1. Explain the linear regression algorithm in detail.

The representation is a linear equation that combines a specific set of input values (x) the solution to which is the predicted output for that set of input values (y). As such, both the input values (x) and the output value are numeric.

The linear equation assigns one scale factor to each input value or column, called a coefficient and represented by the capital Greek letter Beta (B). One additional coefficient is also added, giving the line an additional degree of freedom (e.g. moving up and down on a two-dimensional plot) and is often called the intercept or the bias coefficient.

For example, in a simple regression problem (a single x and a single y), the form of the model would be:

y = B0 + B1*x

In higher dimensions when we have more than one input (x), the line is called a plane or a hyper-plane. The representation therefore is the form of the equation and the specific values used for the coefficients.

When a coefficient becomes zero, it effectively removes the influence of the input variable on the model and therefore from the prediction made from the model (0 * x = 0).

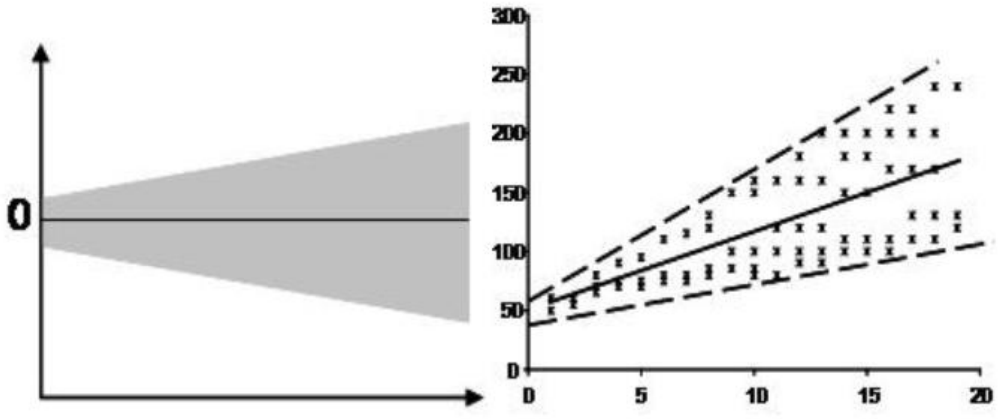
2. What are the assumptions of linear regression regarding residuals?

1. Linear regression requires the relationship between the independent and dependent variables to be **linear**. The linearity assumption can best be tested with scatterplots.

2. Linear regression analysis requires that the errors between observed and predicted values (i.e., the residuals of the regression) should be **normally distributed**. This assumption may be checked by looking at a histogram or a Q-Q-Plot.

3. Linear regression assumes that there is **no multicollinearity** in the data. Multicollinearity occurs when the independent variables are too highly correlated with each other.

4. The last assumption of linear regression is **homoscedasticity**. A scatterplot of residuals versus predicted values is good way to check for homoscedasticity. There should be no clear pattern in the distribution; if there is a cone-shaped pattern (as shown below), the data is heteroscedastic.



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

Coefficient of Correlation: It is the degree of relationship between two variables say x and y. It can go between -1 and 1. 1 indicates that the two variables are moving in unison. They rise and fall together

and have perfect correlation. -1 means that the two variables are in perfect opposites. One goes up and other goes down, in perfect negative way. If they are not correlated then the correlation value can still be computed which would be 0. The correlation value always lies between -1 and 1.

A correlation greater than 0.8 is generally described as strong, whereas a correlation less than 0.5 is generally described as weak.

Coefficient of determination: It shows percentage variation in y which is explained by all the x variables together. Higher the better. It is always between 0 and 1. It can never be negative – since it is a squared value. It is also known as R squared value.

The coefficient of determination is the ratio of the explained variation to the total variation.

For example, if $r = 0.922$, then $r^2 = 0.850$, which means that 85% of the total variation in *y* can be explained by the linear relationship between *x* and *y* (as described by the regression equation). The other 15% of the total variation in *y* remains unexplained.

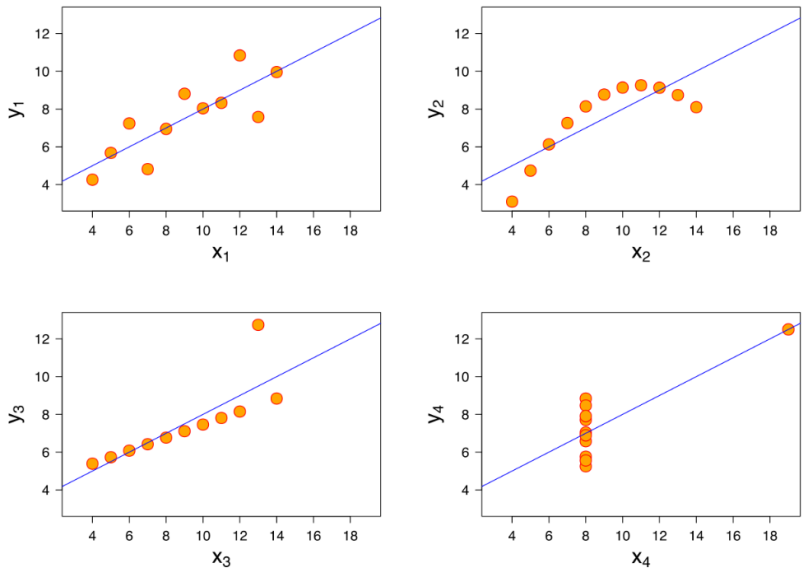
4. Explain the Anscombe's quartet in detail.

