Linear Regression Subjective Questions:

1. **Explain the linear regression algorithm in detail.**

   Linear regression is an algorithm used to find the relation between two continuous variables. Here one is independent and other is dependent variables. If there is one independent variable then it is called Simple Linear Regression else it is Multiple Linear Regression. Simple linear regression follows the equation of Y= β0+ β1X where as Multiple Linear regression follows Y= β0+ β1X1+ β2X2+…..

   Here β0 is the intercept (the place where line touches the y-axis) and
   β1 is called Coefficients and X1,X2…. are called variables.

   In Linear Regression, dependent variable is always continuous. While implementing Linear Regression should satisfy few assumptions then only we can say that Linear Regression is suitable for that scenario. There are evaluation metrics of RMSE, R2, Adj R2 and Prob(F-Statistic) will tell us how much good our model is.

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

   Mean of the Residual is zero

   Errors should be Normally distributed.

   There should be no auto-correlation between errors

   Error terms should have constant variance.

   No correlation between independent variables and residuals

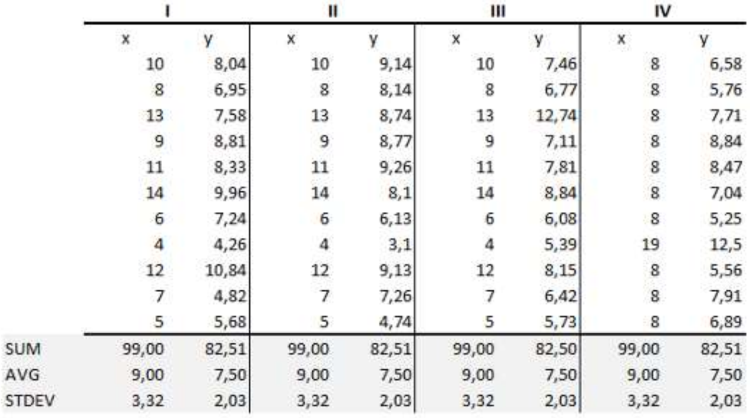
3. **What is the coefficient of correlation and the coefficient of determination?**

Coefficient of Correlation : It mainly tells us how well two variables are in unison. If both are increasing then we say that they both are positively correlated. If both are decreasing then we can say that both are negatively correlated. It ranges from -1 to 1 where -1 stands for negative correlation and 1 stands for positive correlation. We are using Pearson's correlation which is denoted by (R), if we do the square of it then it will become coefficient of determination(R2).

Coefficient of Determination(R2): It tells us how much variance in the data can be explained by the independent variables. It ranges from 0-1. It will never be negative. It is easier to interpret in regression problems.

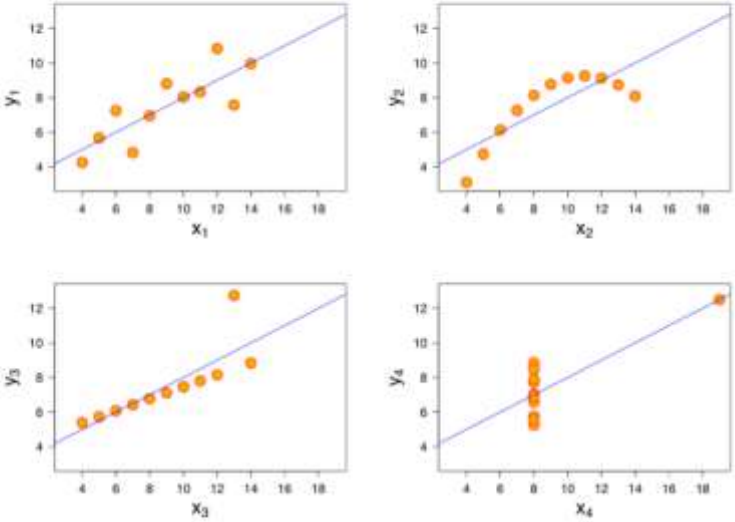
4. **Explain the Anscombe's quartet in detail.**

   Anscombe's Quartet was developed by statistician Francis Anscombe and it consists of four datasets each contain 11 (x,y) pairs. Each will have same descriptive statistics but when they plot the graph then their visualization is totally different from one other. Let's see the below image to identify the datasets



	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

So, from the above datasets it is clear that sum , Avg and Standard deviation is same for all the four datasets. But if plot them on x,y plane then each dataset tells us a different story. Let's understand it better from the below graphs



So, here

first plot shows that points are near to the regression line and the model is good.

Second plot shows that it is not normally distributed and it follows exponential trend

Third plot shows that it is linear but having outliers.

Fourth plot shows that one outlier is enough to produce a high correlation coefficient.

So, mainly Anscombe's quartet states the importance of visualization in data analysis. Since looking at the data tells us clear picture and its structure.

