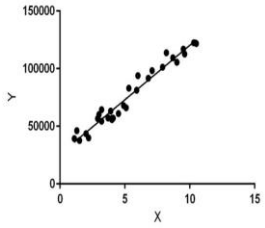
1. **Explain the linear regression algorithm in detail.**

Ans: Linear Regression is a machine learning algorithm based on supervised learning. 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.



Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y (output). Hence, the name is Linear Regression.
In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.

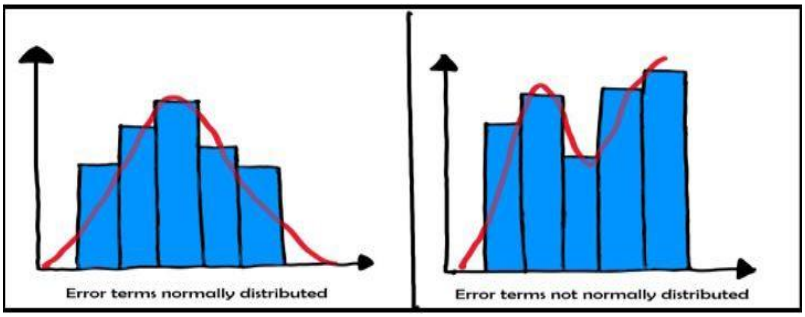
$$y = \theta_1 + \theta_2.x$$

While training the model we know:
x: input training data (univariate – one input variable (parameter))
y: labels to data (supervised learning)
When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best θ_1 and θ_2 values.
θ_1: intercept **θ_2:** coefficient of x
Once we find the best θ_1 and θ_2 values, we get the best fit line. So when we are finally using our model for prediction, it will predict the value of y for the input value of x.

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

Ans: a) **Normality assumption**: It is assumed that the error terms, ε (i), are normally distributed. If the residuals are not normally distributed, their randomness is lost, which

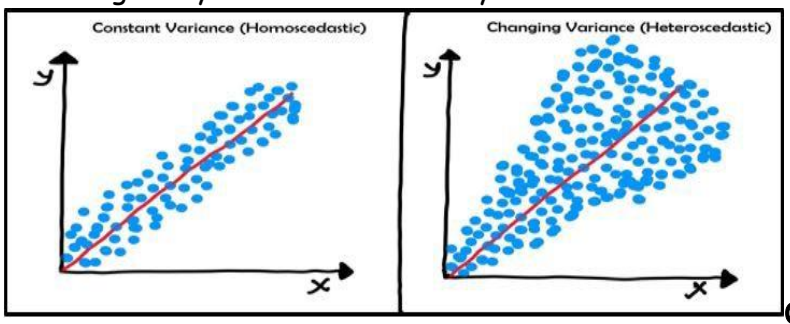
implies that the model is not able to explain the relation in the data.



Normality Assumption

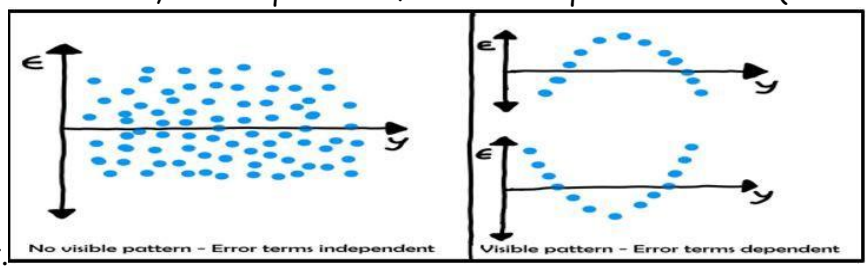
b) **Zero mean assumption**: It is assumed that the residuals have a mean value of zero, i.e., the error terms are normally distributed around zero.

c) **Constant variance assumption**: It is assumed that the residual terms have the same (but unknown) variance, σ2. This assumption is also known as the assumption of homogeneity or homoscedasticity.



