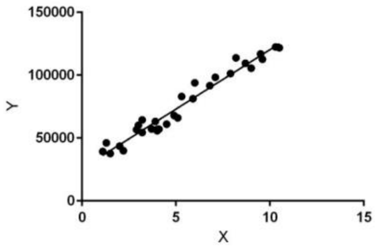
SUBJECTIVE QUESTIONS

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.

Hypothesis function for Linear Regression:

$$y = \theta_1 + \theta_2.x$$

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 θ_1 and θ_2 values.
θ_1: intercept
θ_2: coefficient of x
Once we find the best θ_1 and θ_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.
By achieving the best-fit regression line, the model aims to predict y value such that the error difference between predicted value and true value is minimum. So, it is very important to update the θ_1 and θ_2 values, to reach the best value that minimize the error between predicted y value (pred) and true y value (y).
Cost function (J) of Linear Regression is the **Root Mean Squared Error (RMSE)** between predicted y value (pred) and true y value (y).

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

1. R² or Coefficient of Determination
2. Residual Standard Error (RSE)

RSS (Residual Sum of Squares): In statistics, it is defined as the total sum of error across the whole sample. It is the measure of the difference between the expected and the actual output. A small RSS indicates a tight fit of the model to the data. It is also defined as follows:

$$RSS = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2$$

TSS (Total sum of squares): It is the sum of errors of the data points from mean of response variable. Mathematically, TSS is:

$$\mathbf{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

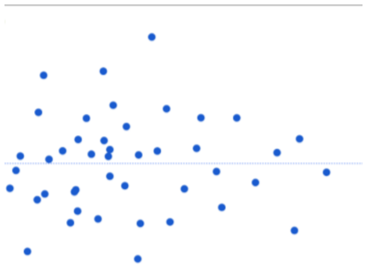
Importance of RSS/TSS:

Think about it for a second. If you know nothing about linear regression and still have to draw a line to represent those points, the least you can do is have a line pass through the mean of all the points as shown below.

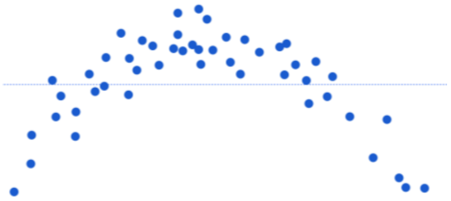
2. What are the assumptions of linear regression regarding residuals?

There are mainly three assumptions of regression regarding residuals, they are :-

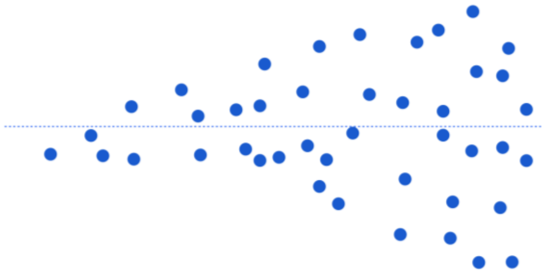
1. Error term are normally distributed with mean term equal to zero.



2. Error term are independent of each other i.e., error of one term is not related with the error of the second.



3. Error term has constant variance which it should be **homoscedasticity not heteroscedasticity** i.e. variance should not change as error value increases or decrease as the error value change and variance should not follow any pattern as the error term change.



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

Coefficient of correlation: - Coefficient of correlation is a numerical measure type of relationship between two variable i.e. how one variable is correlated with other variable. Coefficient of correlation ranges between -1 and 1. Coefficient of correlation equal to 1 that both variable are positively correlated that means if one variable increase other variable also increases and if Coefficient of correlation is -1 that mean both variable are highly negatively correlated that means if one variable increase other variable decreases. A coefficient of zero indicates there is no discernable relationship between fluctuations of the variables .so we can conclude that Coefficient of correlation is statistical measure to see how change in one value changes the value of other variable.

