1. Explain the linear regression algorithm in detail.
   - Simple linear regression is a type of regression analysis where the number of independent variables is one and there is a linear relationship between the independent(x) and dependent(y) variable. There is a best fit straight line. Based on the given data points, we try to plot a line that models the points the best. The line can be modelled based on the linear equation shown below.

   y = b0 + b1*x

   The motive of the linear regression algorithm is to find the best values for b0 and b1.

   - Multiple linear regression is a type of regression analysis where the number of independent variables is more than one and there is a linear relationship between the independent variables (x) and dependent(y) variable. The line can be modelled based on the equation shown below.

   y = b0 + b1*x 1+b2*x2+b3*x3+…..+bnxn

   The motive of the linear regression algorithm is to find the best values for the constant b0 and coefficients b1,b2,etc.

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

   To use residuals in a linear regression model, some of the following assumptions are used:
   - We need to have at least one linear parameter mapping with the other.
   - Mean of the all residual values should approximately be zero.
   - There might be no perfect relationship (correlation) with all the explanatory variables.
   - Residuals are not correlated.

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

   Coefficient of correlation (R value) is the measure of dependability between two variables. Calculated using difference of Predicted Y and actual Y value. Predicted Y is the Y point which falls on the best fit line. It varies from -1 to +1. If x and y have strong correlation, R value is positive. If x and y have reverse relation if the R value is negative. If x and y has no correlation, R value tends to be close to zero. Whereas, Coefficient of Determination(R^2) gives the parameter of how a direct variation between X and Y is. R^2 varies between 0 to 1 . 1 signifies the variable is closely related and Y can be best explained by that X value. R^2 value close to 1 represents the strong fit of values on the line, and 0 signifies that points are scarcely placed on the graph.

4. Explain the Anscombe’s quartet in detail.

   Anscombe’s quartet actually is pair 4 data sets with different data points and behavior in it. Although they are from the same dataset, they appear very different when graphed for the best fit line. Each datasets include 11 pair of data point pair, when plotted provides different best fit, with illustrate and analyze the type of relationship between variables and we can conclude to realistic variables.

5. What is Pearson's R?

   Pearson's R(P value) is the measure of strength and association of variables. It varies between -1 to +1 It is done by mapping the scatterplot and seeing the points how far are the points from straight fit line. Closer the points from straight line, means the higher correlation. If points are scattered, it means the variables (points) are weakly correlated. One approach is to do the t test which can signify the normal to hyper vent association of the variables and correlations.

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

   Scaling is modifying the scale of data to fit the graph in a readable manner so that the dependent unit can have a fair relation with the result and visually can be pointed. Scaling is done for variables to increase or decrease the magnitude to have a meaningful plot. Normalized scaling is done in the confined range where entire graph shifts to fit the necessity. Whereas in standardized scaling, the reference point is the mean of the graph, and the graph is shifted based on that center.

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

   VIF has formula as 1/(1- R^2) Which says if R^2 is one, the denominator becomes zero and the VIF becomes infinite. In literal terms, If the correlation is strong, VIF is infinite. And weekly correlated variables is considered if they have greater than VIF of 5.

8. What is the Gauss-Markov theorem?

   The Gauss Markov theorem tells us that if a certain set of assumptions are met, the ordinary least squares estimate for regression coefficients gives you the *best linear unbiased estimate (BLUE)* possible.
   Gauss Markov Assumptions
   There are five Gauss Markov assumptions (also called *conditions*):

   1. Linearity: the parameters we are estimating using the OLS method must be themselves linear.
   2. Random: our data must have been randomly sampled from the population.
   3. Non-Collinearity: the regressors being calculated aren't perfectly correlated with each other.
   4. Exogeneity: the regressors aren't correlated with the error term.
   5. Homoscedasticity: no matter what the values of our regressors might be, the error of the variance is constant.

   Purpose of the Assumptions
   The Gauss Markov assumptions guarantee the validity of ordinary least squares for estimating regression coefficients.
   Checking how well our data matches these assumptions is an important part of estimating regression coefficients. When you know where these conditions are violated, you may be able to plan ways to change your experiment setup to help your situation fit the ideal Gauss Markov situation more closely.