Constant Variance

d) **Independent error assumption**: It is assumed that the residual terms are independent of each other, i.e. their pair-wise co-variance is zero. This means that there is no correlation between the residuals and the predicted values, or among the residuals themselves. If some correlation is present, it implies that there is some relation that the regression model is not able to identify. If the independent variables are not linearly independent of each other, the uniqueness of the least square's solution (or normal equation solution) is lost.



Independent Error

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

Ans: The correlation coefficient is a statistical measure of the strength of the relationship between the relative movements of two variables. The values range between -1.0 and 1.0. A correlation of -1.0 shows a perfect negative correlation, while a correlation of 1.0 shows a perfect positive correlation. A correlation of 0.0 shows no linear relationship between the movements of the two variables.

For example, a correlation coefficient could be calculated to determine the level of correlation between the price of crude oil and the stock price of an oil-producing company, such as Exxon Mobil Corporation. Since oil companies earn greater profits as oil prices rise, the correlation between the two variables is highly positive.

The coefficient of determination (denoted by R^2) is interpreted as the proportion of the variance in the dependent variable that is predictable from the independent variable. The coefficient of determination is the square of the correlation (r) between predicted y scores and actual y scores; thus, it ranges from 0 to 1. An R^2 of 0 means that the dependent variable cannot be predicted from the independent variable. An R^2 of 1 means the dependent variable can be predicted without error from the independent variable. An R^2 between 0 and 1 indicates the extent to which the dependent variable is predictable. For e.g. An R^2 of 0.20 means that 20 percent is predictable.

The coefficient of determination for a linear regression model with one independent variable is: $$R^2 = \{(1/N)^* \Sigma [(x_i - x)^*(y_i - y)]/ (\sigma x * \sigma_y)\}^2$$

Where N is the number of observations used to fit the model, Σ is the summation symbol x_i is the x value for observation I, x is the mean x value, y_i is the y value for observation i, y is the mean y value, σ_x is the standard deviation of x and σ_y is the standard deviation of y.

4. Explain the Anscombe's quartet in detail.

Ans: 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, and we must emphasize completely, when they are graphed. Each graph tells a different

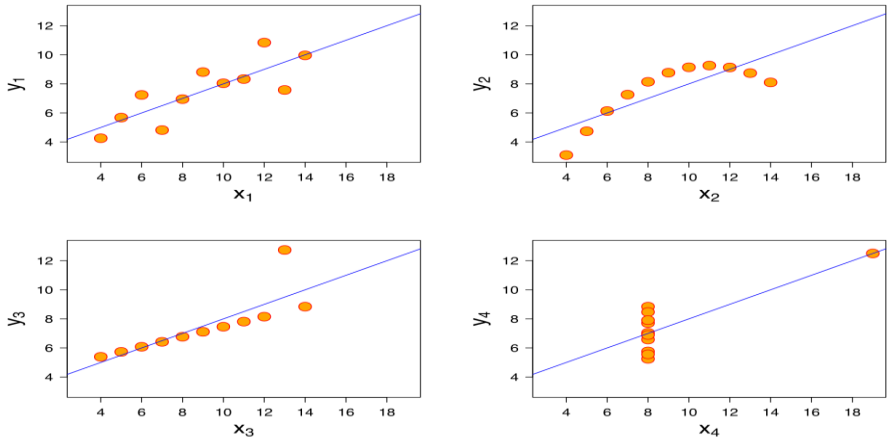
story irrespective of their similar summary statistics.

	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: Mean of x is 9 and mean of y is 7.50 for each dataset. Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset. The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset.

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.

