Linear Regression: Subjective Questions

Question 1: Explain the linear regression algorithm in detail.

Answer 1: Regression is a family of Machine Learning Models/Algorithms, in which
The output variable to be predicted is a continuous variable, e.g. scores of a student.
Regression falls under supervised learning methods – in which you have the previous years' data with labels and you use that to build the model.
Regression is the most commonly used predictive analysis model.
Accurately predicting future outcomes has applications across industries — in economics, finance, business, medicine, engineering, education and even in sports & entertainment are some use case scenarios. Given the wide range of applications and its critical importance, Regression forms an important mode.

Linear Regression are of two types:

1. **Simple Linear Regression** : 1 dependent variable (interval or ratio), 1 independent variable (interval or ratio or dichotomous)
2. **Multiple Linear Regression** : 1 dependent variable (interval or ratio) , 2+ independent variables (interval or ratio or dichotomous)

Simple Linear Regression
The simple linear regression which explains the relationship between a dependent variable and one independent variable using a straight line. The straight line is plotted on the scatter plot of these two points. The standard equation of the regression line is given by the following expression: $Y = \beta_0 + \beta_1.X$
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, found by subtracting predicted value of dependent variable from actual value of dependent variable

Multiple Linear Regression
It's a statistical technique to understand the relationship between one dependent variable and several independent variables (explanatory variables).
The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X.
The standard equation of the regression line is given by the following expression: $Y = \beta_0 + \beta_1.X1 + \beta 2.X2 + ….. + \beta n.Xn$

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

- **R2 or Coefficient of Determination** : R2 = 1 - (RSS / TSS)
    a. RSS (Residual Sum of Squares): It is defined as the total sum of error across the whole sample.
    b. TSS (Total sum of squares): It is the sum of errors of the data points from mean of response variable.

2. **Residual Standard Error (RSE):** It's the square root of the residual sum of squares divided by the residual degrees of freedom.

Question 2: What are the assumptions of linear regression regarding residuals?

Answer 2: The assumptions of simple linear regression were:

- Residual (error) terms should not have correlation between them. Absence of this phenomenon is known as Autocorrelation.
- Error terms are normally distributed (not X, Y) , with 0 mean
- Error terms are independent of each other
- Error terms have constant variance (homoscedasticity)

Question 3: What is the coefficient of correlation and the coefficient of determination?

Answer 3: Coefficient of correlation measure how strong a relationship is between two variables is. . The coefficient value ranges between -1 and 1, where:

- **1 indicates a strong positive relationship**: means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other.
- **-1 indicates a strong negative relationship** : means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other
- **A result of zero indicates no relationship at all**: means that for every increase, there isn't a positive or negative increase. The two just aren't related.

They are various ways to compute these, the most common being:

1. Pearson: Pearson R's used as a standard
2. Kendall: Kendall Tau correlation coefficient
3. Spearman : Spearman rank correlation

Coefficient of Determination also referred as R-squared is the proportion of the variance in the dependent variable that is predictable from the independent variable(s). It's way of checking the accuracy of our model. R2 is a number which explains what portion of the given data variation is explained by the developed model. It always takes a value between 0 & 1. In general term, it provides a measure of how well actual outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model, i.e. expected outcomes. Overall, the higher the R-squared, the better the model fits your data.
Mathematically, it is represented as: R2 = 1 - (RSS / TSS).

It mainly represents the percentage of the data that is closest to the line of the best fit. If the regression line passes exactly through every point on the scatter plot, it would be able to explain all of the variation. The further the line is away from the points, the less it is able to explain.

