SUBJECTIVE QUESTIONS

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

   Linear regression is a linear model that derives a linear relationship between variables i.e. the input and output variables. This means that the output variable, y, can be calculated by using a linear combination of input variables, X.

   In case of one input variable, the model is called a simple linear regression. On the other hand, in case of multiple input variables, the model is called multiple linear regression.

   Different techniques can be used to train the linear regression model, the most common of which is called the **Ordinary Least Squares** method.

   <u>Representation of Linear Regression</u>

   The linear model as stated above includes a set of input and output variables. In addition to these variables, the model consists of scale factor for each input variable, called a **coefficient**. One additional coefficient is also added to give the line a degree of freedom, and is often called an **intercept** or the **bias coefficient**.

   For example – y = B1.x + B0

   Here, y is the output variable. B1 is the coefficient, x is the input variable and B0 is the bias coefficient

   <u>Linear Regression Learning</u>

   1) <u>Ordinary Least Squares</u>

      This technique seeks to minimize the sum of squared residuals. This means that given a regression line through the data we calculate the distance from each data point to the regression line, square it, and sum all of the squared errors together. This is the quantity that ordinary least squares seeks to minimize.

      This approach treats the data as a matrix and uses linear algebra operations to estimate the optimal values for the coefficients. It means that all of the data must be available and you must have enough memory to fit the data and perform matrix operations.

      It is unusual to implement the Ordinary Least Squares procedure yourself unless as an exercise in linear algebra. It is more likely that you will call a procedure in a linear algebra library. This procedure is very fast to calculate.

   2) <u>Gradient Descent</u>

      When there are one or more inputs you can use a process of optimizing the values of the coefficients by iteratively minimizing the error of the model on your training data.

      This operation is called Gradient Descent and works by starting with random values for each coefficient. The sum of the squared errors are calculated for each pair of input and output values. A learning rate is used as a scale factor and the coefficients are updated in the direction towards minimizing the error. The process is repeated until a minimum sum squared error is achieved or no further improvement is possible.

      When using this method, you must select a learning rate (alpha) parameter that determines the size of the improvement step to take on each iteration of the procedure.

Gradient descent is often taught using a linear regression model because it is relatively straightforward to understand. In practice, it is useful when you have a very large dataset either in the number of rows or the number of columns that may not fit into memory.

3) Regularization
There are extensions of the training of the linear model called regularization methods. These seek to both minimize the sum of the squared error of the model on the training data (using ordinary least squares) but also to reduce the complexity of the model (like the number or absolute size of the sum of all coefficients in the model).
Two popular examples of regularization procedures for linear regression are:
   - Lasso Regression: where Ordinary Least Squares is modified to also minimize the absolute sum of the coefficients (called L1 regularization).
   - Ridge Regression: where Ordinary Least Squares is modified to also minimize the squared absolute sum of the coefficients (called L2 regularization).

These methods are effective to use when there is collinearity in your input values and ordinary least squares would overfit the training data.

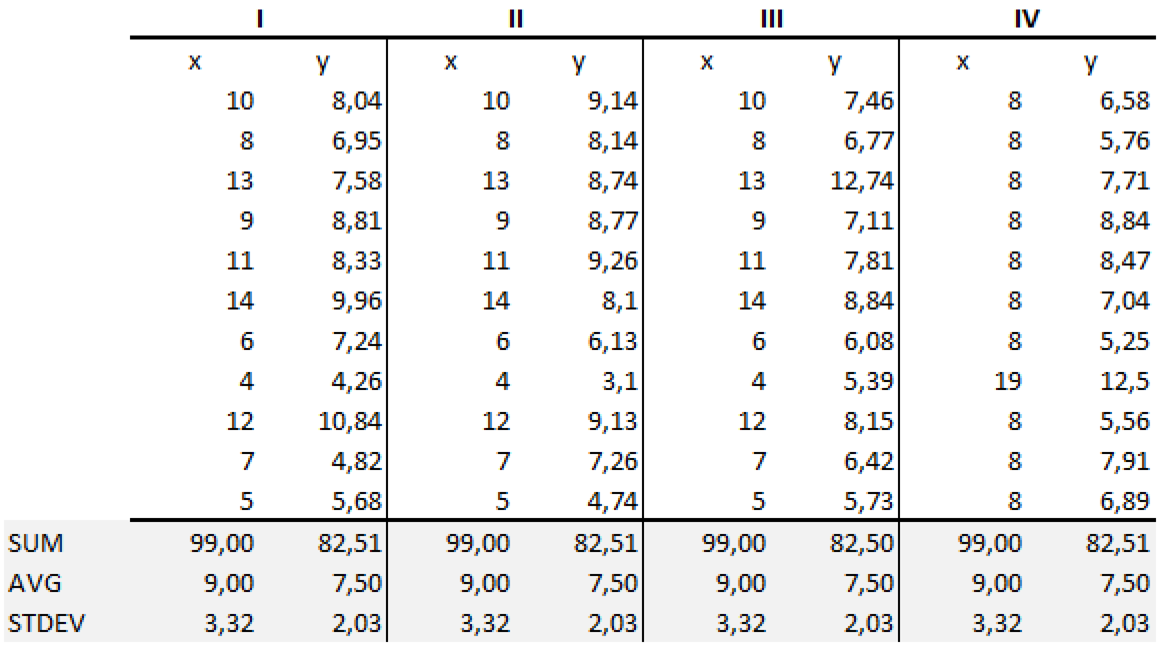
2. **What are the assumptions of linear regression regarding residuals?**
Following are the assumptions made for residuals in linear regression:
   1) Linearity of residuals
   On plotting a graph of residuals and y-values, the plot follows a linear pattern that shows linearity assumption is met.
   2) Independence of residuals
   There is no relation between the residual values of the data in a linear regression model
   3) Normal distribution of residuals
   If we draw a histogram of residuals and examine the normality of residuals, if the residuals are not skewed, that means that the assumption of linearity is satisfied.
   4) Equal variance of residuals
   On plotting the residual values against y-values, if the residuals do not fan out in a triangular fashion, that means that the equal variance assumption is met.