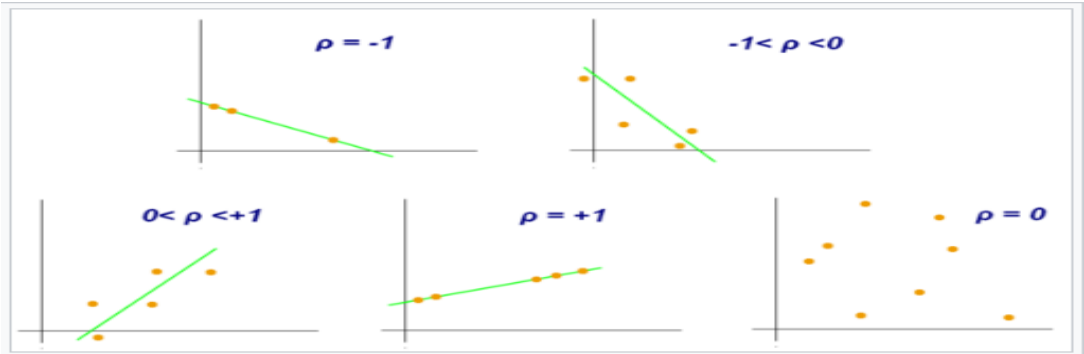


This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

5. What is Pearson's R?

Ans: In statistics, the Pearson correlation coefficient, also referred to as Pearson's r,or the bivariate correlation, is a statistic that measures linear correlation between two

variables X and Y. It has a value between +1 and −1. A value of +1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation.



Examples of scatter diagrams with different values of correlation coefficient (ρ)

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

Ans: Scaling is the method used to standardize or normalize the features of data in a fixed range. Since, the range of values of data may vary widely, it becomes a necessary step in data preprocessing while using machine learning algorithms.

If scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

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

Here, X_{max} and X_{min} are the maximum and the minimum values of the feature respectively. When the value of X is the minimum value in the column, the numerator will be 0, and hence X' is 0. On the other hand, when the value of X is the maximum value in the column, the numerator is equal to the denominator and thus the value of X' is 1. If the value of X is between the minimum and the maximum value, then the value of X' is between 0 and 1.

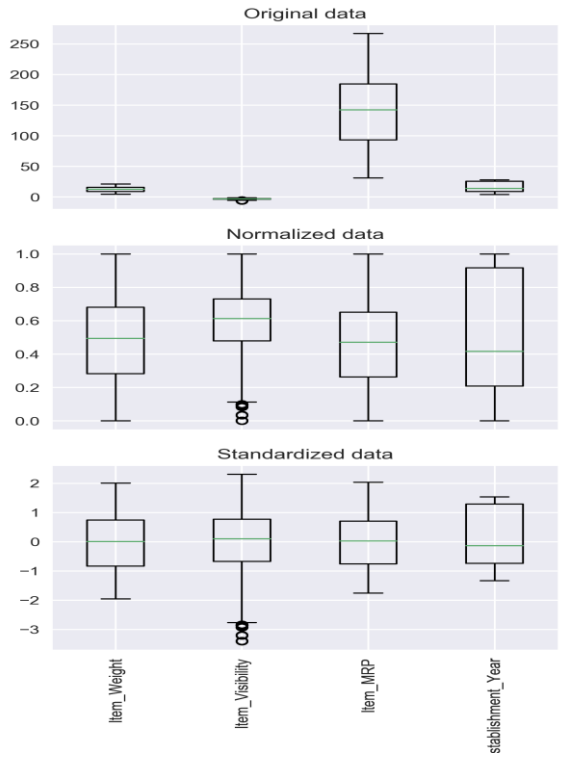
Normalization is good to use when we know that the distribution of our data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbours and Neural Networks

Standardization is another scaling technique where the values are centered on the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

$$X' = \frac{X - \mu}{\sigma}$$

μ is the mean of the feature values and σ is the standard deviation of the feature values. Note that in this case, the values are not restricted to a particular range..

Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if we have outliers in our data, they will not be affected by standardization.



The comparison between an unscaled and scaled data using boxplots.

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

<u>Ans:</u> An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

8. What is the Gauss-Markov theorem?

Ans: Gauss Markov theorem tells us that if a certain set of assumptions are met, the ordinary least squares estimate for regression coefficients gives us the best linear unbiased estimate possible.

Gauss Markov Assumptions:

There are five Gauss Markov assumptions (also called conditions):

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.

The Gauss Markov assumptions guarantee the validity of ordinary least squares for estimating regression coefficients.

In practice, the Gauss Markov assumptions are rarely all met perfectly, but they are still useful as a benchmark, and because they show us what 'ideal' conditions would be. They also allow us to pinpoint problem areas that might cause our estimated regression coefficients to be inaccurate or even unusable.