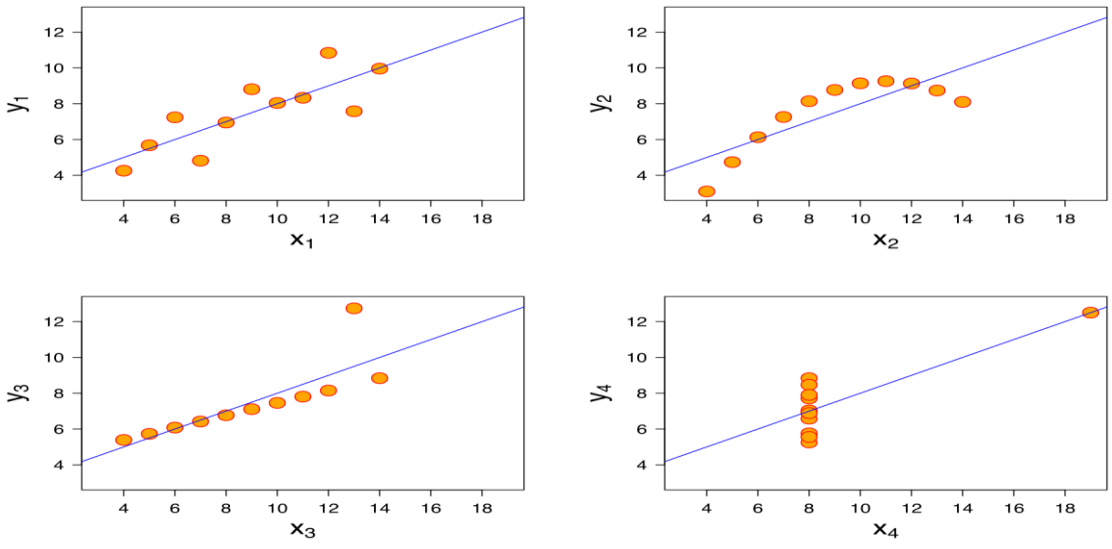
Question 4: Explain the Anscombe's quartet in detail.

Answer 4 : *Anscombe's Quartet* is a group of four datasets that appear to be similar when using typical summary statistics, yet tell four different stories when graphed .The pairs are :

	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

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:



- Dataset I appears to have clean and well-fitting linear models.
- Dataset II is not distributed normally.
- In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
- Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

Question 5: What is Pearson's R?

Answer 5: Pearson's R (or Pearson correlation coefficient, for short) is a measure of the strength of a linear association between two variables. The full name is the **Pearson Product Moment Correlation (PPMC)**. It shows the linear relationship between two sets of data. In simple terms, it answers the question, *Can I draw a line graph to represent the data?* Two letters are used to represent the Pearson correlation: Greek letter rho (ρ) for a population and the letter "r" for a sample.

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

Answer 6: Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step

We need to scale features because of two reasons:

1. Ease of interpretation
2. Faster convergence for gradient descent methods

We scale the features using two very popular method:

1. **Standardizing**: The variables are scaled in such a way that their mean is zero and standard deviation is one.
2. **MinMax 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.

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

Answer 7: VIF represents and measures co-linearity between predictor variables. The higher the VIF, the higher the multicollinearity. VIF = 1.0 means the independent variables are orthogonal to each other. VIF = infinity, there is perfect correlation between the variables. The common

heuristic for VIF is that while a VIF greater than 10 is definitely high, a VIF of greater than 5 should also not be ignored and inspected appropriately

Question 8: What is the Gauss-Markov theorem?

Answer 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. The Gauss-Markov theorem famously states that OLS is BLUE. BLUE is an acronym for Best Linear Unbiased Estimator