3. **What is the coefficient of correlation and the coefficient of determination?**
Coefficient of Correlation
The coefficient of correlation measures the strength and direction of linear relationship between two variables. It is sometimes referred to as the pearson product moment correlation coefficient.
The range of values of coefficient of correlation range from -1 to 1. A value of -1 refers to highly negative correlation among two variables, while a value of 1 represents a high positive correlation. A value of 0 signifies no linear correlation between the two variables.
The coefficient of correlation is represented by ‘r’.

Coefficient of Determination
The coefficient of determination gives a proportion of the variance 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.

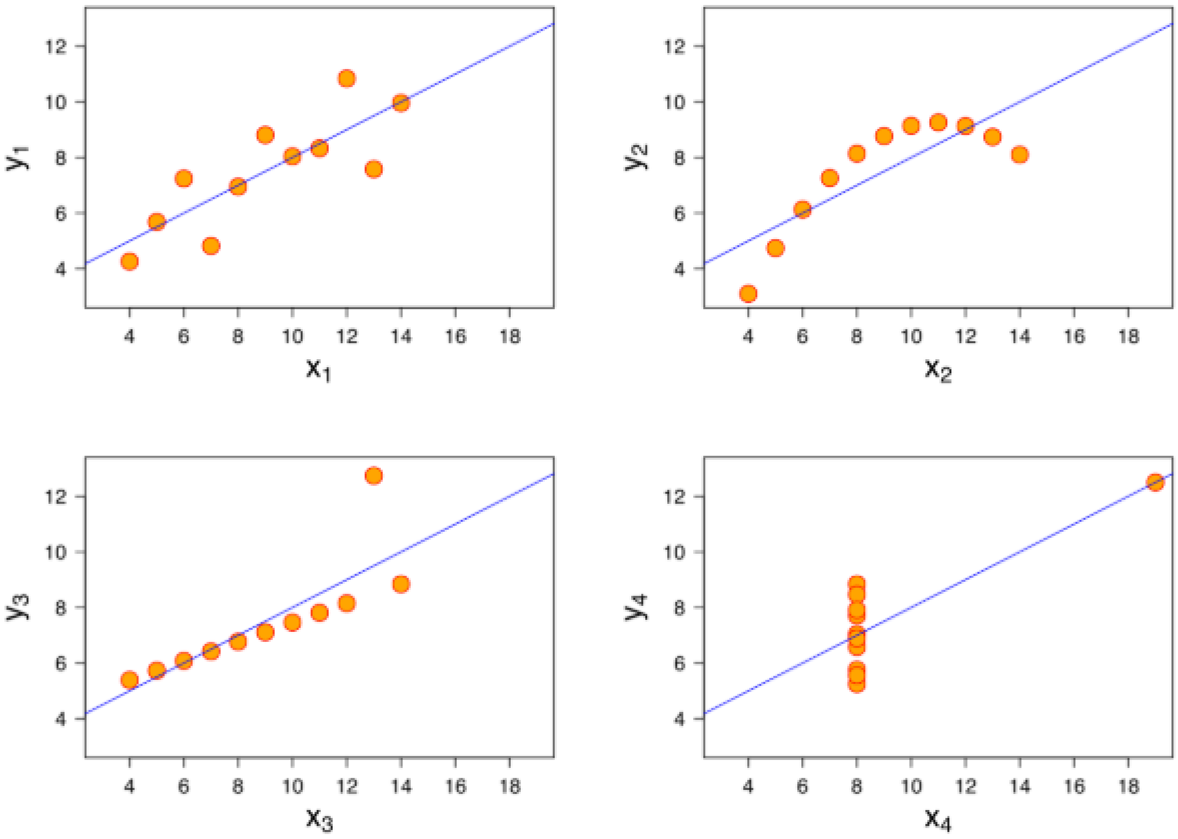
The value of coefficient of determination ranges from 0 to 1 and denotes strength of the linear relation between them.
Coefficient of determination is denoted by ‘r^2’.

