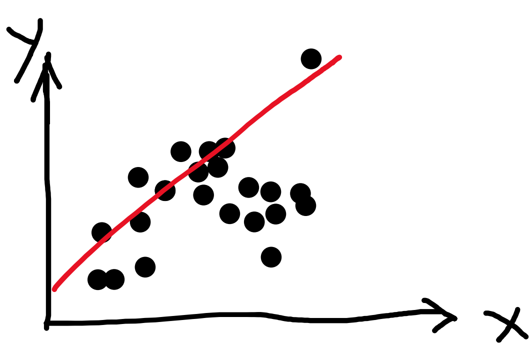
1. Explain the linear regression algorithm in detail.
2. What are the assumptions of linear regression regarding residuals?
3. What is the coefficient of correlation and the coefficient of determination?
4. Explain the Anscombe's quartet in detail.
5. What is Pearson's R?
6. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
7. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
8. What is the Gauss-Markov theorem?
9. Explain the gradient descent algorithm in detail.
10. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

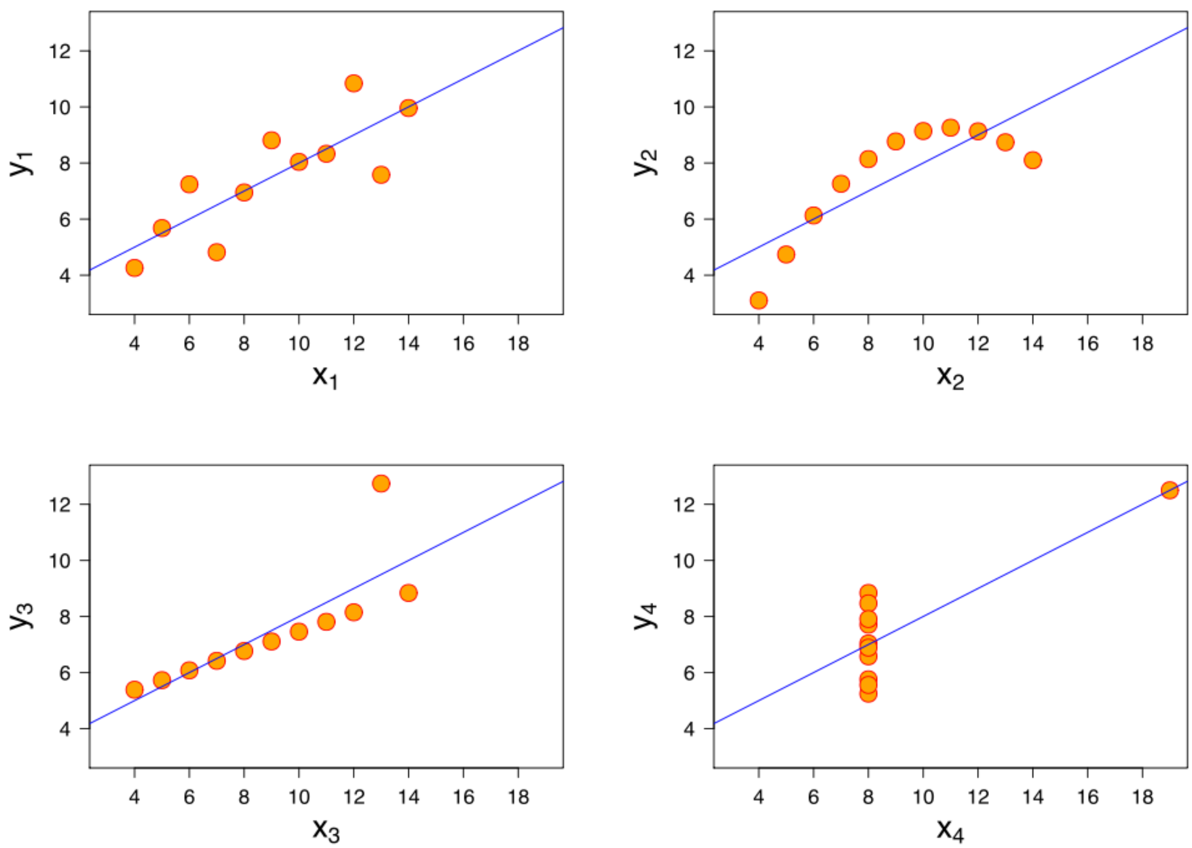
Answers

1. 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. It's used to predict values within a continuous range, (e.g. sales, price) rather than trying to classify them into categories (e.g. cat, dog). There are two main types:
   a. Simple Regression : Simple linear regression uses traditional slope-intercept form, where m and c and x represents our input data and y represents our prediction.
      - Y = mx + c
   b. Multivariable regression : These model is used for establishing the relationship between a dependent variable and more than one independent variable.
      - f(x,y,z) = w1x+w2y+w3z