$y_i = x_i' \beta + \varepsilon_i$

and generated by the ordinary least squares estimate is the best linear unbiased estimate possible if

- $E\{\varepsilon_i\} = 0, i = 1, \ldots, N$
- $\{\varepsilon_1 \ldots\ldots \varepsilon_n\}$ and $\{x_1 \ldots\ldots, x_N\}$ are independent
- $cov\{\varepsilon_i, \varepsilon_j\} = 0, i, j = 1 \ldots, N\ I \neq j.$
- $V\{\varepsilon_1 = \sigma^2, i = 1, \ldots N$

The first of these assumptions can be read as "The expected value of the error term is zero." The second assumption is collinearity, the third is exogeneity, and the fourth is homoscedasticity.

9. Explain the gradient descent algorithm in detail.

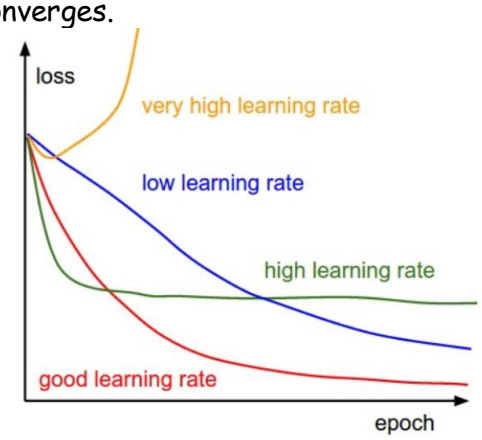
Ans: Gradient Descent is the most common optimization algorithm in machine learning and deep learning. It is a first-order optimization algorithm. This means it only takes into account the first derivative when performing the updates on the parameters. On each iteration, we update the parameters in the opposite direction of the gradient of the objective function J (w) w.r.t the parameters where the gradient gives the direction of the steepest ascent. The size of the step we take on each iteration to reach the local minimum

is determined by the learning rate a. Therefore, we follow the direction of the slope downhill until we reach a local minimum.

For the sake of simplicity, let's assume that the logistic regression model has only two parameters: weight w and bias b.
1. Initialize weight w and bias b to any random numbers.
2. Pick a value for the learning rate a. The learning rate determines how big the step would be on each iteration.

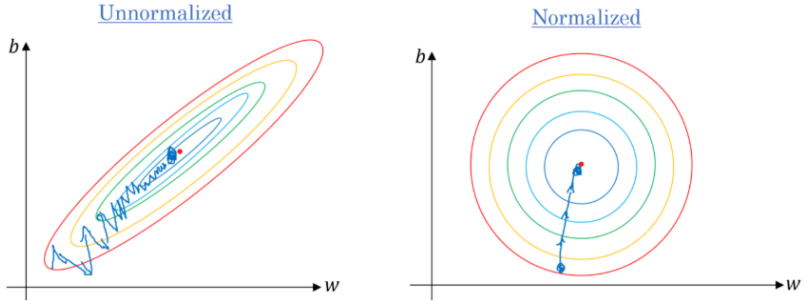
- If a is very small, it would take long time to converge and become computationally expensive.
- If a is large, it may fail to converge and overshoot the minimum.

Therefore, plot the cost function against different values of a and pick the value of a that is right before the first value that didn't converge so that we would have a very fast learning algorithm that converges.



- The most commonly used rates are: 0.001, 0.003, 0.01, 0.03, 0.1, and 0.3.

3. Make sure to scale the data if it's on a very different scales. If we don't scale the data, the level curves (contours) would be narrower and taller which means it would take longer time to converge.



Gradient descent: normalized versus unnormalized level curves.