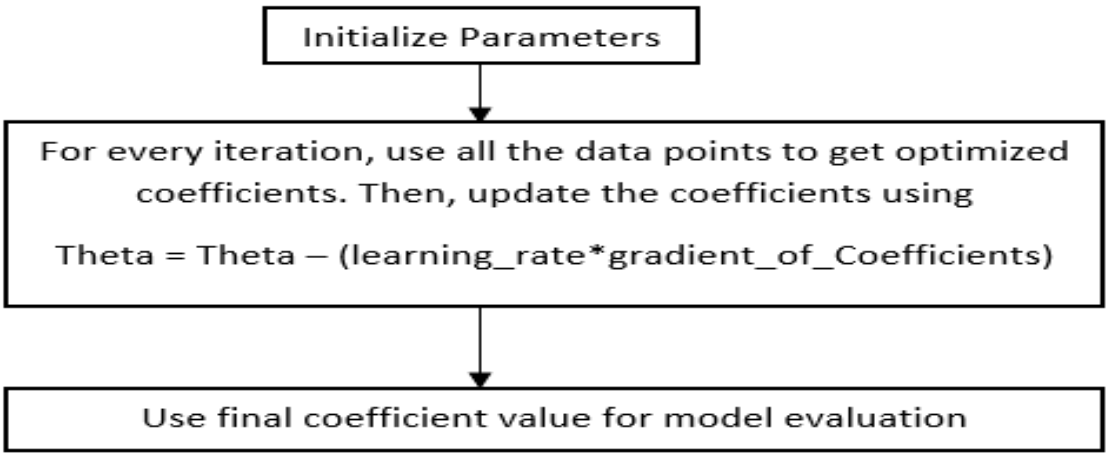
The classical assumptions are:

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. **Exogeneity**: 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

Question 9: Explain the gradient descent algorithm in detail.

Answer 9: Gradient descent algorithm is one of the most popular optimization algorithms for finding optimal parameters for the model.

It's an iterative machine learning optimization algorithm to reduce the cost function so that we have models that makes accurate predictions. Cost function(C) or Loss function measures the difference between the actual output and predicted output from the model. Cost function are a convex function.

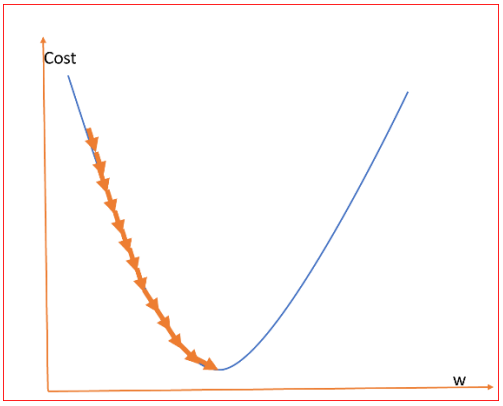


We randomly initialize all the weights/co-efficient for a model to a value close to zero but not zero. We calculate the gradient, $\partial c/\partial \omega$ which is a partial derivative of cost with respect to weight. α is learning rate, helps adjust the weights/coefficients with respect to gradient descent, we need to update the /coefficients for all the variables simultaneously.

Learning rate controls how much we should adjust the weights with respect to the loss gradient. Learning rates are randomly initialized.

$$w := w - \alpha \frac{\partial c}{\partial \omega}$$

Our goal is to minimize the cost function to find the optimized value for weights, we run multiple iterations with different weights and calculate the cost to arrive at a minimum cost as shown below :



Different types of Gradient descents are

- **Batch Gradient Descent**: The entire dataset to compute the gradient of the cost function for each iteration of the gradient descent and then update the weights.
- **Stochastic Gradient Descent**: We use a single data point or example to calculate the gradient and update the weights with every iteration. We first need to shuffle the dataset so that we get a completely randomized dataset.
- **Mini batch Gradient Descent**: It's a variation of stochastic gradient descent where instead of single training example, mini-batch of samples is used. Mini batch gradient descent is widely used and converges faster and is more stable. Batch size can vary depending on the dataset.

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

Answer 10 : 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. It's a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution

Q-Plot is used for residual analysis and validate a key assumption of Linear Regression, which is Error Terms (Residuals Errors) have a normal distribution. Q-plots are ubiquitous in statistics.

We plot a linear regression model, & check if the points are approximately on the line. If not, then residual aren't Gaussians and hence the errors, violating the assumption.