   In practice, the Gauss Markov assumptions are rarely all met perfectly, but they are still useful as a benchmark, and because they show us what 'ideal' conditions would be. They also allow us to pinpoint problem areas that might cause our estimated regression coefficients to be inaccurate or even unusable.
   The Gauss-Markov Assumptions In Algebra

We can summarize the Gauss-Markov Assumptions succinctly in algebra, by saying that a linear regression model represented by
$y_i = x_i' \beta + \varepsilon_i$
and generated by the ordinary least squares estimate is the best linear unbiased estimate (BLUE) possible if

- $E\{\varepsilon_i\} = 0, i = 1, . . . , N$
- $\{\varepsilon_1 …… \varepsilon_n\}$ and $\{x_1 ….., x_N\}$ are independent
- $\mathrm{cov}\{\varepsilon_i, \varepsilon_j\} = 0$, i, j = 1,…., N I ≠ j.
- $V\{\varepsilon_1 = \sigma^2$, i= 1, ….N

The first of these assumptions can be read as "The expected value of the error term is zero.". The second assumption is collinearity, the third is exogeneity, and the fourth is homoscedasticity.

9. Explain the gradient descent algorithm in detail.

Gradient descent is an optimization algorithm used to find the values of parameters (coefficients) of a function (f) that minimizes a cost function (cost).

Gradient descent is best used when the parameters cannot be calculated analytically (e.g. using linear algebra) and must be searched for by an optimization algorithm.

Gradient Descent Procedure

The procedure starts off with initial values for the coefficient or coefficients for the function. These could be 0.0 or a small random value.

coefficient = 0.0

The cost of the coefficients is evaluated by plugging them into the function and calculating the cost.

cost = f(coefficient)

or

cost = evaluate(f(coefficient))

The derivative of the cost is calculated. The derivative is a concept from calculus and refers to the slope of the function at a given point. We need to know the slope so that we know the direction (sign) to move the coefficient values in order to get a lower cost on the next iteration.

delta = derivative(cost)

Now that we know from the derivative which direction is downhill, we can now update the coefficient values. A learning rate parameter (alpha) must be specified that controls how much the coefficients can change on each update.

coefficient = coefficient – (alpha * delta)

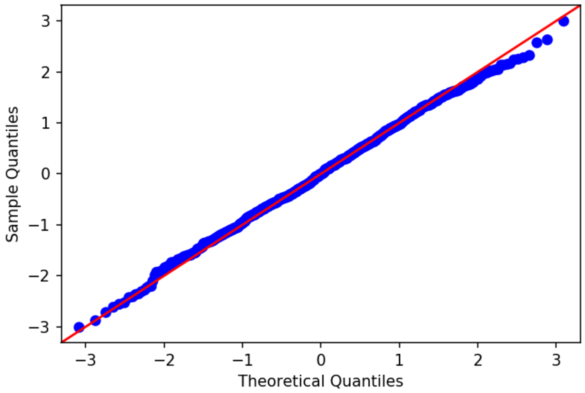
This process is repeated until the cost of the coefficients (cost) is 0.0 or close enough to zero to be good enough.

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

The “QQ” in QQ plot means quantile-quantile — that is, the QQ plot compares the quantiles of our data against the quantiles of the desired distribution (defaults to the normal distribution, but it can be other distributions too as long as we supply the proper quantiles).

Quantiles are breakpoints that divide our numerically ordered data into equally proportioned buckets. For example, you’ve probably heard of percentiles before — percentiles are quantiles that divide our data into 100 buckets (that are ordered by value), with each bucket containing 1% of observations. Quartiles are quantiles that divide our data into 4 buckets (0–25%, 25–50%, 50–75%, 75–100%). They’re a quick and visual way to assess whether a variable is normal or not.

e.g., In the below figure, the blue dots fall pretty cleanly on the red line. That means that our data is normally distributed.


Sample Quantiles
Theoretical Quantiles