5. **What is Pearson's R?**

Pearson's R is a statistical measure to find the relation between two continuous variables. It is the best method to identify the association between two variables of interest as it is using covariance in the formula.

Correlation = Covariance(X,Y) / SQRT(Var(X)* Var(Y))

It also gives us magnitude of the associations as well as direction of the relation.

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

Scaling is a process to normalize the independent variables. Scaling makes the data to be in the same unit irrespective of their value.

There are different types of scaling techniques. Normalized scaling makes the data to be normally distributed and It ranges the value between 0-1 where as Standardized scaling refers to, it scales the data having mean of 0 and standard deviation of 1.

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

VIF = 1/(1-R2)

If there are any two variables highly correlated then the value will be 1.

So VIF = 1/(1-1) ➔1/0 ➔ infinity

So, the squared multiple correlation of any predictor variable with other predictors approaches unity then VIF will be infinity.

8. **What is the Gauss-Markov theorem?**

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. The sampling distributions are centered on the actual population value and are the tightest possible distributions. Finally, these aren't just the best estimates that OLS can produce, but the best estimates that any linear model estimator can produce. It famously states that OLS is BLUE means Best Linear Unbiased Estimator.

There are five Gauss Markov assumptions (also called conditions):

Linearity: Parameters we are estimating using the OLS method must be themselves linear.

Random: Our data must have been randomly sampled from the population.

Non-Collinearity: Regressors being calculated aren't perfectly correlated with each other.

Exogeneity: The regressors aren't correlated with the error term.

Homoscedasticity: No matter what the values of our regressors might be, the error of the variance is constant.

Gauss Markov assumptions guarantee the validity of OLS for estimating regression coefficients. Checking how well our data matches these assumptions is important part of estimating regression coefficients. We can plan to change our experiment if there is any violation. In practice, Gauss Markov assumptions are rarely all met. But still they are considered as a bench mark for ideal situations.

9. Explain the gradient descent algorithm in detail.

Cost function gives us the best fit line to our dataset by minimizing the errors and it does by taking the help of optimizer. Gradient descent is an optimizer to reduce the cost function. It's main principle is to find the global optimum to minimize the cost function.

Functionality of Gradient Descent:

It first initialize the weight at a point and draw tangential line to it(slope). It takes the derivative of the slope and the derivative gives the direction to which the point should move.

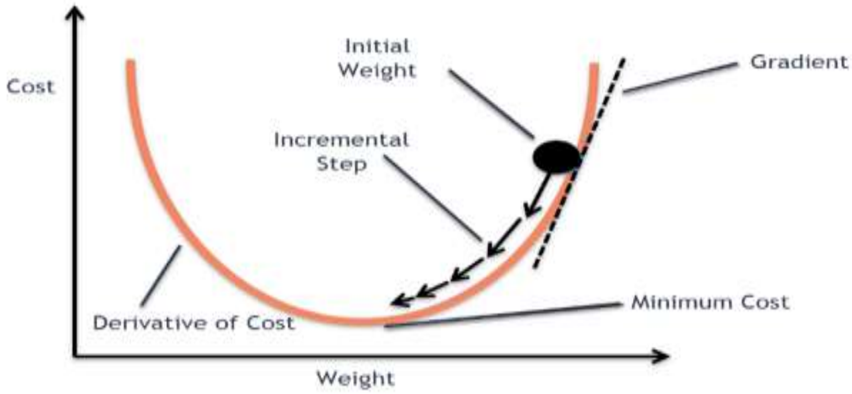


Figure1: Gradient Descent Algorithm [2]

Now there is a learning rate parameter(α) which gives the size of the step to be taken to reach global optimum. If alpha is too large then gradient descent will overshoots and fail to reach global optimum. If alpha is too small then it takes baby steps to reach global optimum. Even though if we fix alpha to a particular value still, alpha will changes automatically when it is reaching global optimum to converge.

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

Q-Q Plot: It is a technique to identify whether two datasets can fit with the same distribution or not. It is a plot of quantiles of first dataset against quantiles of second dataset. Here quantile mean the fraction of the points below the given value. 0.4 or 40% quantile is the point at which data falls below 40% and 60% falls above that value. A reference line at 45 degrees also plotted. If two datasets has same distribution then both datapoints will be falls close to that reference line. The greater deviation from this reference line indicates that two datasets are not following the same distribution.

Use of Q-Q Plot:

1. Sample size doesn't need to be equal to use q-q plot. So it can easily apply to all sizes of dataset to identify their distribution.

2. Many distributional aspects like outliers, symmetry etc.. can also be checked simultaneously.

Importance of Q-Q plot in Linear Regression:

In linear regression after fitting the model q-q plot checks if the points lie approximately lie on the line, and if they don’t then the residuals are not normally distributed.