1. Explain the linear regression algorithm in detail.

Linear Regression is a statistical technique used to model the relationship between a dependent variable and one or more independent variables. The primary goal is to predict the dependent variable based on the independent variables. The model assumes a linear relationship between the variables.

- **Simple Linear Regression**: Involves one independent variable and is described by the equation:

  $$Y = \beta0+ \beta1x+ \epsilon$$

  where:

  - y is the dependent variable
  - $\beta0$ is the y-intercept
  - $\beta1$ is the slope of the line
  - x is the independent variable
  - ϵ is the error term
- **Multiple Linear Regression**: Involves two or more independent variables and is described by the equation:

  $Y = \beta0+\beta1x1+\beta2x2+\ldots+\beta nxn+\epsilon$

  where x1,x2,…,xn are independent variables and $\beta1,\beta2,\ldots,\beta n$ and $\beta1,\beta2,\ldots,\beta n$ are their respective coefficients.

Training the Model: The coefficients (β) are estimated using methods like Ordinary Least Squares (OLS), which minimizes the sum of squared residuals (the difference between observed and predicted values).

Prediction: Once the model is trained, it can predict the dependent variable using the estimated coefficients.

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

Linear regression makes several key assumptions about residuals (the differences between observed and predicted values):

1. **Linearity**: The relationship between the independent and dependent variables is linear.
2. **Independence**: Residuals are independent of each other. There should be no autocorrelation (especially in time series data).
3. **Homoscedasticity**: The residuals have constant variance across all levels of the independent variables. There should be no pattern in the spread of residuals.
4. **Normality**: Residuals should be normally distributed. This is important for hypothesis testing and confidence intervals.

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

Coefficient of Correlation (r): Measures the strength and direction of the linear relationship between two variables. It ranges from -1 to 1, where:

a. R = 1: Perfect positive linear relationship
b. R = −1: Perfect negative linear relationship
c. R = 0: No linear relationship

Coefficient of Determination (R^2): Indicates the proportion of the variance in the dependent variable that is predictable from the independent variables. It ranges from 0 to 1.where 1 means exaplains all the variance and zero means no variance explained at all.

4. Explain the Anscombe's quartet in detail.

Anscombe's Quartet is a set of four datasets that have nearly identical simple descriptive statistics but very different distributions and appearances when graphed. It was created by Francis Anscombe in 1973 to illustrate the importance of graphing data before analyzing it and to show how summary statistics alone can be misleading.

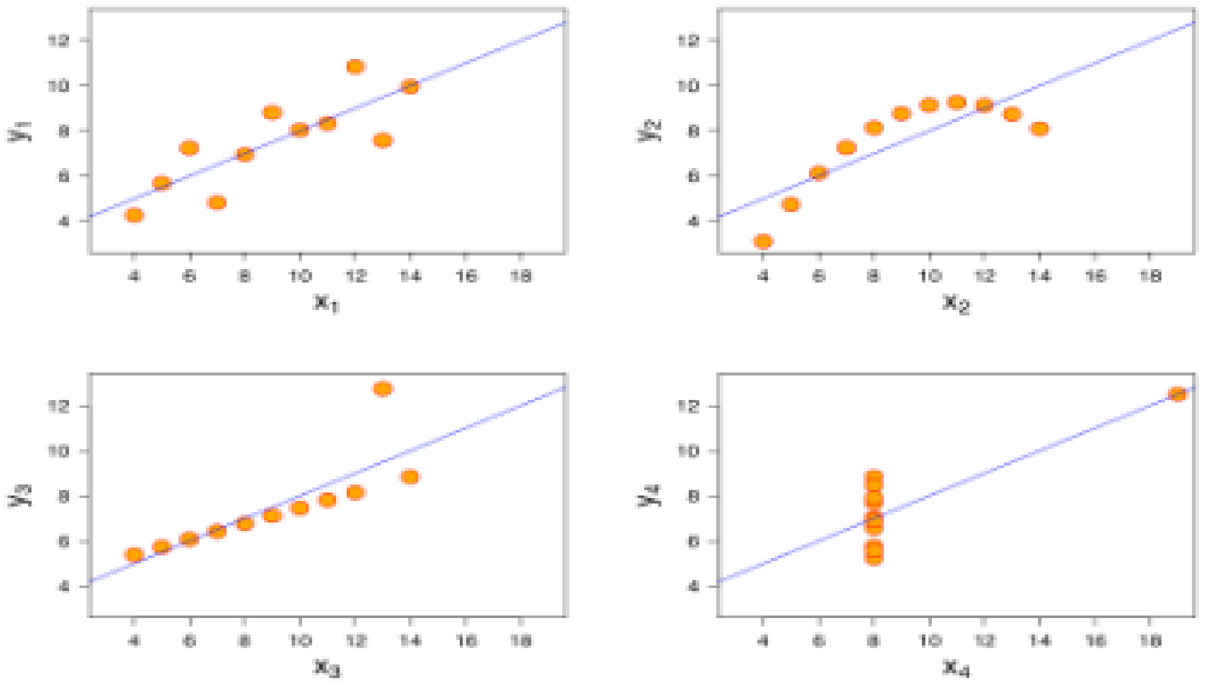
Each dataset in the quartet consists of 11 data points and has the following properties:

