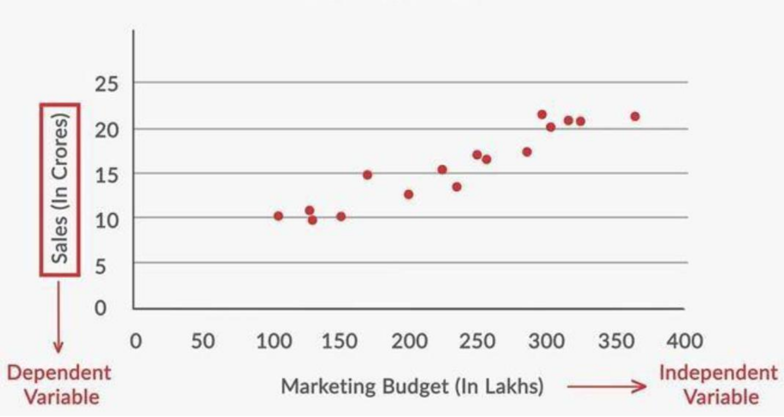
LR Subjective Question & Answer-

1. Explain the linear regression algorithm in detail.

Linear Regression is a machine learning algorithm based on **supervised 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.

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.
In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.

The standard equation of the regression line is given by the following expression: $Y = \beta_0 + \beta_1 X$ known as Regression Line.



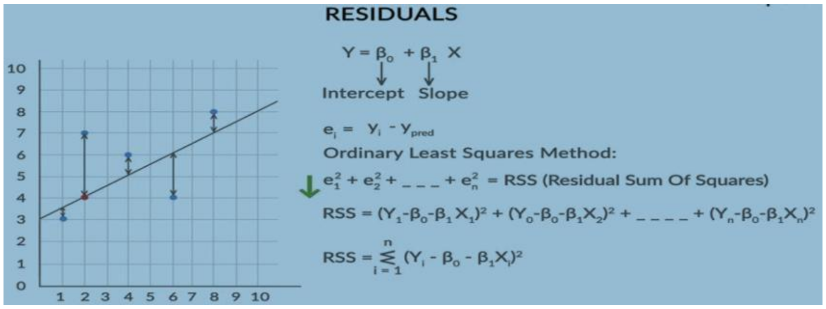
While training the model we are given:
x: input training data (univariate – one input variable(parameter))
y: labels to data (supervised learning)

When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best β_0 and β_1 values.

β_0: intercept
β_1: coefficient of X

Once we find the best β_1 and β_2 values, we get the best fit line. So, when we are finally using our model for prediction, it will predict the value of y for the input value of x.

The best-fit line is found by minimizing the expression of RSS (Residual Sum of Squares) which is equal to the sum of squares of the residual for each data point in the plot. Residuals for any data point is found by subtracting predicted value of dependent variable from actual value of dependent variable.



RESIDUALS

$Y = \beta_o + \beta_1 X$

$\downarrow$ Intercept $\downarrow$ Slope

$e_i = y_i - y_{pred}$

Ordinary Least Squares Method:

$\downarrow$ $e_1^2 + e_2^2 + \_\_\_ + e_n^2$ = RSS (Residual Sum Of Squares)

$RSS = (Y_1 - \beta_o - \beta_1 X_1)^2 + (Y_o - \beta_o - \beta_1 X_2)^2 + \_\_\_\_ + (Y_n - \beta_o - \beta_1 X_n)^2$

$RSS = \sum_{i=1}^{n} (Y_i - \beta_o - \beta_1 X_i)^2$

The strength of the linear regression model can be assessed using 2 metrics:

1. R^2 or Coefficient of Determination
2. Residual Standard Error (RSE)

There are two types of Linear Regression based on number of dependent variables:

1.Simple Linear Regression &
2.Multiple Linear Regression.

