1. Explain the linear regression algorithm in detail.

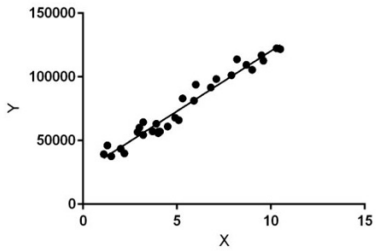
Linear Regression is a machine learning algorithm based on **supervised learning**.
This Regression models target prediction value based on independent variables, basically

- Find out the relationship between independent and dependent variable.
- Explain change in dependent variable with change in the value of Predictors.

There are two types of Linear Regression.

i. Simple Linear regression
ii. Multiple Linear Regression

It is mostly used for finding out the relationship between variables and forecasting
(Guarantee's interpolation of data not extrapolation)
It Shows correlation between variables not causation between variables.
It is a type of Parametric **Regression.** (No of dependent variable is fixed)



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.

Hypothesis function for Linear Regression :

$$Y = \beta 0 + \beta 1 X$$

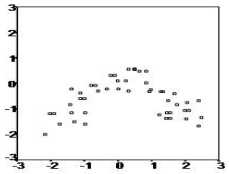
While training the model we are given :
x: input training data (univariate – one input variable(parameter))
y: labels to data (supervised learning)
When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best β1 and β0 values.
β0: intercept
β1: coefficient of x
Once we find the best β1 and β0 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?

Linear regression is an analysis that assesses whether one or more predictor variables explain the dependent (criterion) variable. The regression has five key assumptions:

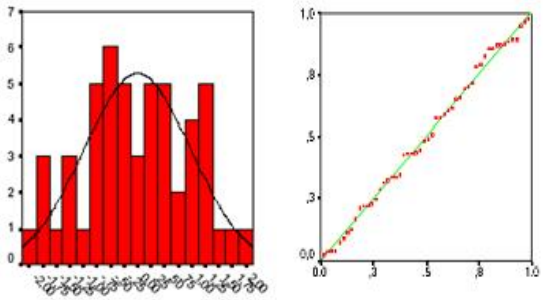
Linear relationship:

linear regression needs the relationship between the independent and dependent variables to be linear. It is also important to check for outliers since linear regression is sensitive to outlier effects. The linearity assumption can best be tested with scatter plots, the following two examples depict cases, where no and little linearity is present.



Multivariate normality :

linear regression analysis requires all variables to be multivariate normal. This assumption can best be checked with a histogram or a Q-Q-Plot. Normality can be checked with a goodness of fit test. When the data is not normally distributed a non-linear transformation might fix this issue.



No or little multicollinearity :

Linear regression assumes that there is little or no multicollinearity in the data. Multicollinearity occurs when the independent variables are too highly correlated with each other.

Multicollinearity may be tested with three central criteria:

1) **Correlation matrix** – when computing the matrix of Pearson's Bivariate Correlation among all independent variables the correlation coefficients need to be smaller than 1.

2) **Tolerance** – the tolerance measures the influence of one independent variable on all other independent variables; the tolerance is calculated with an initial linear regression analysis. Tolerance is defined as $T = 1 - R^2$ for these first step regression analysis. With $T < 0.1$ there might be multicollinearity in the data and with $T < 0.01$ there certainly is.

3) **Variance Inflation Factor (VIF)** – the variance inflation factor of the linear regression is defined as $VIF = 1/T$. With $VIF > 5$ there is an indication that multicollinearity may be present; with $VIF > 10$ there is certainly multicollinearity among the variables.

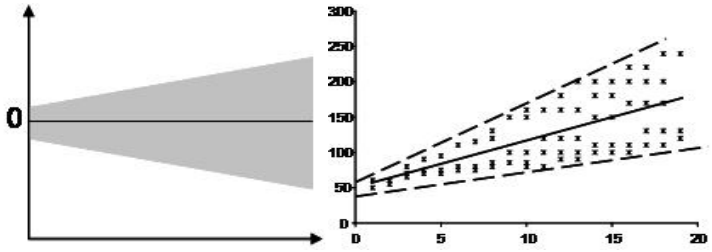
The simplest way to address the problem is to remove independent variables with high VIF values.