Coefficient of determination: - In statistic coefficient of determination or R^2 is measure through which we can access how good the linear regression model is in prediction of the dependent variable.
Coefficient of determination explain the variance of the dependent variable that is predicted or explained by linear regression model. Coefficient of determination value indicate how a model is

good fit. Let understand with an example if value of r^2 is .40 indicates that 40 percent of the variation in the outcome has been explained just by predicting the outcome using the covariates included in the model

.Higher the value r^2 better the fit of the line and better is the model. We can add more independent variable to increase the value r^2 and make prediction of our linear regression model better.

4. **Explain the Anscombe's quartet in detail.**

Anscombe's Quartet

Perhaps the most elegant demonstration of the dangers of summary statistics is Anscombe's Quartet. It's a group of four datasets that appear to be similar when using typical summary statistics, yet tell four different stories when graphed. Each dataset consists of eleven (x,y) pairs as follows:

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

All the summary statistics you'd think to compute are close to identical:

The average x value is 9 for each dataset

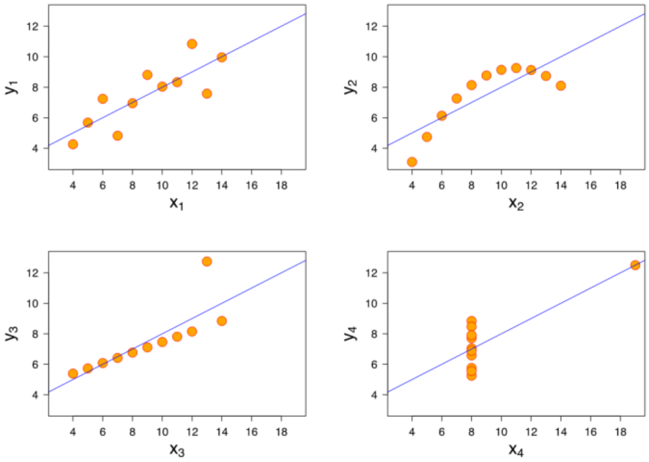
The average y value is 7.50 for each dataset

The variance for x is 11 and the variance for y is 4.12

The correlation between x and y is 0.816 for each dataset

A linear regression (line of best fit) for each dataset follows the equation y = 0.5x + 3

So far these four datasets appear to be pretty similar. But when we plot these four data sets on an x/y coordinate plane, we get the following results:



Now we see the real relationships in the datasets start to emerge. Dataset I consists of a set of points that appear to follow a rough linear relationship with some variance. Dataset II fits a neat curve but doesn't follow a linear relationship (maybe its quadratic?). Dataset III looks like a tight linear relationship between x and y, except for one large outlier. Dataset IV looks like x remains constant, except for one outlier as well. Computing summary statistics or staring at the data wouldn't have told us any of these stories. Instead, it's important to visualize the data to get a clear picture of what's going on.

5. What is Pearson's R?

Pearson's R is a numerical term that measure the strength of linear relationship between two variable. How good one variable is related to another variable and we can predict change in one variable how influence the change in other variable. Pearson's R is also known as Pearson product-moment correlation coefficient (PPMCC). Pearson's R value is from -1 to +1, +1 total positive linear correlation and -1 means total negative linear correlation and 0 is no linear relationship at all.

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

Scaling is feature which is used to normalize the range of the variable in the data set. we can scaling is procedure which is used to convert range of different variable in to specified range for all the variable. We understand scaling better by taking an example of geely automotive model wherein we have used feature scaling method to scale variables like (fueltype, doornumber, enginelocation and companymidtier) etc. Here we say a huge difference in the range of the variable and this difference in the range of the variable will affect the coefficients as obtained by fitting the regression model might be very large or very small as compared to the other coefficients and it became annoying at the time of model evaluation so it is advised to scale all the variable accordingly.