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

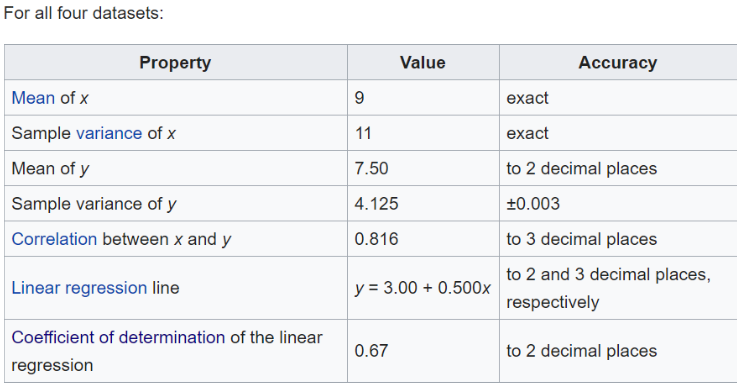
3. What is the coefficient of correlation and the coefficient of determination?

Coefficient of correlation is a statistical measure of the degree to which changes to the value of one variable predict change to the value of another. In positively correlated variables, the value increases or decreases in tandem. In negatively correlated variables, the value of one increase as the value of the other decreases. Correlation coefficients are expressed as values between +1 and -1.

The coefficient of determination is a measure used in statistical analysis that assesses how well a model explains and predicts future outcomes. It is indicative of the level of explained variability in the data set. The coefficient of determination, also commonly known as "R-squared," is used as a guideline to measure the accuracy of the model.
The coefficient of determination is used to explain how much variability of one factor can be caused by its relationship to another factor. It is relied on heavily in trend analysis and is represented as a value between 0 and 1.

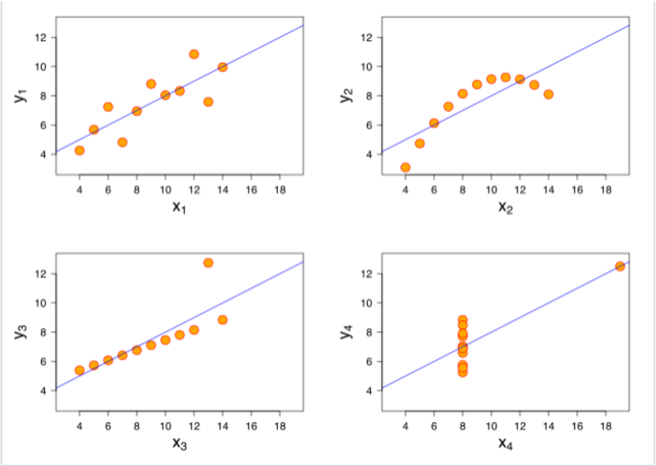
4. Explain the Anscombe’s quartet in detail.

Anscombe's quartet 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. They were constructed in 1973 by the statisticianFrancis 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."



For all four datasets:

Property	Value	Accuracy
Mean of *x*	9	exact
Sample variance of *x*	11	exact
Mean of *y*	7.50	to 2 decimal places
Sample variance of *y*	4.125	±0.003
Correlation between *x* and *y*	0.816	to 3 decimal places
Linear regression line	*y* = 3.00 + 0.500*x*	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression	0.67	to 2 decimal places



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated and following the assumption of normality.

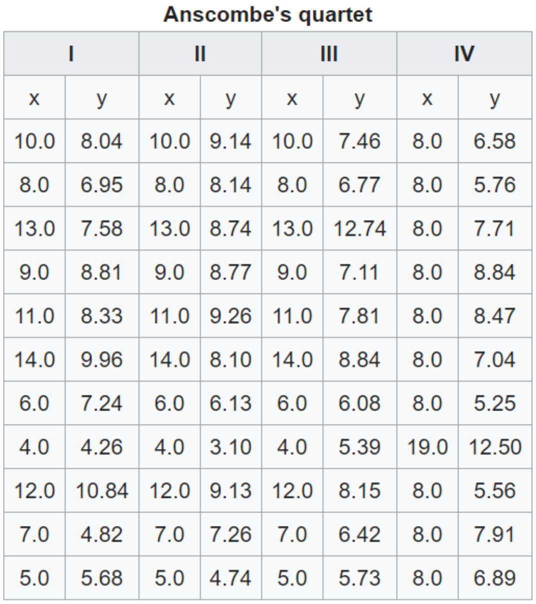
The second graph (top right) is not distributed normally; while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.