4. **Explain the Anscombe’s quartet in detail.**
Anscombe’s Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x,y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.



	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

The summary statistics show that the means and the variances were identical for x and y across the groups. When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story.



- 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.

5. **What is Pearson's R?**
   Pearson's R is also called Pearson product moment correlation which is explained in question 3.
6. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**
   Real world dataset contains features that highly vary in magnitudes, units, and range. Scaling is performed to get these features at a common scale of magnitudes, units and range to avoid significance of one variable over other just based on their units or range.

There are two types of scaling:

a) Normalized Scaling

Normalization rescales the values into a range of [0,1]. This might be useful in some cases where all parameters need to have the same positive scale. However, the outliers from the data set are lost.

$$X_{changed} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

b) Standardized Scaling

Standardization rescales data to have a mean (μ) of 0 and standard deviation (σ) of 1 (unit variance).

$$X_{changed} = \frac{X - \mu}{\sigma}$$

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

The **variance inflation factor** *(VIF)* quantifies the extent of correlation between one predictor and the other predictors in a model. It is used for diagnosing *collinearity/multicollinearity*. Higher values signify that it is difficult to impossible to assess accurately the contribution of predictors to a model.

The extent to which a predictor is correlated with the other predictor variables in a linear regression can be quantified as the *R-squared* statistic of the regression where the predictor of interest is predicted by all the other predictor variables ( ). The *variance inflation* for a variable is then computed as:

$$VIF = \frac{1}{1 - R^2}$$

Thus, the VIF becomes infinite in a case when R-squared stat becomes 1. This happens when the predicted values exactly match the actual values of the output variables and there is no variance between these two variables.

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

The Gauss–Markov theorem states that in a linear regression model in which the errors are uncorrelated, have equal variances and expectation value of zero, the best linear unbiased estimator (BLUE) of the coefficients is given by the ordinary least squares (OLS) estimator, provided it exists.

Here "best" means giving the lowest variance of the estimate, as compared to other unbiased, linear estimators.

The errors do not need to be normal, nor do they need to be independent and identically distributed (only uncorrelated with mean zero and homoscedastic with finite variance). The requirement that the estimator be unbiased cannot be dropped, since biased estimators exist with lower variance. See, for example, the James–Stein estimator (which also drops linearity) or ridge regression.

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

Gradient descent is an optimization algorithm used to find the values of parameters (coefficients) of a function (f) that minimizes a cost function (cost). Gradient descent is best used when the parameters cannot be calculated analytically (e.g. using linear algebra) and must be searched for by an optimization algorithm.

Example –

Think of a large bowl like what you would eat cereal out of or store fruit in. This bowl is a plot of the cost function (f). A random position on the surface of the bowl is the cost of the current values of the coefficients (cost). The bottom of the bowl is the cost of the best set of coefficients, the minimum of the function.

The goal is to continue to try different values for the coefficients, evaluate their cost and select new coefficients that have a slightly better (lower) cost.

Repeating this process enough times will lead to the bottom of the bowl and you will know the values of the coefficients that result in the minimum cost.

Process of gradient descent

The procedure starts off with initial values for the coefficient or coefficients for the function. These could be 0.0 or a small random value.

coefficient = 0.0

The cost of the coefficients is evaluated by plugging them into the function and calculating the cost.

cost = f(coefficient)

or

cost = evaluate(f(coefficient))

The derivative of the cost is calculated. The derivative is a concept from calculus and refers to the slope of the function at a given point. We need to know the slope so that we know the direction (sign) to move the coefficient values in order to get a lower cost on the next iteration.

delta = derivative(cost)

Now that we know from the derivative which direction is downhill, we can now update the coefficient values. A learning rate parameter (alpha) must be specified that controls how much the coefficients can change on each update.

coefficient = coefficient – (alpha * delta)

This process is repeated until the cost of the coefficients (cost) is 0.0 or close enough to zero to be good enough.

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

    Q-Q (quantile-quantile) plots allow us to compare distributions by plotting their quantiles against each other.

    Importance of Q-Q plots:

    - Interpretation of the points on the plot: a point on the chart corresponds to a certain quantile coming from both distributions (again in most cases empirical and theoretical).
    - On a Q-Q plot, the reference line is dependent on the location and scale parameters of the theoretical distribution. The intercept and slope are equal to the location and scale parameters respectively.
    - A linear pattern in the points indicates that the given family of distributions reasonably describes the empirical data distribution.
    - Q-Q plot gets very good resolution at the tails of the distribution but worse in the center (where probability density is high)
    - Q-Q plots do not require specifying the location and scale parameters of the theoretical distribution, because the theoretical quantiles are computed from a standard distribution within the specified family.
    - The linearity of the point pattern is not affected by changing location or scale parameters.
    - Q-Q plots can be used to visually evaluate the similarity of location, scale, and skewness of the two distributions.