1. Explain the linear regression algorithm in detail.
2. What are the assumptions of linear regression regarding residuals?
3. What is the coefficient of correlation and the coefficient of determination?
4. Explain the Anscombe's quartet in detail.
5. What is Pearson's R?
6. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
7. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
8. What is the Gauss-Markov theorem?
9. Explain the gradient descent algorithm in detail.
10. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

ANSWERS

1. **Linear Regression** is a machine learning algorithm based on **supervised learning**. It is perhaps one of the most well known and well understood algorithms in statistics and machine learning. It performs a **regression task**. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.
2. There are five assumptions of linear regressions -:
    a. Linear Relationship between the features and target.
    b. Little or No Multicollinearity between the features.
    c. Homoscedasticity Assumption.
    d. Normal distribution of error terms.
    e. Little or No autocorrelation in the residuals.
3. The quantity *r*, called the ***linear correlation coefficient***, measures the strength and the direction of a linear relationship between two variables. The linear correlation coefficient is sometimes referred to as the *Pearson product moment correlation coefficient* in honor of its developer Karl Pearson. The mathematical formula for computing ***r*** is:

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\left(\sum x^2\right) - \left(\sum x\right)^2}\sqrt{n\left(\sum y^2\right) - \left(\sum y\right)^2}}$$

The ***coefficient of determination***, r^2, is useful because it gives the proportion of the variance (fluctuation) of one variable that is predictable from the other variable. It is a measure that allows us to determine how certain one can be in making predictions from a certain model/graph. The *coefficient of determination* is the ratio of the explained variation to the total variation. The *coefficient of determination* is such that $0 \leq r^2 \leq 1$, and denotes the strength of the linear association between *x* and *y*.

4. **Anscombe's Quartet** was developed by statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough. It comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (*x*,*y*) points.
5. **Pearson's R**, the Pearson product-moment correlation coefficient or the bivariate correlation, is a measure of the linear correlation between two variables X and Y. It is denoted by r. Basically, a Pearson product-moment correlation attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient, *r*, indicates how far away all these data points are to this line of best fit (i.e., how well the data points fit this new model/line of best fit).
6. Scaling is an important technique in Machine Learning and it is one of the most important steps during the preprocessing of data before creating a machine learning model. This can make a difference between a weak machine learning model and a strong one. In scaling *(also called **min-max scaling**)*, you transform the data such that the features are within a specific range e.g. [0, 1]. 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. The two most important scaling techniques is Standardization and Normalization.

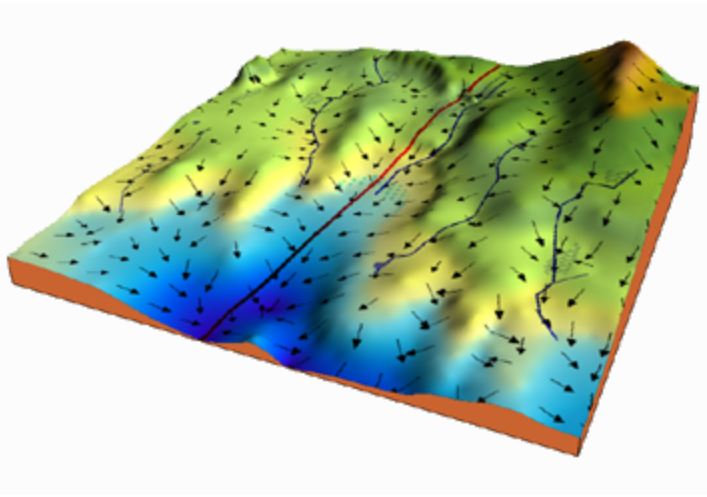
   Normalization: Normalization is the process of rescaling one or more attributes to the range of 0 to 1. This means that the largest value for each attribute is 1 and the smallest value is 0.

   Standardization: typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

7. The ***variance inflation factor*** *(VIF)* quantifies the extent of correlation between one predictor and the other predictors in a model. This is particularly problematic in two scenarios, where:
   1. The focus of the model is on making inferences regarding the relative importance of the predictors.
   2. The model is to be used to make predictions in a different data set, in which the correlations may be different.
8. The Gauss-Markov theorem states that if your linear regression model satisfies the first six classical assumptions, then ordinary least squares (OLS) regression produces unbiased estimates that have the smallest variance of all possible linear estimators.
9. Gradient descent is an optimization algorithm used to minimize some function by iteratively moving in the direction of steepest descent as defined by the negative of the gradient. In machine learning, we use gradient descent to update the parameters of our model. Parameters refer to coefficients in Linear Regression and weights in neural networks.

   Consider the 3-dimensional graph below in the context of a cost function. Our goal is to move from the mountain in the top right corner (high cost) to the dark blue sea in the bottom left (low

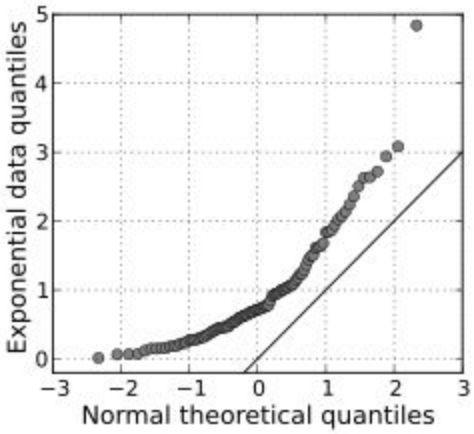
cost). The arrows represent the direction of steepest descent (negative gradient) from any given point–the direction that decreases the cost function as quickly as possible.



Starting at the top of the mountain, we take our first step downhill in the direction specified by the negative gradient. Next we recalculate the negative gradient (passing in the coordinates of our new point) and take another step in the direction it specifies. We continue this process iteratively until we get to the bottom of our graph, or to a point where we can no longer move downhill–a local minimum.

The size of these steps is called the *learning rate*. With a high learning rate we can cover more ground each step, but we risk overshooting the lowest point since the slope of the hill is constantly changing. With a very low learning rate, we can confidently move in the direction of the negative gradient since we are recalculating it so frequently. A low learning rate is more precise, but calculating the gradient is time-consuming, so it will take us a very long time to get to the bottom.

10. Q Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. 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 the two data sets come from a common distribution, the points will fall on that reference line.



A Q Q plot showing the 45 degree reference line. Image: skbkekas|Wikimedia Commons.

The image above shows quantiles from a theoretical normal distribution on the horizontal axis. It's being compared to a set of data on the y-axis. This particular type of Q Q plot is called a **normal quantile-quantile (QQ) plot.** The points are not clustered on the 45 degree line, and in fact follow a curve, suggesting that the sample data is not normally distributed.