Linear Regression Assignment Questions and Answers

Question 1: Explain the linear regression algorithm in detail.

Answer: Linear Regression is a machine learning algorithm which is based on **supervised learning** category. It finds a best linear-fit relationship on any given data, between independent (Target) and dependent (Predictor) variables. In other words, it creates the best straight-line fitting to the provided data to find the best linear relationship between the independent and dependent variables. Mostly it uses **Sum of Squared Residuals** Method.

Linear regression is of the 2 types:

i. **Simple Linear Regression:** It explains the relationship between a dependent variable and only one independent variable using a straight line. The straight line is plotted on the scatter plot of these two points.

Formula for the Simple Linear Regression:

$Y=\beta 0+\beta 1X1 +\epsilon$

ii. **Multiple Linear Regression:** It shows the relationship between one dependent variable and several independent variables. The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X. It fits a 'hyperplane' instead of a straight line.

Formula for the Multiple Linear Regression:

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_pX_p+\epsilon$$

The equation of the best fit regression line $Y = \beta_0 + \beta_1X$ can be found by the following two methods:

· Differentiation

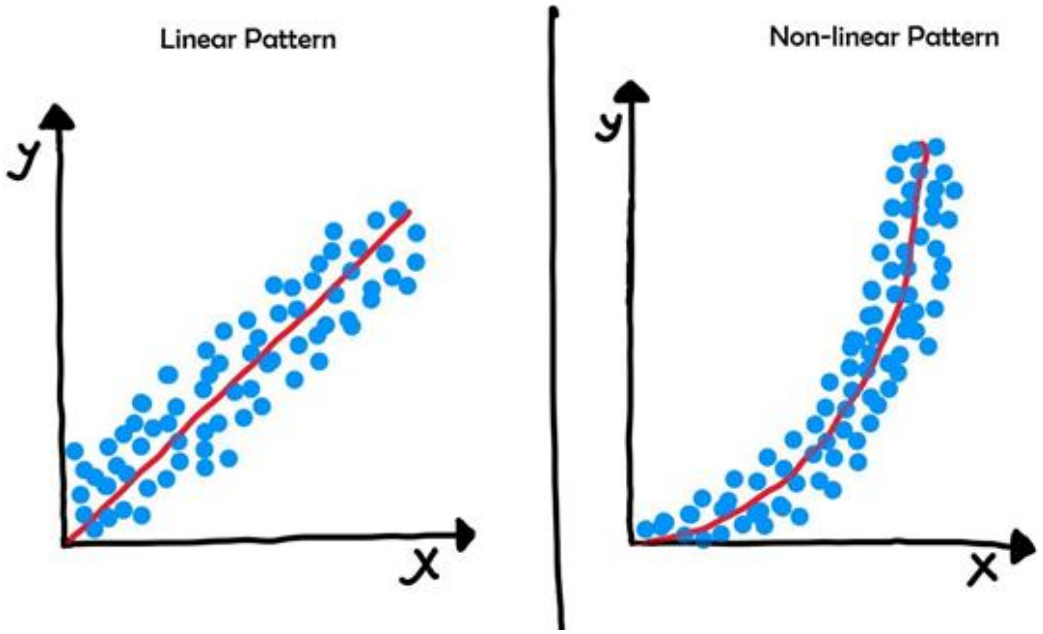
· Gradient descent

We can use statsmodels or SKLearn libraries in Python for the linear regression.