No auto-correlation:

Linear regression analysis requires that there is little or no autocorrelation in the data. Autocorrelation occurs when the residuals are not independent from each other. For instance, this typically occurs in stock prices, where the price is not independent from the previous price.

Homoscedasticity:

The scatter plot is good way to check whether the data are homoscedastic (meaning the residuals are equal across the regression line). The following scatter plots show examples of data that are not homoscedastic (i.e., heteroscedastic):



The Goldfeld-Quandt Test can also be used to test for heteroscedasticity. The test splits the data into two groups and tests to see if the variances of the residuals are similar across the groups. If homoscedasticity is present, a non-linear correction might fix the problem.

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

The correlation coefficient (r) is a statistical measure that calculates the strength of the relationship between the relative movements of two variables. The values range between -1.0 and 1.0. A calculated number greater than 1.0 or less than -1.0 means that there was an error in the correlation measurement. 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 relationship between the movement of the two variables.

The coefficient of determination (R^2 or r-squared) is a statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by the independent variable. In other words, the coefficient of determination tells one how well the data fits the model (the goodness of fit).

Mathematically, the coefficient of determination can be found using the following formula:

$$\text{Coefficient of Determination } (R^2) = 1 - \frac{SS_{\text{regression}}}{SS_{\text{total}}}$$

Where:

$SS_{regression}$ – the sum of squares due to regression (explained sum of squares)

SS_{total} – the total sum of squares.

The total sum of squares measures the variation in the observed data (data used in regression modelling). The sum of squares due to regression measures how well the regression model represents the data that were used for modelling.

4. Explain the Anscombe's quartet in detail.

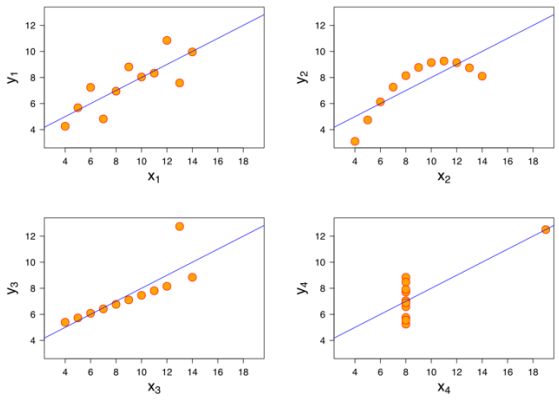
Anscombe's Quartet was developed by statistician Francis Anscombe.
It comprises four datasets, each containing eleven (x,y) pairs and each datasets share the same descriptive statistics.
But Each graph tells a different story irrespective of their similar summary statistics.

	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 denotes importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset. This type of scenario's called Anscombe's quartet

5. What is Pearson's R?

Pearson's correlation coefficient is the test statistics that measures the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship.

Assumptions:

- **Independent of case:** Cases should be independent to each other.
- **Linear relationship:** Two variables should be linearly related to each other. This can be assessed with a scatterplot: plot the value of variables on a scatter diagram, and check if the plot yields a relatively straight line.
- **Homoscedasticity:** the residuals scatterplot should be roughly rectangular-shaped.

Properties:

- **Limit:** Coefficient values can range from +1 to -1, where +1 indicates a perfect positive relationship, -1 indicates a perfect negative relationship, and a 0 indicates no relationship exists..
- **Pure number:** It is independent of the unit of measurement. For example, if one variable's unit of measurement is in inches and the second variable is in quintals, even then, Pearson's correlation coefficient value does not change.
- **Symmetric:** Correlation of the coefficient between two variables is symmetric. This means between X and Y or Y and X, the coefficient value of will remain the same.

Degree of correlation:

- **Perfect:** If the value is near ± 1, then it said to be a perfect correlation: as one variable increases, the other variable tends to also increase (if positive) or decrease (if negative).
- **High degree:** If the coefficient value lies between ± 0.50 and ± 1, then it is said to be a strong correlation.
- **Moderate degree:** If the value lies between ± 0.30 and ± 0.49, then it is said to be a medium correlation.
- **Low degree:** When the value lies below ± .29, then it is said to be a small correlation.
- **No correlation:** When the value is zero.