It comprises four datasets, each containing eleven (x,y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize COMPLETELY, when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

Quartet's Summary Stats

When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story.



- Dataset I appears to have clean and well-fitting linear models.

- Dataset II is not distributed normally.
- In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
- Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

5. What is Pearson's R?

It is the test statistics that measures the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship.

Assumptions are:

1. Both variables should be normally distributed.
2. There should be no significant outliers
3. Each variable should be continuous
4. The two variables have a linear relationship
5. The observations are paired observations
6. Homoscedascity

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

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units.

If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.
Techniques to perform Feature Scaling**:**

- **Min-Max Normalization:** This technique re-scales a feature or observation value with distribution value between 0 and 1.

$$X_{new} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

- **Standardization:** It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.

$$X_{new} = \frac{X_i - X_{mean}}{\text{Standard Deviation}}$$

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

VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity.

〖VIF〗_2=1/(1-R_2^2)

If there is perfect correlation, then VIF = infinity.

8. What is the Gauss-Markov theorem?

The Gauss Markov theorem says that, under certain conditions, the ordinary least squares (OLS) estimator of the coefficients of a linear regression model is the best linear unbiased estimator (BLUE), that is, the estimator that has the smallest variance among those that are unbiased and linear in the observed output variables.

Assumptions:

The regression model is:

$$y = X\beta + \varepsilon$$

Where,

- y is an N*1 vector of observations of the output variable (N is the sample size);
- X is an N*K matrix of inputs (K is the number of inputs for each observation);
- beta is a K*1 vector of regression coefficients;
- ε is an vector of errors.

The OLS estimator of beta is

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

9. Explain the gradient descent algorithm in detail.

step 1: 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

step 2: 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))

step 3: The derivative of the cost is calculated. The derivative 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)

step 4: Now that we know from the derivative which direction is downhill, we can now update the coefficient values. A learning rate parameter must be specified that controls how much the coefficients can change on each update.

coefficient = coefficient – (alpha * delta)

step 5: 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*.

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

Few advantages:

- It can be used with different sample sizes.
- Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It is used to check following scenarios:

If two data sets

1. come from populations with a common distribution
2. have common location and scale
3. have similar distributional shapes
4. have similar tail behavior

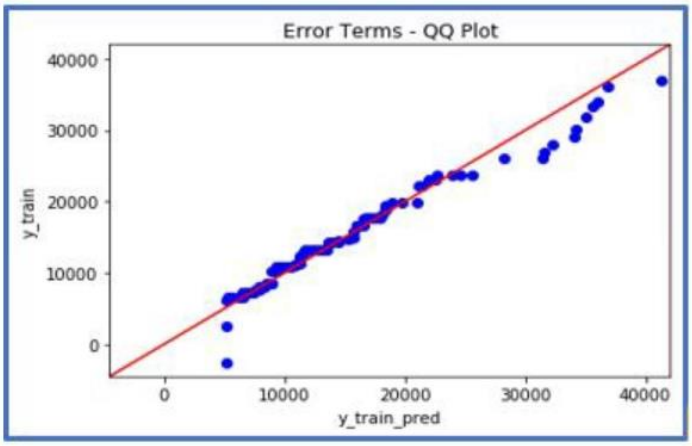
Interpretation*:*

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

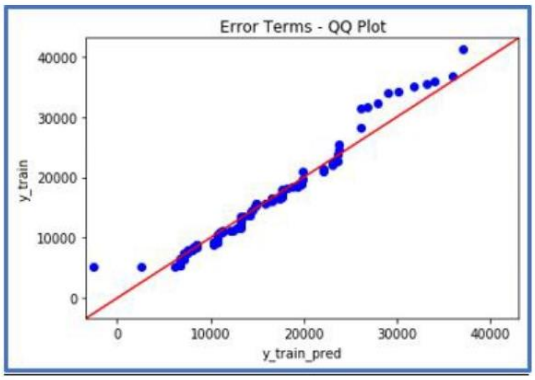
Below are the possible interpretations for two data sets.

1. **Similar distribution**: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis

2. **Y-values < X-values:** If y-quantiles are lower than the x-quantiles



3. **X-values < Y-values:** If x-quantiles are lower than the y-quantiles.



4. Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x –axis.