. It must be noted that the vif for the corresponding descriptor variables also goes to infinite.

8. What is the Gauss-Markov theorem?

The Gauss-Markov theorem states that if a linear regression model satisfies the seven classical assumptions then ordinary least squares regression produce the unbiased estimates that have least variance of all possible linear estimators.

The assumptions are as follows:

a. A linear relationship exists between dependent and independent variables
b. Residuals are normally distributed
c. Residuals have 0 mean
d. Residuals have constant variance
e. Residuals are independent of each other
f. The independent variables are measured without errors
g. The independent variables are independent of each other

9. Explain the gradient descent algorithm in detail.

Gradient descent is the iterative algorithm used to minimize the loss function in various machine learning algorithms. The steps are as follows:

1. Start with a random value of the parameters theta(s). Generally zero is taken
2. Next we find out the estimate of the dependent variables in at the current value of theta(s) using the independent variables from training set
3. Then we compute the loss function based on this predicted value of dependent variable (from step 2) and the corresponding known values of dependent variable from the training set
4. After this we compute the gradient of the loss function wrt each parameter theta
5. If the value of loss function is very small or below a certain decided threshold we stop, else we go to step 6.
6. This gradient tells us the direction and size of step to take for each parameter theta to achieve the highest minimization. We subtract this gradient times some learning rate from the present value of the parameters theta.
7. We go back to step 2.

10. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Quantile-quantile plot (Q-Q plot) is used to determine if two datasets have same underlying distribution and whether they come from same source or not.

It is used for diagnosing the residuals of a linear regression model, mainly to check if the observations come from a normal distribution which is one of the basic assumptions of linear regression.