Formula to calculate Pearson's R:

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

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

It is a step of Data Pre Processing which is applied to independent variables or features of data. It basically helps to normalise the data within a particular range. Sometimes, it also helps in speeding up the calculations in an algorithm.

Real world dataset contains features that highly vary in magnitudes, units, and range. Normalisation should be performed when the scale of a feature is irrelevant or misleading and not should Normalise when the scale is meaningful.

The algorithms which use Euclidean Distance measure are sensitive to Magnitudes. Here feature scaling helps to weigh all the features equally. Formally, If a feature in the dataset is big in scale compared to others then in algorithms where Euclidean distance is measured this big scaled feature becomes dominating and needs to be normalized.

Min-Max Normalization: This technique re-scales a feature or observation value with distribution value between 0 and 1.

$$X_{new} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

Standardization: It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.

$$X_{new} = \frac{X_i - X_{mean}}{\text{Standard Deviation}}$$

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:

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

R^2 is a statistic that will give some information about the goodness of fit of a model. In regression, the R^2 coefficient of determination is a statistical measure of how well the regression predictions approximate the real data points. An R^2 of 1 indicates that the regression predictions perfectly fit the data.

If R^2 value 1 then VIF value will come as infinity as per the above mathematical equation.

8. What is the Gauss-Markov theorem?

The Gauss-Markov theorem states that if we 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.

The Gauss-Markov Theorem: OLS is BLUE!

The Gauss-Markov theorem famously states that OLS is BLUE. BLUE is an acronym for the following**: Best Linear Unbiased Estimator**

The definition of "best" refers to the minimum variance or the narrowest sampling distribution. More specifically, when we model satisfies the assumptions, OLS coefficient estimates follow the tightest possible sampling distribution of unbiased estimates compared to other linear estimation methods.

Gauss-Markov Theorem OLS Estimates and Sampling Distributions

The best estimates are those that are unbiased and have the minimum variance. When a model satisfies the assumptions, the Gauss-Markov theorem states that the OLS procedure produces unbiased estimates that have the minimum variance. The sampling distributions are centered on the actual population value and are the tightest possible distributions.

9. Explain the gradient descent algorithm in detail.

Gradient Descent is the most common first-order optimization algorithm in *machine learning* and *deep learning*. This means it only takes into account the first derivative when performing the updates on the parameters.

In linear regression, the model targets to get the best-fit regression line to predict the value of y based on the given input value (x). While training the model, the model calculates the cost function which measures the Root Mean Squared error between the predicted value (pred) and true value (y). The model targets to minimize the cost function.

To minimize the cost function, the model needs to have the best value of θ_1 and θ_2. Initially model selects θ_1 and θ_2 values randomly and then iteratively update these value in order to minimize the cost function until it reaches the minimum. By the time model achieves the minimum cost function, it will have the best θ_1 and θ_2 values. Using these finally updated values of θ_1 and θ_2 in the hypothesis equation of linear equation, model predicts the value of x in the best manner it can.

Below is the calculation how θ_1 and θ_2 values get updated in optimise manner :

Linear Regression Cost Function:

$$J = \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

$$minimize\frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

Gradient Descent Algorithm For Linear Regression:

Cost Function

$$J(\Theta_0, \Theta_1) = \frac{1}{2m}\sum_{i=1}^{m}[h_\Theta(x_i) - y_i]^2$$

Predicted Value; True Value

Gradient Descent

$$\Theta_j = \Theta_j - \alpha\frac{\partial}{\partial\Theta_j}J(\Theta_0, \Theta_1)$$

Learning Rate

Now,

$$\frac{\partial}{\partial\Theta}J_\Theta = \frac{\partial}{\partial\Theta}\frac{1}{2m}\sum_{i=1}^{m}[h_\Theta(x_i) - y]^2$$

$$= \frac{1}{m}\sum_{i=1}^{m}(h_\Theta(x_i) - y)\frac{\partial}{\partial\Theta_j}(\Theta x_i - y)$$

$$= \frac{1}{m}(h_\Theta(x_i) - y)x_i$$

Therefore,

$$\Theta_j := \Theta_j - \frac{\alpha}{m}\sum_{i=1}^{m}[(h_\Theta(x_i) - y)x_i]$$

$$\theta_j = \theta_j - \alpha\frac{\partial}{\partial\theta_j}J(\theta)$$

Now,

$$\frac{\partial}{\partial\theta}J_\theta = \frac{\partial}{\partial\theta}\frac{1}{2m}\sum_{i=1}^{m}\left[h_\theta(x_i) - y_i\right]^2$$

$$\frac{\partial}{\partial\theta}J_\theta = \frac{1}{m}\sum_{i=1}^{m}\left(h_\theta(x_i) - y_i\right)\cdot\frac{\partial}{\partial\theta_j}\left(\theta x_i - y_i\right)$$