Scale the data to have μ = 0 and σ = 1. Below is the formula for scaling each example:

$$\frac{x_i - \mu}{\sigma}$$

4. On each iteration, take the partial derivative of the cost function J(w) w.r.t each parameter (gradient):

$$\frac{\partial}{\partial w} J(w) = \nabla_w J$$

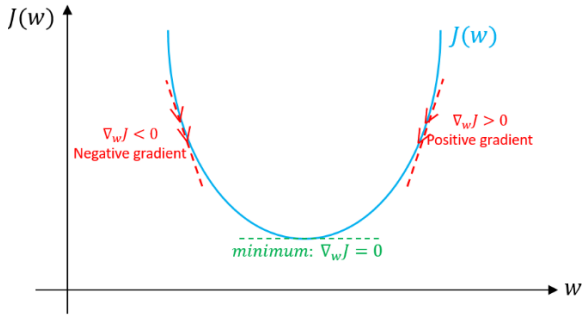
$$\frac{\partial}{\partial b} J(w) = \nabla_b J$$

The update equations are:

$$w = w - \alpha \nabla_w J$$

$$b = b - \alpha \nabla_b J$$

For the sake of illustration, let's assume we don't have bias. If the slope of the current value of w > 0, this means that we are to the right of optimal w*. Therefore, the update will be negative, and will start getting close to the optimal values of w*. However, if it's negative, the update will be positive and will increase the current values of w to converge to the optimal values of w*



Gradient descent. An illustration of how gradient descent algorithm uses the first derivative of the loss function to follow downhill it's minimum.

- Continue the process until the cost function converges. That is, until the error curve becomes flat and doesn't change.
- In addition, on each iteration, the step would be in the direction that gives the maximum change since its perpendicular to level curves at each step.

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

Ans: Quantile-Quantile (Q-Q) plot, is 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.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

Few advantages:

a) It can be used with sample sizes also

b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

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 tail behaviour.

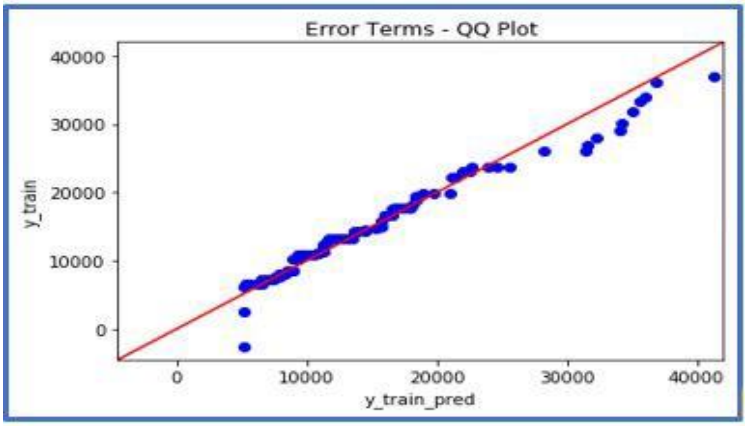
Interpretation:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

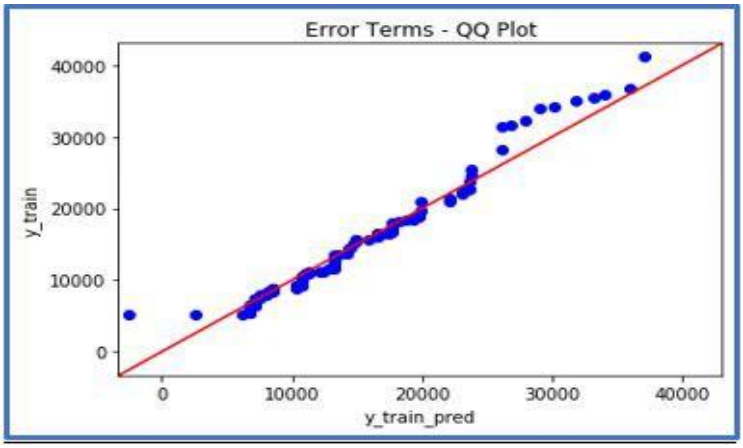
Below are the possible interpretations for two data sets.

a) **Similar distribution**: If all point of quantiles lies on or close to 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 point of quantiles lies away from the straight line at an angle of 45 degree from x –axis

statsmodels.api provide **qqplot** and **qqplot_2samples** to plot Q-Q graph for single and two different data sets respectively.