In the given graph, this regression technique finds out a linear relationship between x (input) and y(output). It predicts a dependent variable value (y) based on a given independent variable (x). And the regression line (red line) is the best fit line for our model.



2. Multivariate Normality–Multiple regression assumes that the residuals are normally distributed. No **Multicollinearity**—Multiple regression assumes that the independent **variables** are not highly correlated with each other. This assumption is tested using Variance Inflation Factor (VIF) values.
3. **Coefficient of correlation** is “R” value which is given in the summary table in the Regression output. R square is also called **coefficient of determination**. Multiply R times R to get the R square value.
4. **Anscombe's quartet** comprises **four** data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x,y) points.



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed.
(Graphs used from Wikipedia)

- The first scatter plot appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.

- The second graph is not distributed normally; while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant.
- In the third graph, the distribution is linear, but should have a different regression line. The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

The datasets are as follows. The x values are the same for the first three datasets.

Anscombe's quartet

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

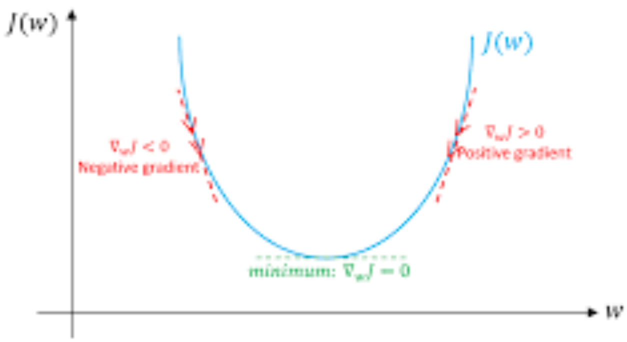
5. The Pearson product-moment correlation coefficient is a measure of the strength of the linear relationship between two variables. It is referred to as Pearson's correlation or simply as the correlation coefficient. Pearson's r can range from -1 to 1.

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. So, Feature Scaling is used to bring all values to same magnitudes and thus, tackle this issue.
   The difference is that, in scaling, you're changing the range of your data while in normalization you're changing the shape of the distribution of your data.

7. If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

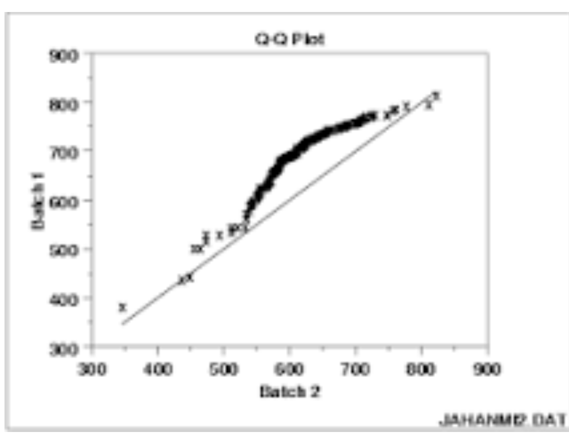
8. The Gauss-Markov theorem states that if your linear regression model satisfies the first six classical assumptions, then ordinary least squares (OLS) regression produces unbiased estimates that have the smallest variance of all possible linear estimators.

9. Gradient descent is a first-order iterative optimization algorithm for finding the local minimum of a differentiable function. To find a local minimum of a function using gradient descent, we take steps proportional to the negative of the gradient (or approximate gradient) of the function at the current point. 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.



(image from google)

10. The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.



(image from google)

If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight, used to check the validity of a distributional assumption for a data set. In general, the basic idea is to compute the theoretically expected value for each data point based on the distribution in question.