Explain the linear regression algorithm in detail.

Linear Regression is a type of machine learning algorithm modelling method which comes under Supervised Learning category under Regression type which means the past data with labels is available to build the model and the output is continuous in nature.

In simple terms, linear regression algorithm helps us in finding the best linear fitting relationship to the given data, i.e., finding the best linear relationship between the independent and dependent variables. Key logic used here is Residual Sum of Squares Method.

Linear regression models can be classified into two types depending upon the number of independent variables:

- Simple linear regression: Used when the number of independent variables is one.
- Multiple linear regression: Used when the number of independent variables is more than one.

The strength of a linear regression model is mainly explained by R^2, where $R^2 = 1 - (RSS/TSS)$.
RSS – Residual Sum of Squares & TSS – Total Sum of Squares.

Steps for Building a linear model

- OLS (Ordinary Least Squares) method in statmodel to fit a line.
- Summary statistics
    - F-statistic, r-squared, coefficients and their p-values.
- Residual Analysis
    - Histogram of the error terms to check normality.
    - Plot of the error terms with X or y to check independence.
- Predictions
    - Making predictions on the test set using the ‘predict' function.

What are the assumptions of linear regression regarding residuals?

Assumptions of Simple Linear Regression.

- There is a linear relationship between X (independent variable) and Y (dependent variable):
    - X and Y should display some sort of a linear relationship; otherwise, there is no use of fitting a linear model between them.
- Error terms are normally distributed.
- Error terms are independent of each other.
    - The error terms should not be dependent on one another.
- Error terms have constant variance (homoscedasticity):
    - The variance should not increase (or decrease) as the error values change.
    - Also, the variance should not follow any pattern as the error terms change.

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

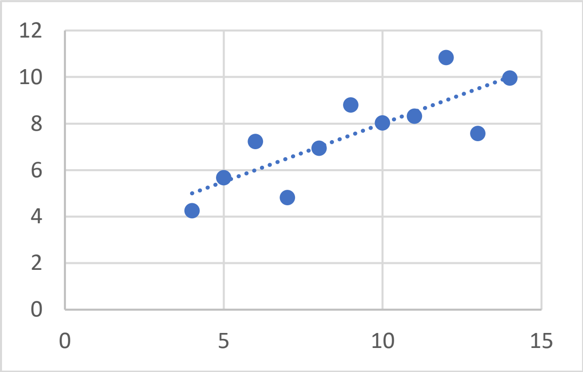
Coefficient of correlation is represented by “R” value which is given in the summary table in the Linear Regression output table whereas coefficient of determination is the value R square. Coefficient of Determination is the square of Coefficient of Correlation.

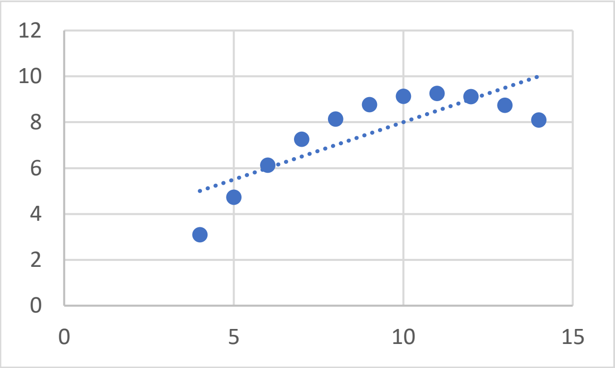
Coefficient of determination (R^2): shows percentage variation in y which is explained by all the x variables together. Higher the better. It is always lies between 0 and 1. It can never be negative – since it is a squared value.

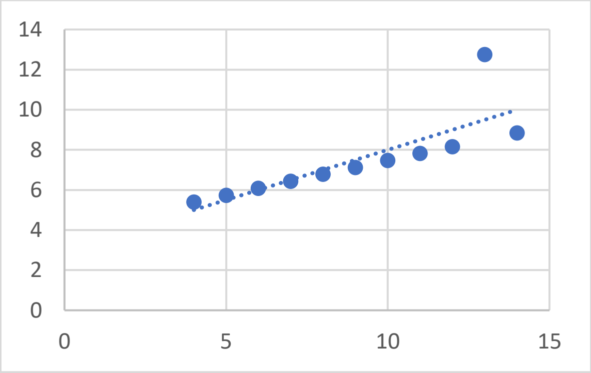
Coefficient of Correlation (R): is the degree of relationship between two variables say x and y. It can go between -1 and 1. 1 indicates that the two variables are moving in unison. They rise and fall together and have perfect correlation. -1 means that the two variables are in perfect opposites. One goes up and other goes down, in perfect negative way.