- **Mean of x**: 9
- **Mean of y**: 7.5
- **Variance of x**: 11
- **Variance of y**: 4.12
- **Correlation between x and y**: 0.816

However, the datasets reveal different relationships and trends when plotted:

1. **Dataset I**: Displays a linear relationship.
2. **Dataset II**: Shows a non-linear relationship with a clear quadratic trend.
3. **Dataset III**: Indicates a linear relationship with one outlier that significantly affects the fit.
4. **Dataset IV**: Shows a nearly linear relationship with a vertical line of points, revealing a strong influence from a single point.

The quartet emphasizes that while summary statistics (mean, variance, correlation) might be similar, the underlying data can be vastly different.



5. What is Pearson’s R?

Pearson’s R (or Pearson correlation coefficient) measures the strength and direction of the linear relationship between two continuous variables. It is a value between -1 and 1:

- **r = 1**: Perfect positive linear relationship
- **r = −1**: Perfect negative linear relationship

- **r = 0**: No linear relationship

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

Scaling is the process of adjusting the range and distribution of feature values. It is performed to ensure that all features contribute equally to the model and to improve the convergence of gradient-based optimization algorithms.

- **Normalized Scaling**: Adjusts the feature values to fit within a specific range, typically [0, 1]. It is done using Min-Max Scaling:

- **Standardized Scaling**: Transforms features to have zero mean and unit variance. It is done using Standardization:

nor**malization** is useful when features have different units or ranges and you want them to be on a common scale. **Standardization** is beneficial when the features have a normal distribution and you want to center them around the mean.

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

Variance Inflation Factor (VIF) measures the extent of multicollinearity in regression models. It quantifies how much the variance of a regression coefficient is inflated due to multicollinearity with other predictors.

A VIF value becomes infinite (or very large) when there is perfect multicollinearity. This means that one or more independent variables are perfectly correlated with each other, causing redundancy. As a result, the model cannot estimate unique contributions of each predictor, leading to an infinite VIF.

8. What is the Gauss-Markov theorem?

The **Gauss-Markov Theorem** states that, in a linear regression model where the errors have expectation zero, are homoscedastic (constant variance), and are uncorrelated, the Ordinary Least Squares (OLS) estimator has the smallest variance among all linear unbiased estimators. In other words, OLS estimators are the Best Linear Unbiased Estimators (BLUE) under these assumptions.

9. Explain the gradient descent algorithm in detail.

Gradient Descent is an iterative optimization algorithm used to minimize a loss function. It is commonly used in machine learning to find the optimal parameters for a model.

Steps in Gradient Descent:

1. **Initialize Parameters**: Start with random or zero values for the parameters.
2. **Compute the Gradient**: Calculate the gradient (partial derivatives) of the loss function with respect to each parameter.
3. **Update Parameters**: Adjust the parameters in the direction opposite to the gradient to reduce the loss. The update is done as: $\theta = \theta - \alpha \partial\theta\partial J(\theta)$ where θ represents the parameters, α is the learning rate, and $J(\theta)$ is the loss function.
4. **Iterate**: Repeat steps 2 and 3 until convergence (i.e., when the change in the loss function is minimal or when a predefined number of iterations is reached).

Learning Rate (α) determines the size of the step taken in the direction of the gradient. A learning rate that is too high might cause the algorithm to overshoot the minimum, while a rate that is too low might make the convergence very slow

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

A **Q-Q Plot** (Quantile-Quantile Plot) is a graphical tool used to assess if a dataset follows a specific theoretical distribution, typically the normal distribution. It compares the quantiles of the sample data to the quantiles of a theoretical distribution.

How to Interpret a Q-Q Plot:

- **If the points lie approximately along a straight line**: The data follows the specified theoretical distribution (e.g., normal distribution).
- **If the points deviate significantly from the line**: The data does not follow the specified distribution.

Importance in Linear Regression:

- **Normality of Residuals**: One of the assumptions of linear regression is that residuals are normally distributed. A Q-Q plot helps in diagnosing whether this assumption holds.
- **Model Validity**: Checking the normality of residuals ensures that the model's estimates and predictions are reliable and valid for statistical inference.