1. Explain the linear regression algorithm in detail.

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.

Linear regression is one of the most commonly used ML algorithms for predictive analysis.

The overall idea of regression is to examine two things:

- Does a set of predictor variables do a good job in predicting an outcome (dependent) variable?
- Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable?

These regression estimates are in turn used to explain the relationship between one dependent variable with one or more independent variables.

The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = b0 + b1*x, where y = estimated dependent variable score, b0 = constant, b1 = regression coefficient, and x = score on the independent variable.

Three major uses for regression analysis are :

- determining the strength of predictors
- forecasting an effect, and
- trend forecasting.

First, the regression might be used to identify the strength of the effect that the independent variable(s) have on a dependent variable. Typical questions are what is the strength of relationship between dose and effect, sales and marketing spending, or age and income.

Second, it can be used to forecast effects or impact of changes. That is, the regression analysis helps us to understand how much the dependent variable changes with a change in one or more independent variables. A typical question is, “how much additional sales income do I get for each additional $1000 spent on marketing?”

Third, regression analysis predicts trends and future values. The regression analysis can be used to get point estimates. A typical question is, “what will the price of gold be in 6 months?”

There are several types of linear regression analyses available to researchers.

Simple linear regression
1 dependent variable (interval or ratio), 1 independent variable (interval or ratio or dichotomous)

Multiple linear regression
1 dependent variable (interval or ratio) , 2+ independent variables (interval or ratio or dichotomous)

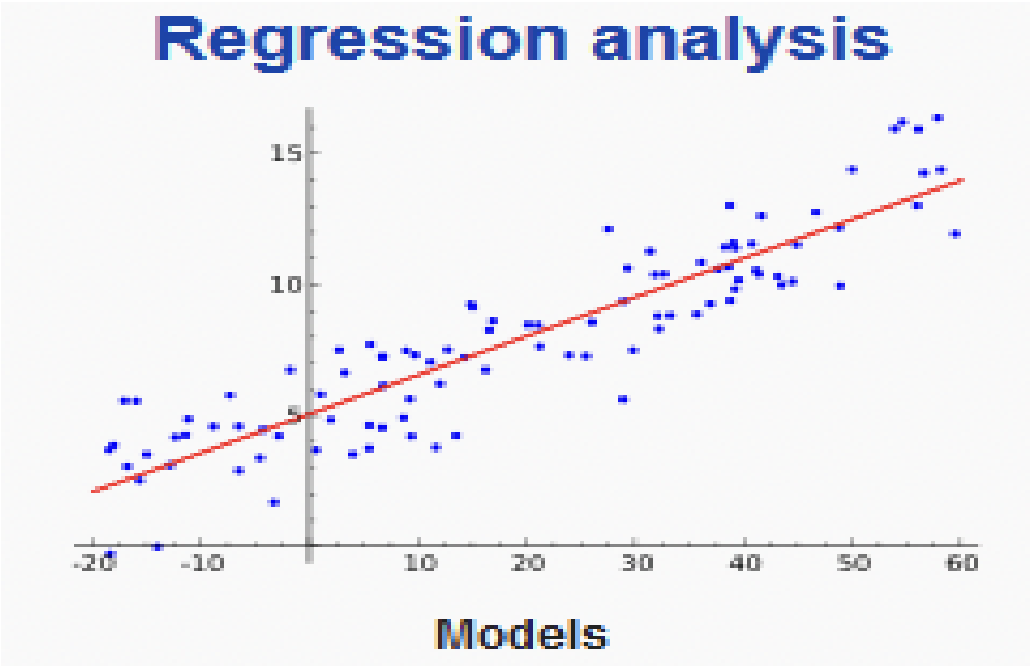
Logistic regression
1 dependent variable (dichotomous), 2+ independent variable(s) (interval or ratio or dichotomous)

Ordinal regression
1 dependent variable (ordinal), 1+ independent variable(s) (nominal or dichotomous)

Multinomial regression
1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio or dichotomous)

Discriminant analysis
1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio)

Linear regression models are often fitted using the least squares approach, but they may also be fitted in other ways, such as by minimizing the "lack of fit" in some other norm (as with least absolute deviations regression), or by minimizing a penalized version of the least squares cost function as in ridge regression (L2-norm penalty) and lasso (L1-norm penalty). Conversely, the least squares approach can be used to fit models that are not linear models. Thus, although the terms "least squares" and "linear model" are closely linked, they are not synonymous.