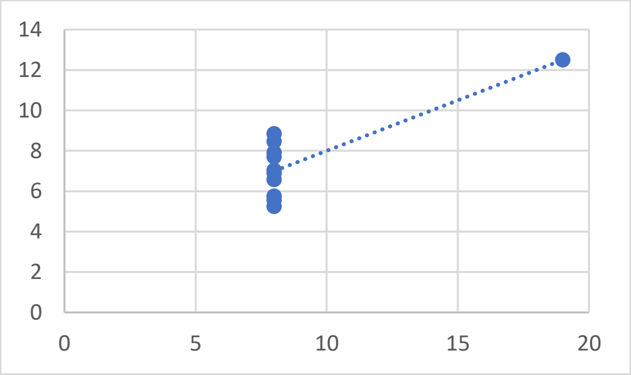
Explain the Anscombe’s quartet in detail.

Anscombe’s quartet comprises of 4 data sets that have identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. For instance here below is an example.









All four data sets have same mean, Sample Variance, Correlation between x & y, similar regression line and R square value. Still the data point are different.

This often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

What is Pearson’s R?

Pearson’s R which is also known as Pearson correlation coefficient is a measure of the linear correlation between two variables X and Y. It is known as correlation coefficient or the bivariate correlation. As usual the value of the correlation lies between -1 and 1.

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations.

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

Feature Scaling important step performed during data preparation especially when we perform multiple Linear Regression when there are multiple features to be compared which could be in different scale. So if the values are the in different scale naturally their coefficients will also be in different scale which is not easy to be compared.

For example, some feature like Area will be in thousands or hundreds, where as other features like no of bathrooms, which take very small values like 1,2,3 etc. Also, the categorical variables that you encoded take either 0 or 1 as their values. Hence, it is important to have everything on the same scale for the model to be easily interpretable.

Min-Max Scaling also called as Normalization usually means to scale a variable to have a values between 0 and 1, while standardization transforms data to have a mean of zero and a standard deviation of 1. The advantage of Standardisation over the other is that it doesn't compress the data between a particular range as in Min-Max scaling. This is useful, especially if there is are extreme data point (outlier).

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

VIF which stands for Variance Inflation Factor can be infinite sometimes, infinite value indicates that the corresponding variable may be expressed perfectly and exactly by a linear combination of other variables.

VIF = 1 / (1 – R2)

When R2 reached 1, VIF reaches infinity.

R2 value of the model when that variable is regressed against all the other independent variables.

What is the Gauss-Markov theorem?

Gauss–Markov theorem states that the ordinary least squares estimator has the lowest sampling variance within the class of linear unbiased estimators, if the errors in the linear regression model are uncorrelated, have equal variances and expectation value of zero.

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.

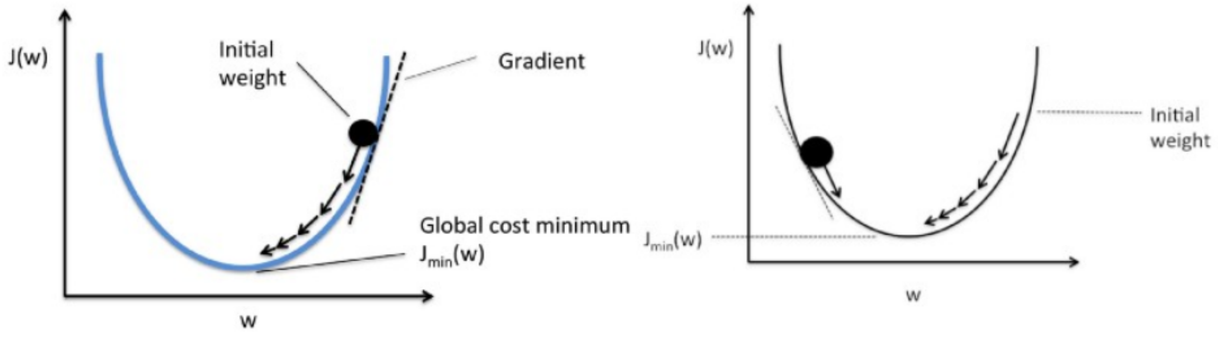
There are five Gauss Markov assumptions:

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. 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.

Explain the gradient descent algorithm in detail.

Gradient descent is an optimisation algorithm. In linear regression, it is used to optimise the cost function and find the values of the β (estimators) corresponding to the optimised value of the cost function.

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 works like a ball rolling down a graph (ignoring the inertia). The ball moves along the direction of the greatest gradient and comes to rest at the flat surface (minima).



Mathematically, the aim of gradient descent for linear regression is to find the solution of ArgMin J(Θ_0,Θ_1), where J(Θ_0,Θ_1) is the cost function of the linear regression. It is given by:

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$$

Here, h is the linear hypothesis model, $h = \Theta_0 + \Theta_1 x$, y is the true output, and m is the number of data points in the training set.

Gradient descent starts with a random solution, and then, based on the direction of the gradient, the solution is updated to the new value, where the cost function has a lower value.

Repeat until convergence:

$$\theta_j = \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}).\, x_j^{(i)} \quad \text{for j = 1,2,...,n}$$

Here ∝ is the learning speed generally defined as 0.01.

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

The Q-Q plot, or quantile-quantile plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. It plots quantiles of the data versus quantiles of a distribution.

The purpose of Q-Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q-Q plot; if both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.