Difference between normalization (min-max) scaling and standardized scaling can be found below:-
Normalization: (x-xmin)/(xmax-xmin)
Standardization: (x-mu)/sigma

In normalization scaling method we scale our variable on the basis of min max scaler using above formula .if the variable is equal to xmax then equation (xmax-xmin)/(xmax-xmin) which is equal to 1 and if variable equal to min then (xmin-xmin)/(xmax-xmin) is equal to zero so we can say that when we normalize scaling value of the variable is between 0 and 1.

In Standardization scaling technique we use mean and standard deviation to scale the variable.in this process we try proceed values of each feature in the data have zero-mean (when subtracting the mean in the numerator) and unit-variance.

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

Variance inflation factor is method used in linear regression to find out how one variable is correlated to another variable.
The formulae of VIF is given by:-
VIF=1/(1-r^2)
1-r^2 = 1/VIF (here VIF is infinite)
R^2 = 1
When one variable is highly correlated with another variable, VIF tends to 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, that is, the estimator that has the smallest variance among those that are unbiased and linear in the observed output variables. Best mean that the lowest variance of the estimator is given by OLS when compared to other unbiased linear estimator. The error do not need to be identically distributed neither need to be independent as well as error do not need to be normal.

9. **Explain the gradient descent algorithm in detail.**

For finding the minimum value of a function we use a first order iterative optimization algorithm which is known as gradient descent. Gradient descent method try to reduce the cost function.

Gradient descent is an iterative method of optimizing an objective function (cost function), by moving toward the negative of the gradient. We can compute minimal $\theta1$ using below formula:

$\theta1=\theta0-\eta(\partial/\partial\theta(J(\theta)))$ where η is the learning rate.

We can understand it better with an example **y=x^2**. When we wanted to find the minimum value then we start with random value of $\theta0=10$ we iterate toward the minimum value. We use the above formula to get to the minimum value taking learning rate 0.1

$(\partial/\partial\theta(J(\theta)))$ of $x^2=2x$ $\theta1 = 10-0.1(2*10)=8$

$\theta2 =8 -0.1(2*8)=6.4$

$\theta3= 6.4 - 0.1(2*6.4)=5.12$

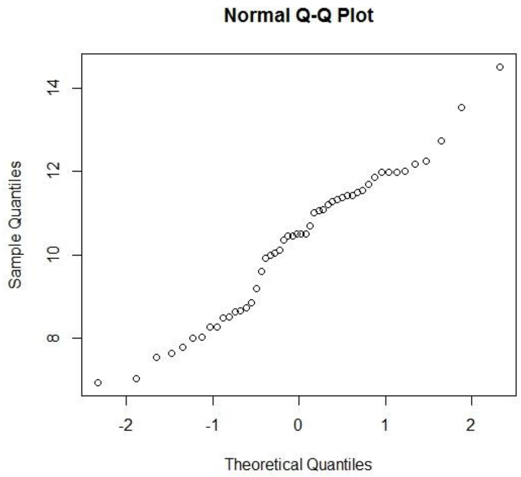
Here negative value represent that we have to move in opposite direction of the function to get the minimum value and learning rate define the rate at which we iterate through. If the learning rate is very low we have to iterate many time to reach which will be very time consuming and if we keep the learning rate high or value will fluctuate between positive and negative value just like a pendulum so we should be very careful in deciding the learning rate. If we continue the above procedure we will get our minimum value of the function through gradient descent.

10. 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 graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential. For example, if we run a statistical analysis that assumes our dependent variable is Normally distributed, we can use a Normal Q-Q plot to check that assumption. It’s just a visual check, not an air-tight proof, so it is somewhat subjective. But it allows us to see at-a-glance if our assumption is plausible, and if not, how the assumption is violated and what data points contribute to the violation.

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that’s roughly straight. Here’s an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.

From these graph we can see that sample quantile is linearly distributed with the theoretical quantile value and we can fit a straight line which show value are normally distributed. Q-Q plot are used to find where the data set have normal distribution or not through graphical method.


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles