Q. Explain the linear regression algorithm in detail?

A. Linear Regression is an algorithm which is used to find the linear relationship between the dependent(target) variable and independent variables(predictors). It means if the value of independent variables is changed, the value of dependent variable is also changed accordingly.

It is represented by-

$$y = mx + c$$

Where Y is the dependent variable ,X is the independent variable,m is the slope of the regression line, b is a constant.

Two Types of regression-

1. Simple Linear Regression
2. Multiple Linear Regression

Simple Linear Regression

Simple linear regression is for finding relationships between two continuous variables. One is an independent variable and the other is dependent variable.

Example:

The dataset contains the two variables 'number of hours studied' and 'marks obtained'. 'Number of horse studied' is an independent variable and 'marks obtained' is dependent variable.

$$Y(pred) = b0 + b1*x$$

Multiple Linear Regression

Multiple linear regression is for finding relationships between more than one independent variable and one dependent variable.

Example

Consider a dataset with height, weight, bmi. Bmi is dependent variable and height and weight is independent variable

$$Y(Bmi) = b0 + b1 * height + b2 * weight$$

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

A. There are four assumptions of linear regression regarding residuals

1. **Normally Distribution**: If we draw a histogram of residuals, and find residuals are normally distributed and not skewed, it means the assumption is True.
2. **Zero mean assumption**: if we draw a histogram of residuals and find the error team are normally distributed around the zero, it means the assumption is True.
3. ***Constant variance assumption***: It means the residuals terms have the same variance,$\sigma 2$. It is also called homogeneity or homoscedasticity.
4. ***Independent error assumption:*** It means the residual terms are independent of each other, their pairwise covariance is zero.

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

A. Coefficient of correlation(r) means linear relationships between two variables x and y. It can be between -1 and 1. + sign indicates positive correlation and - sign indicated negative correlation. If it is zero then there is no correlation between the variables. while coefficient of determination (R-squared) explains the variation.The value of R^2 is between 0 and 1. R-squared distinguishes between the two variables based on their roles in the regression.

Q. Explain the Anscombe's quartet in detail.

A.

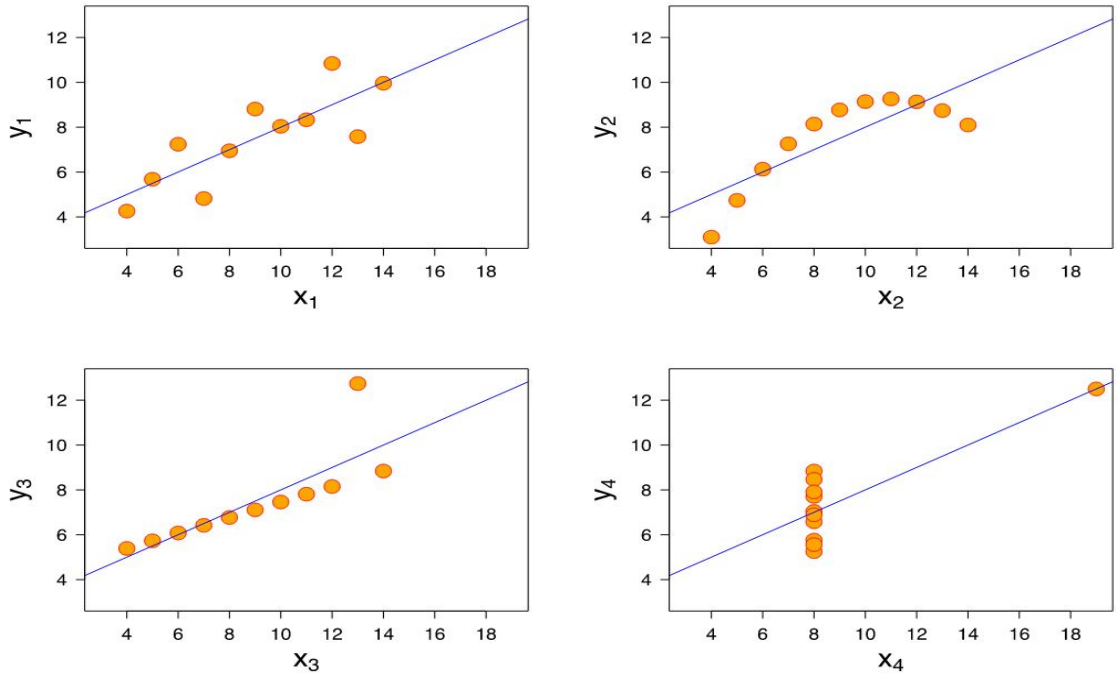
Consider 4 set with each 11 data-points

summary statistics for all the 4 sets

Summary

Set	mean(X)	sd(X)	mean(Y)	sd(Y)	cor(X,Y)
1	9	3.32	7.5	2.03	0.816
2	9	3.32	7.5	2.03	0.816
3	9	3.32	7.5	2.03	0.816
4	9	3.32	7.5	2.03	0.817

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:



- The first scatter plot is a simple linear relationship between the two variables.
- The second graph was not distributed normally,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.
- Finally, the fourth graph shows an example when one high-leverage point is enough to produce a high correlation coefficient.