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 below key assumptions:

- There should be a linear and additive relationship between dependent (response) variable and independent (predictor) variable(s). A linear relationship suggests that a change in response Y due to one-unit change in X1 is constant, regardless of the value of X1. An additive relationship suggests that the effect of X1 on Y is independent of other variables.
- There should be no correlation between the residual (error) terms. Absence of this phenomenon is known as Autocorrelation.
- The independent variables should not be correlated. Absence of this phenomenon is known as multicollinearity.
- The error terms must have constant variance. This phenomenon is known as homoskedasticity. The presence of non-constant variance is referred to heteroskedasticity.
- The error terms must be normally distributed.

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

R square or coefficient of determination : shows percentage variation in y which is explained by all the x variables together. Higher the better. It is always between 0 and 1. It can never be negative – since it is a squared value.

It is easy to explain the R square in terms of regression. It is not so easy to explain the R in terms of regression.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.850[a]	.723	.690	4.57996

a. Predictors: (Constant), weight, horsepower

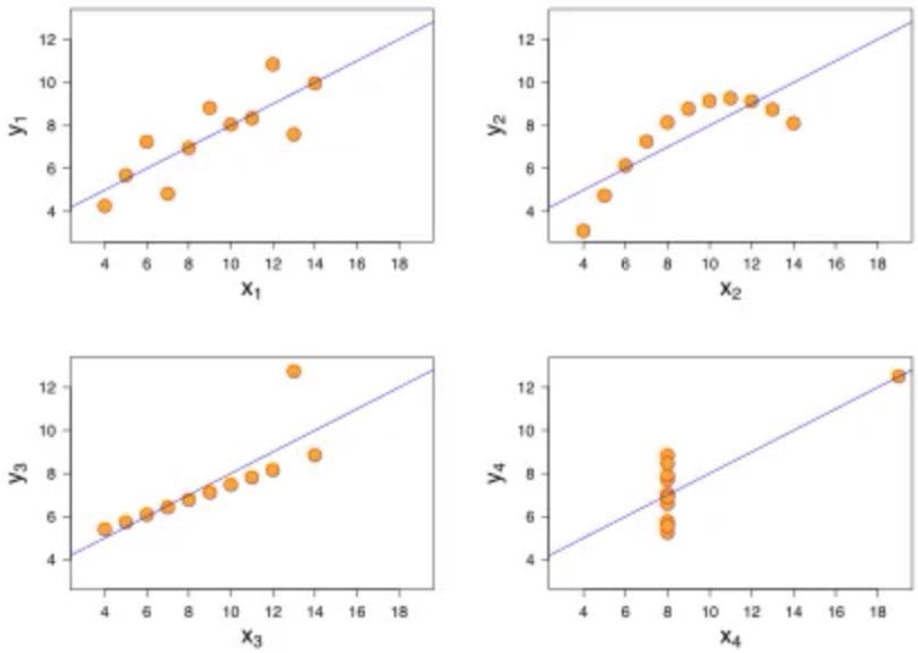
b. Dependent Variable: mpg

Coefficient of Correlation is the R value i.e. .850 (or 85%). Coefficient of Determination is the R square value i.e. .723 (or 72.3%). R square is simply square of R i.e. R times R.

Coefficient of Correlation: is the degree of relationship between two variables say x and y. It can go between -1 and 1. 1 indicates that the two variables are moving in unison. They rise and fall together and have perfect correlation. -1 means that the two variables are in perfect opposites. One goes up and other goes down, in perfect negative way. Any two variables in this universe can be argued to have a correlation value. If they are not correlated then the correlation value can still be computed which would be 0. The correlation value always lies between -1 and 1 (going thru 0 – which means no correlation at all – perfectly not related). Correlation can be rightfully explained for simple linear regression – because you only have one x and one y variable. For multiple linear regression R is computed, but then it is difficult to explain because we have multiple variables involved here. That's why R square is a better term. You can explain R square for both simple linear regressions and also for multiple linear regressions.

4. Explain the Anscombe's quartet in detail.

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (*x*, *y*) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analysing it and the effect of outliers on statistical properties.