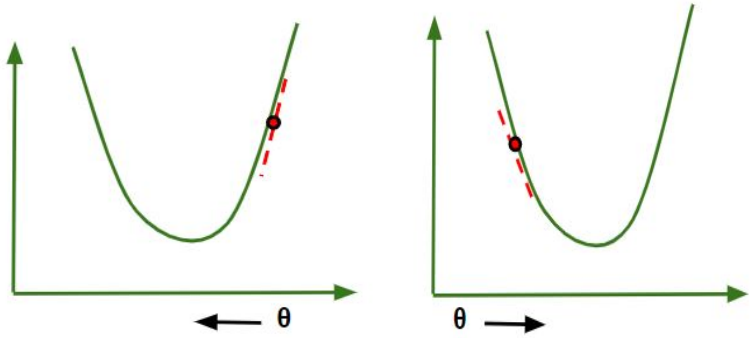
$$\frac{\partial}{\partial\theta}J_\theta = \frac{1}{m}\sum_{i=1}^{m}\left[\left(h_\theta(x_i) - y\right)x_i\right]$$

Therefore,

$$\theta_j := \theta_j - \frac{\alpha}{m}\sum_{i=1}^{m}\left[\left(h_\theta(x_i) - y_i\right)x_i\right]$$

-> θ_j : Weights of the hypothesis.
-> $h_\theta(x_i)$: predicted y value for i^{th} input.
-> j : Feature index number (can be 0, 1, 2, ..., n).
-> α : Learning Rate of Gradient Descent.

We graph cost function as a function of parameter estimates i.e. parameter range of our hypothesis function and the cost resulting from selecting a particular set of parameters. Gradient Descent step downs the cost function in the direction of the steepest descent. Size of each step is determined by parameter **α** known as **Learning Rate**.
In the Gradient Descent algorithm, we can infer two points :

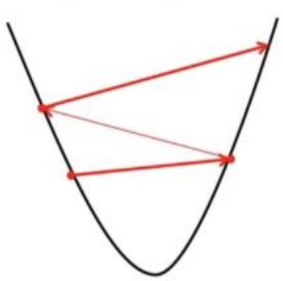


If slope is +ve : $\theta_j = \theta_j$ – (+ve value). Hence value of θ_j decreases.

If slope is -ve : $\theta_j = \theta_j$ – (-ve value). Hence value of θ_j increases.

The choice of correct learning rate is very important as it ensures that Gradient Descent converges in a reasonable time. :

1.If we choose **α to be very large**, Gradient Descent can overshoot the minimum. It may fail to converge or even diverge.



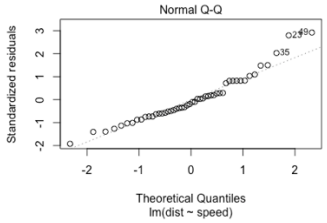
2.If we choose α to be very small, Gradient Descent will take small steps to reach local minima and will take a longer time to reach minima.

For linear regression Cost Function graph is always convex shaped.

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

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. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.
A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

QQ-plots are ubiquitous in statistics. It fit a linear regression model, check if the points lie approximately on the line, and if they don't, your residuals aren't Gaussian and thus your errors aren't either. This implies that for small sample sizes, we can't assume our estimator $\hat{\beta}$ is Gaussian either, so the standard confidence intervals and significance tests are invalid.



The points approximately fall on the line.On the x-axis are the theoretical quantiles of a standard normal. That is, we sort the n points, and then for each I , using the standard normal quantile function we find the x so that $P_{\text{std norm}}(X \leq x) = \frac{i-0.5}{n}$. For this dataset, for the case of the leftmost point, we have that i=1 and n=50.