So it is better to visualize the data to analyse.

Q. What is Pearson's R?

A. Pearson's correlation coefficient (r) is used to find the relationship between two variables.For ex- height and weight. It measures the strength of association between the variables. Pearson's correlation coefficient (r) data ranges from -1 to +1. We can find the correlation using the scatter plot.

r=-1 : A perfect straight line with a negative slope.
r=0 : no linear relationship between the variables.
r=1 : a perfect straight line with a positive slope

Positive correlation indicates that both variables increase or decrease together, whereas negative correlation indicates that as one variable increases, so the other decreases.

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

A. Scaling is technique which is used to standardize the data in a fixed range.It is generally performed during the data preprocessing step.If an algorithm is not using feature scaling method then it can consider higher value is high and smaller value is small, regardless of the unit of the values.
For Example: it considers 2000 meter is greater than 10km which is not true. This is the reason why scaling is performed

Normalization: This technique re-scales a feature with values between 0 and 1.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Standardization: It re-scales a feature value so that it has distribution with 0 mean value and variance equals 1.

$$x' = \frac{x - \bar{x}}{\sigma}$$

Standardization affected the dummy variables while normalization did not.

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

A. VIF indicates the correlation between the variables. If there is perfect correlation, then the VIF = infinity. Ex-

$$VIF = 1 / (1 – R2)$$

If R2 is more which means this feature is correlated with other features.When R2 reaches 1, VIF reaches infinity. Thus, always remove features with VIF > 5.

Q. What is the Gauss-Markov theorem?

A. It states that if assumptions are satisfied,then the ordinary least squares estimate for regression coefficients gives you the *best linear unbiased estimate (BLUE)* possible.

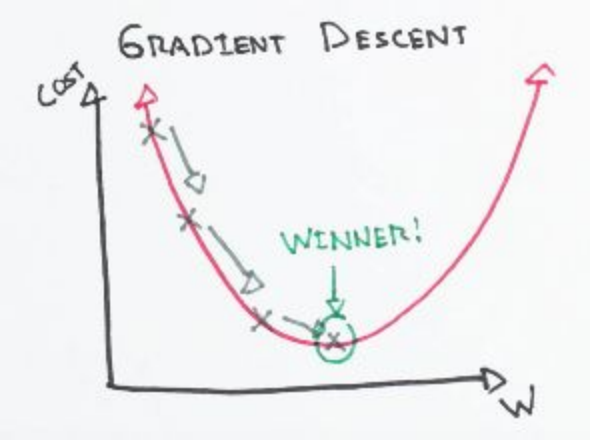
Assumptions of the Gauss - Markov theorem

1. **Linearity**: the parameters we are estimating must be linear.
2. **Random**: Data must have been randomly sampled from the dataset.
3. **Non-Collinearity**: variables are not perfectly correlated with each other.
4. **Exogeneity**: the regressors not be correlated with the error term.
5. **Homoscedasticity**: the error of the variance is always constant.

Q. Explain the gradient descent algorithm in detail?

A. Gradient descent is an optimization technique which is used to find the values of parameters (coefficients) of a function (f) that minimizes a cost function (cost).It iteratively moves in the direction of steepest descent as defined by the negative of the gradient.

Starting at the top point , we take our first step downhill in the direction specified by the negative gradient. Then again we recalculate the negative gradient and take another step in the direction it specifies. We continue this process iteratively until we get to the bottom of our graph, or to a point where we can no longer move downhill–a local minimum.



Algorithm for gradient descent :

Let $h_\theta(x)$ be the hypothesis for linear regression. Then, the cost function is given by:

Let Σ represents the sum of all training examples from i=1 to m.

$$J_{train}(\theta) = (1/2m)\ \Sigma(\ h_\theta(x_{(i)})\ -\ y_{(i)})^2$$

Repeat {

$\theta j = \theta j - (\text{learning rate}/m) * \Sigma(\ h_\theta(x_{(i)})\ -\ y_{(i)})x_{j(i)}$

For every j =0 …n

}

Where $x_{j(i)}$ Represents the j^{th} feature of the i^{th} training example. So if *m* is very large, then the derivative term fails to converge at the global minimum.

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

A. A **Q–Q plot** is a plot of the quantiles of two distributions against each other.It is used to compare the two distributions.

It is used to check following scenarios:

If two data sets —

i. come from populations with a common distribution

ii. have common location and scale

iii. have similar distributional shapes

iv. have similar behavior

interpretations for two data sets:

a) **Similar distribution**: If all points of quantiles lie on a straight line at an angle of 45 degree from x -axis.

b) **Y-values < X-values:** If y-quantiles are lower than the x-quantiles.

c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.

d) **Different distribution:** If all points of quantiles lie away from the straight line at an angle of 45 degree from x -axis.