The first scatter plot (top left) appears to be a simple linear regression, corresponding to two variables correlated and following the assumption of normality. The second graph (top right) is not distributed normally; while an obvious relationship between the two variables can be observed, it is not linear, and the Pearson's Correlation Coefficient is not relevant. In the third graph (bottom left), the distribution is linear, but with a different regression line, which is offset by the one outlier, which exerts enough influence to alter the regression line and lower the correlation coefficient from 1 to 0.816. Finally, the

fourth graph (bottom right) shows another example when one outlier is enough to produce a high correlation coefficient, even though the relationship between the two variables is not linear.

For all four datasets:

Property	Value
Mean of x in each case:	9 (exact)
Variance of x in each case:	11 (exact)
Mean of y in each case:	7.50 (to 2 decimal places)
Variance of y in each case:	4.122 or 4.127 (to 3 decimal places)
Correlation between x and y in each case:	0.816 (to 3 decimal places)
Linear regression line in each case:	$y = 3.00 + 0.500x$ (to 2 and 3 decimal places, respectively)

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyse 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. What is Pearson's R?

It 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.

In a sample it is denoted by r and is by design constrained as follows:

Furthermore,
1. Positive values denote positive linear correlation.
2. Negative values denote negative linear correlation.
3. 0 denotes no linear correlation.
4. The closer the value is to 1 or -1, the stronger the linear correlation. It Is given by,

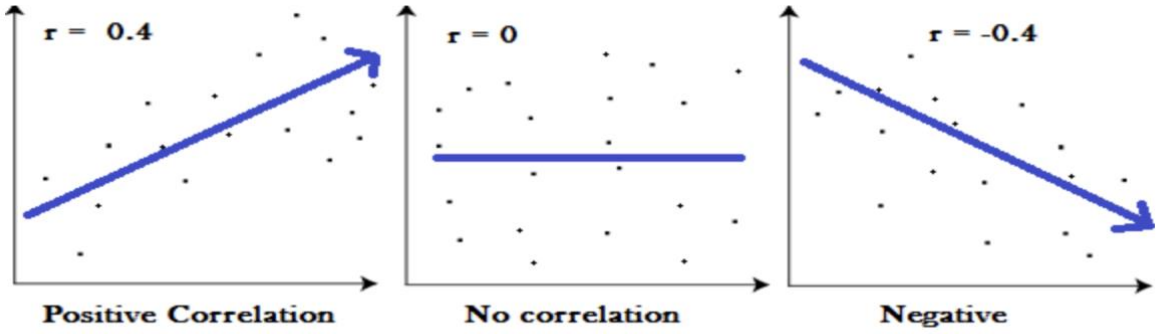
There are several types of correlation coefficient formulas.

One of the most commonly used formulas in stats is Pearson's correlation coefficient formula. If you're taking a basic stats class, this is the one you'll probably use:

$$\rho_{X,Y} = \frac{\mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y]}{\sqrt{\mathrm{E}[X^2] - [\mathrm{E}[X]]^2}\,\sqrt{\mathrm{E}[Y^2] - [\mathrm{E}[Y]]^2}}.$$

Correlation coefficient formulas are used to find how strong a relationship is between data. The formulas return a value between -1 and 1, where:

- 1 indicates a strong positive relationship.
- -1 indicates a strong negative relationship.
- A result of zero indicates no relationship at all.



Graphs showing a correlation of -1, 0 and +1

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

Scaling (also known as **data normalization**) is the method used to standardize the range of features of data. Since, the range of values of data may vary widely, it becomes a necessary step in data pre-processing while using machine learning algorithms.

In scaling *(also called **min-max scaling**)*, you transform the data such that the features are within a specific range e.g. [0, 1].

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

where x' is the normalized value.

It is mainly performed because most of the times, your dataset will contain features highly varying in magnitudes, units and range. But since, most of the machine learning algorithms use Euclidian distance between two data points in their computations, this is a problem. If left alone, these algorithms only take in the magnitude of features neglecting the units. The results would vary greatly between different units, 5kg and 5000gms. The features with high magnitudes will weigh in a lot more in the distance calculations than features with low magnitudes. To supress this effect, we need to bring all features to the same level of magnitudes. This can be achieved by scaling.

Normalized Scaling:

Also known as min-max scaling or min-max normalization, is the simplest method and consists in rescaling the range of features to scale the range in [0, 1] or [−1, 1]. Selecting the target range depends on the nature of the data. The general formula for a min-max of [0, 1] is given as:

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

where x is an original value, x' is the normalized value. For example, suppose that we have the students' weight data, and the students' weights span [160 pounds, 200 pounds]. To rescale this data, we first subtract 160 from each student's weight and divide the result by 40 (the difference between the maximum and minimum weights).