In the third graph (bottom left), the distribution is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.

Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

The quartet is still 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.

The datasets are as follows. The *x* values are the same for the first three datasets.



Anscombe's quartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

5. What is Pearson's R?

In statistics, the **Pearson correlation coefficient** (**PCC**) also referred to as **Pearson's *r***, the **Pearson product-moment correlation coefficient** (**PPMCC**) or the **bivariate correlation**.
Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier *product-moment* in the name.

Pearson's correlation coefficient when applied to a population is commonly represented by the Greek letter ρ (rho) and may be referred to as the *population correlation coefficient* or the *population Pearson correlation coefficient*. Given a pair of random variables(X,Y), the formula for ρ[7] is:

$$\boxed{\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} \qquad (Eq.1)}$$

where:
cov is the covariance
SIGMAx is the standard deviation of X
SIGMAy is the standard deviation of Y.

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. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step.

Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions will not work properly without normalization. For example, many classifiers calculate the distance between two points by the Euclidean distance. If one of the features has a broad range of values, the distance will be governed by this particular feature. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance.

Another reason why feature scaling is applied is that gradient descent converges much faster with feature scaling than without it.

Difference between Normalized & Standardized scaling-

In Normalized scaling the variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.

$$x = \frac{x - min(x)}{max(x) - min(x)}$$

In case of Standardized scaling the variables are scaled in such a way that their mean is zero and standard deviation is one.

$$x = \frac{x - mean(x)}{sd(x)}$$

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

Variance inflation factors show the degree to which a regression coefficient will be affected because of the variable's redundancy with other independent variables. As the squared multiple correlation of any predictor variable with the other predictors approaches unity, the corresponding VIF becomes infinite.

$$ V.I.F. = 1 / (1 - R^2). $$

If the VIF of a variable is infinite, it means there is one or more variable having perfect correlation with this particular variable. Hence you should drop the variable having infinite VIF.

For any predictor orthogonal (independent) to all other predictors, the variance inflation factor is 1.0. VIFi thus provides us with a measure of how many times larger the variance of the ith regression coefficient will be for multicollinear data than for orthogonal data (where each VIF is 1.0). If the VIF's are not unusually larger than 1.0, multicollinearity is not a problem. An advantage of knowing the VIF for each variable is that it gives a tangible idea of how much of the variances of the estimated coefficients are degraded by the multicollinearity. VIF's may be printed using the VI=Y option.

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.

In a regression model where $E\{i\} = 0$ and variance $\sigma\ 2\{i\} = \sigma\ 2 < \infty$ and i and j are uncorrelated for all i and j the least squares estimators b0 and b1 are unbiased and have minimum variance among all unbiased linear estimators.

The theorem states that b1 has minimum variance among all unbiased linear estimators of the form

$\beta\hat{}\ 1 = XciYi$

As this estimator must be unbiased we have

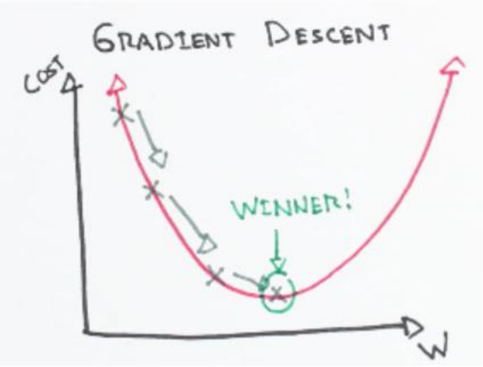
$E\{\beta\hat{}\ 1\} = Xci\ E\{Yi\} = \beta 1$

$=Xci(\beta 0 + \beta 1Xi) = \beta 0\ Xci + \beta 1\ XciXi = \beta 1$ 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 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.

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.



Key concepts are –

Learning Rate -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.

Cost Function-

A Loss Functions tells us "how good" our model is at making predictions for a given set of parameters. The cost function has its own curve and its own gradients. The slope of this curve tells us how to update our parameters to make the model more accurate.

