1.Explain the linear regression algorithm in detail.

The equation for simple linear regressions is, well, pretty simple:

$$y = b_0 + b_1 * x_1$$

Let's break this equation down.

y is the dependent variable. This is what you're trying to explain and to understand how (in which way) it depends on something else. In this example, our dependent variable would be spy potential.

x is the independent variable. We assume that this is causing the dependent variable to change in some way. This might not be a direct cause, but it implies an association between the variables and we obviously want to know more about that. So in our spy example, the independent variable could be something like IQ or physical fitness.

b_1 is what's known as the coefficient for the independent variable. This is a fancy way of expressing how a unit change in **x** (higher IQ, for example) affects a unit change in **y** (more spy potential). This also controls the angle or slope of the line, so the steeper the line, the higher your spy potential becomes per extra IQ point and the shallower the slope/gradient, the less spy potential a candidate receives per additional IQ point.

b_0 is the constant term, or point where your trendline crosses the horizontal axis.

If you have multiple variables the equation looks almost the same, except you add more independent variables into the mix:

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + \ldots + b_n * x_n$$

For each additional independent variable that you're looking to find a relationship for you add an extra coefficient (b) and independent variable (x) to your formula

First, the relationship between the variables in your data needs to be linear. This means that they can be plotted along a line. Once plotted, the difference between the real and predicted values (residuals) needs to be more or less constant (homoscedastic).

At the same time, the residuals must be independent of each other and the predictors (independent variables) should not be highly correlated.

You can determine whether your data meets these conditions by plotting it and then doing a bit of digging into its structure.

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

The four assumptions are:
Linearity of residuals
Independence of residuals
Normal distribution of residuals
Equal variance of residuals

Linearity – we draw a scatter plot of residuals and y values. Y values are taken on the vertical y axis, and standardized residuals (SPSS calls them ZRESID) are then plotted on the horizontal x axis. If the scatter plot follows a linear pattern (i.e. not a curvilinear pattern) that shows that linearity assumption is met.
Independence – we worry about this when we have longitudinal dataset. Longitudinal dataset is one where we collect observations from the same entity over time, for instance stock price data – here we collect price info on the same stock i.e. same entity over time.
We generally have two types of data: cross sectional and longitudinal. Cross -sectional datasets are those where we collect data on entities only once. For example we collect IQ and GPA information from the students at any one given time (think: camera snap shot). Longitudinal data set is one where we collect GPA information from the same student over time (think: video). In cross sectional datasets we do not need to worry about Independence assumption. It is "assumed" to be met.

Normality: we draw a histogram of the residuals, and then examine the normality of the residuals. If the residuals are not skewed, that means that the assumption is satisfied.

Equality of variance: We look at the scatter plot which we drew for linearity (see above) – i.e. y on the vertical axis, and the ZRESID (standardized residuals) on the x axis. If the residuals do not fan out in a triangular fashion that means that the equal variance assumption is met.

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

R^2 = coefficient of determination = 1-SS.res/SS.tot = SS.reg/SS.tot

(SS = "sum of squares", SS.res = SS of the residuals, SS.tot=total SS, SS.reg=SS of the regression="explained sum of Squares; Note: SS.tot = SS.reg+SS.res).

R^2 is the fraction of explained SS from the total SS. Since VAR = SS/n and the sample size (n) is the same for all (tot, reg, and res), R^2 can be interpreted as the proportion of variance explained by the regression. This means: Without knowing your predictors (just the response), this data has some variance, for instance say 2000. If you know can explain a part of the variation in the response by your predictors (i.e. by your regression), you will get an R^2 value between 0 and 1, for instance say 0.8. This indicates that the residuals of your regression have a variance that 80% smaller than the variance of the response: 2000*(1-0.8)=2000*0.2=400. Thus 80% of the variance of the response are explained by the regression, and 20% of this variance are left over in the residuals.

r = (Pearson's) coefficient of correlation = COV(x,y)/(s(x)s(y))

COV(x,y) is the covariance between x and y, s(x) and s(y) are the standard deviations of x and y, respectively. r is a measure of the strength of a linear dependency between two variables (x and y). In principle, it is a scaled (normalized) measure of the covariance.

Only in a simple linear regression with one (metric) predictor is r = sqrt(R^2)

4.Explain the Anscombe's quartet in detail.

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (*x*,*y*) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analysing it and the effect of outliers on statistical properties

5.What is Pearson's R?

Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

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

Scaling is a method used to normalize the range of independent variables or features of data.

Standardized scaling : In scaling *(also called min-max scaling)*, you transform the data such that the features are within a specific range e.g. [0, 1].

$x' = (x - xmin)/(xmax - xmin)$, where x' is the normalized value.

Scaling is important in the algorithms such as support vector machines (SVM) and k-nearest neighbors (KNN) where distance between the data points is important.

Normalized Scaling : The point of normalization is to change your observations so that they can be described as a normal distribution. Normal distribution (Gaussian distribution), also known as the **bell curve**, is a specific statistical distribution where a roughly equal observations fall above and below the mean, the mean and the median are the same, and there are more observations closer to the mean.

$x' = (x - xmean)/(xmax - xmin)$

For normalization, the maximum value you can get after applying the formula is 1, and the minimum value is 0. So, all the values will be between 0 and 1

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

VIF = 1/ (1 – R^2), so VIF will be infinite when R squared is 1.

An $\mathbf{R^2}$ of **1** indicates that the regression predictions perfectly fit the data.

8.What is the Gauss-Markov theorem?

The theorem states that :

b1 has minimum variance among all unbiased linear estimators of the form

$\hat{\beta}_1 = \sum c_i Y_i$

As this estimator must be unbiased we have

$E\{\hat{\beta}_1\} = \sum c_i E\{Y_i\} = \beta_1$

$= \sum c_i(\beta_0 + \beta_1 X_i) = \beta_0 \sum c_i + \beta_1 \sum c_i X_i = \beta_1$ I

This imposes some restrictions on the ci 's.

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.

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.

You can see how simple gradient descent is. It does require you to know the gradient of your cost function or the function you are optimizing, but besides that, it's very straightforward. Next we will see how we can use this in machine learning algorithms.

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

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.

A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.