To rescale a range between an arbitrary set of values [a, b], the formula becomes:

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

where a, b are the min-max values.

Mean normalization

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

where x is an original value, x' is the normalized value. There is another form of the mean normalization which is when we divide by the standard deviation which is also called standardization.

Standardized Scaling:

In machine learning, we can handle various types of data, e.g. audio signals and pixel values for image data, and this data can include multiple dimensions. Feature standardization makes the values of each feature in the data have zero-mean (when subtracting the mean in the numerator) and unit-variance. This method is widely used for normalization in many machine learning algorithms (e.g., support vector machines, logistic regression, and artificial neural networks). The general method of calculation is to determine the distribution mean and standard deviation for each feature. Next we subtract the mean from each feature. Then we divide the values (mean is already subtracted) of each feature by its standard deviation.

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

Where x is the original feature vector, $\bar{x}$ = average (x) is the mean of that feature vector, and σ is its standard deviation. Scaling to unit length

Another option that is widely used in machine-learning is to scale the components of a feature vector such that the complete vector has length one. This usually means dividing each component by the Euclidean length of the vector:

$$x' = \frac{x}{\|x\|}$$

In some applications (e.g. Histogram features) it can be more practical to use the L1 norm (i.e. Manhattan Distance, City-Block Length or Taxicab Geometry) of the feature vector. This is especially important if in the following learning steps the Scalar Metric is used as a distance measure.

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

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:

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

When R-squared reaches 1, VIF reaches infinity. When R-squared reaches 1 then it means multicollinearity exists. Different variables are highly correlated with each other.

8. What is the Gauss-Markov theorem?

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

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.

We can summarize the Gauss-Markov Assumptions succinctly in algebra, by saying that a linear regression model represented by

$y_i = x_i'\beta + \varepsilon_i$

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

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

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.

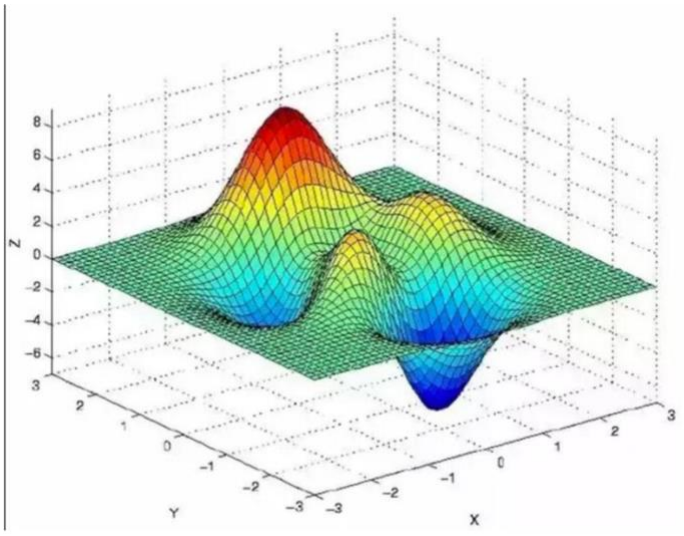
Gradient descent is a first-order iterative optimization algorithm for finding the minimum of a function. To find a local minimum of a function using gradient descent, one takes steps proportional to the negative of the gradient (or approximate gradient) of the function at the current point. If, instead, one takes steps proportional to the positive of the gradient, one approaches a local maximum of that function; the procedure is then known as gradient ascent.

Gradient descent is also known as steepest descent. However, gradient descent should not be confused with the method of steepest descent for approximating integrals.

Suppose you are at the top of a mountain, and you have to reach a lake which is at the lowest point of the mountain (aka valley). A twist is that you are blindfolded and you have zero visibility to see where you are headed. So, what approach will you take to reach the lake?

The best way is to check the ground near you and observe where the land tends to descend. This will give an idea in what direction you should take your first step. If you follow the descending path, it is very likely you would reach the lake.

To represent this graphically, notice the below graph.



Let us now map this scenario in mathematical terms.

Suppose we want to find out the best parameters (θ1) and (θ2) for our learning algorithm. Similar to the analogy above, we see we find similar mountains and valleys when we plot our “cost space”. Cost space is nothing but how our algorithm would perform when we choose a particular value for a parameter.

So on the y-axis, we have the cost J(θ) against our parameters θ1 and θ2 on x-axis and z-axis respectively. Here, hills are represented by red region, which have high cost, and valleys are represented by blue region, which have low cost.