Given the cost function:

$$f(m,b) = \frac{1}{N}\sum_{i=1}^{n}(y_i - (mx_i + b))^2$$

The gradient can be calculated as:

$$f'(m,b) = \begin{bmatrix} \frac{df}{dm} \\ \frac{df}{db} \end{bmatrix} = \begin{bmatrix} \frac{1}{N}\sum -2x_i(y_i - (mx_i + b)) \\ \frac{1}{N}\sum -2(y_i - (mx_i + b)) \end{bmatrix}$$

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

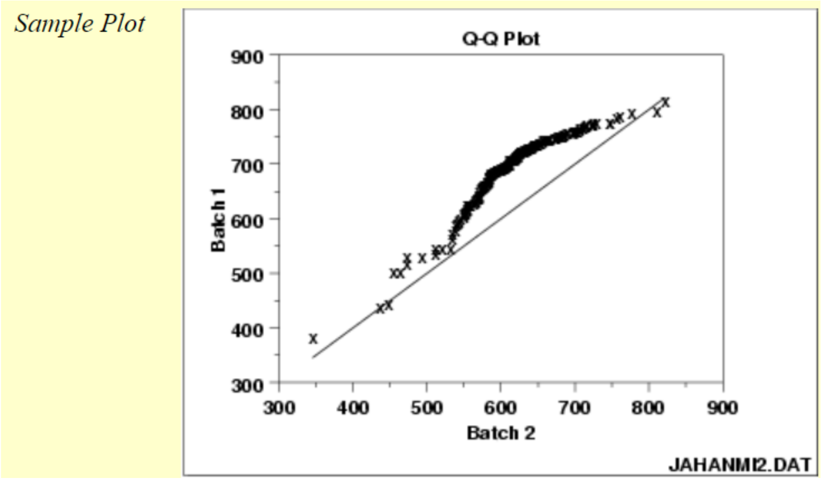
The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

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.

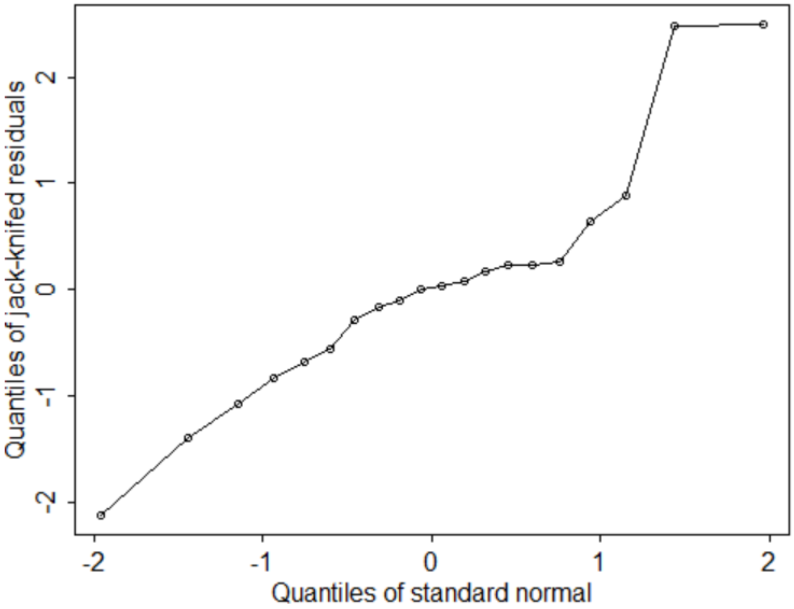
A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come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

The advantages of the q-q plot are:

The sample sizes do not need to be equal.
Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot. For example, if the two data sets come from populations whose distributions differ only by a shift in location, the points should lie along a straight line that is displaced either up or down from the 45-degree reference line.
The q-q plot is like a probability plot. For a probability plot, the quantiles for one of the data samples are replaced with the quantiles of a theoretical distribution.



In general, Q-Q plots showing curvature indicate skew distributions, with downward concavity corresponding to negative skewness (long tail to the left) and upward concavity indicating positive skewness. On the other hand, S-shaped Q-Q plots indicate heavy tails, or an excess of extreme values, relative to the normal distribution.


Quantiles of jack-knifed residuals
Quantiles of standard normal