Now there are many types of gradient descent algorithms. They can be classified by two methods mainly:

On the basis of data ingestion:

- Full Batch Gradient Descent Algorithm
- Stochastic Gradient Descent Algorithm

In full batch gradient descent algorithms, you use whole data at once to compute the gradient, whereas in stochastic you take a sample while computing the gradient.

On the basis of differentiation techniques

- First order Differentiation
- Second order Differentiation

Gradient descent requires calculation of gradient by differentiation of cost function. We can either use first order differentiation or second order differentiation.

Challenges in executing Gradient Descent :

Gradient Descent is a sound technique which works in most of the cases. But there are many cases where gradient descent does not work properly or fails to work altogether. There are three main reasons when this would happen:

Data Challenges

1. If the data is arranged in a way that it poses a non-convex optimization problem. It is very difficult to perform optimization using gradient descent. Gradient descent only works for problems which have a well-defined convex optimization problem.
2. Even when optimizing a convex optimization problem, there may be numerous minimal points. The lowest point is called global minimum, whereas rest of the points are called local minima. Our aim is to go to global minimum while avoiding local minima.
3. There is also a saddle point problem. This is a point in the data where the gradient is zero but is not an optimal point. We don't have a specific way to avoid this point and is still an active area of research.

Gradient Challenges

If the execution is not done properly while using gradient descent, it may lead to problems like vanishing gradient or exploding gradient problems. These problems occur when the gradient is too small or too large. And because of this problem the algorithms do not converge.

Implementation Challenges

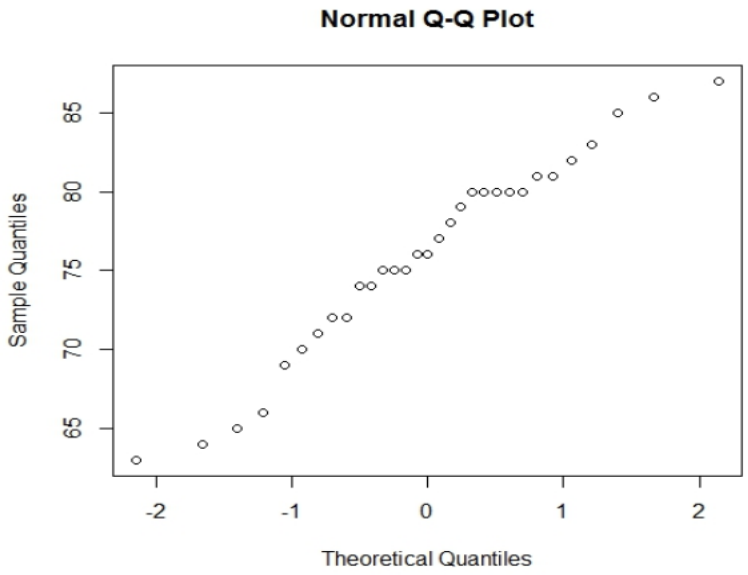
1. Most of the neural network practitioners don't generally pay attention to implementation, but it's very important to look at the resource utilization by networks. For example: When implementing gradient descent, it is very important to note how many resources you would require. If the memory is too small for your application, then the network would fail.
2. Also, it is important to keep track of things like floating point considerations and hardware / software prerequisites.

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

The Q-Q plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential. For example, if we run a statistical analysis that assumes our dependent variable is Normally distributed, we can use a Normal Q-Q plot to check that assumption. It's just a visual check, not an air-tight proof, so it is somewhat subjective. But it allows us to see at-a-glance if our assumption is plausible, and if not, how the assumption is violated and what data points contribute to the violation.

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. Here's an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.

This plot shows if residuals are normally distributed. Do residuals follow a straight line well or do they deviate severely? It's good if residuals are lined well on the straight dashed line.



Q-Q plots take your sample data, sort it in ascending order, and then plot them versus quantiles calculated from a theoretical distribution. The number of quantiles is selected to match the size of your sample data. While Normal Q-Q Plots are the ones most often used in practice due to so many statistical methods assuming normality, Q-Q Plots can actually be created for any distribution.

QQ-plots are ubiquitous in statistics. Most people use them in a single, simple way: 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, you can't assume your estimator is Gaussian either, so the standard confidence intervals and significance tests are invalid.

Submitted By: Chandan Kumar