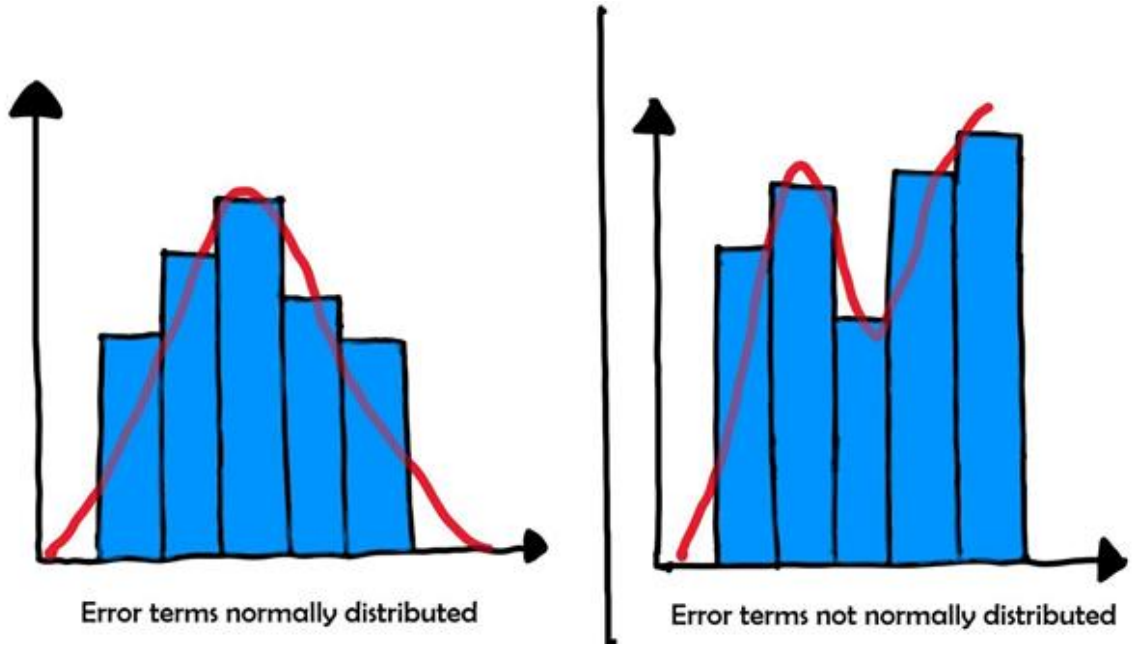
Question 2: What are the assumptions of linear regression regarding residuals?

Answer: At the time of building a linear model, we assume that the target variable and predictor variables are linearly dependent. But, apart from these, below are few assumptions in linear regression model:

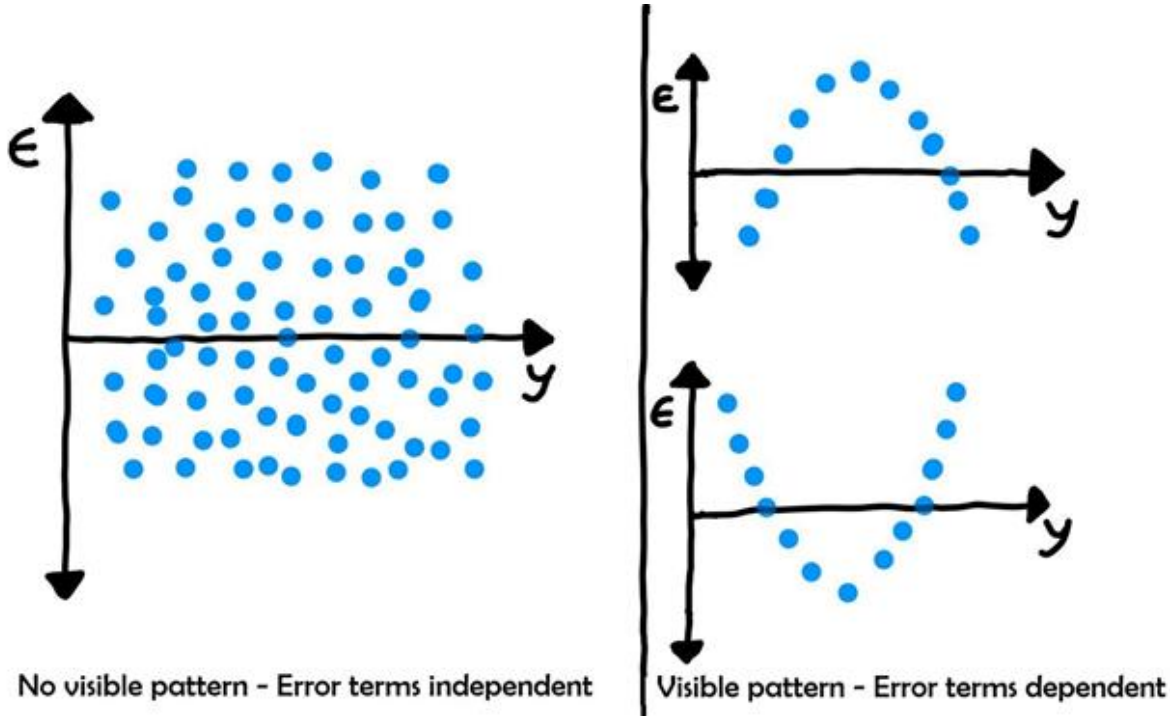
1. **Linear relationship between X and y:** X and Y should always display some sort of a linear relationship; otherwise, there will not be any use of fitting a linear model between them.



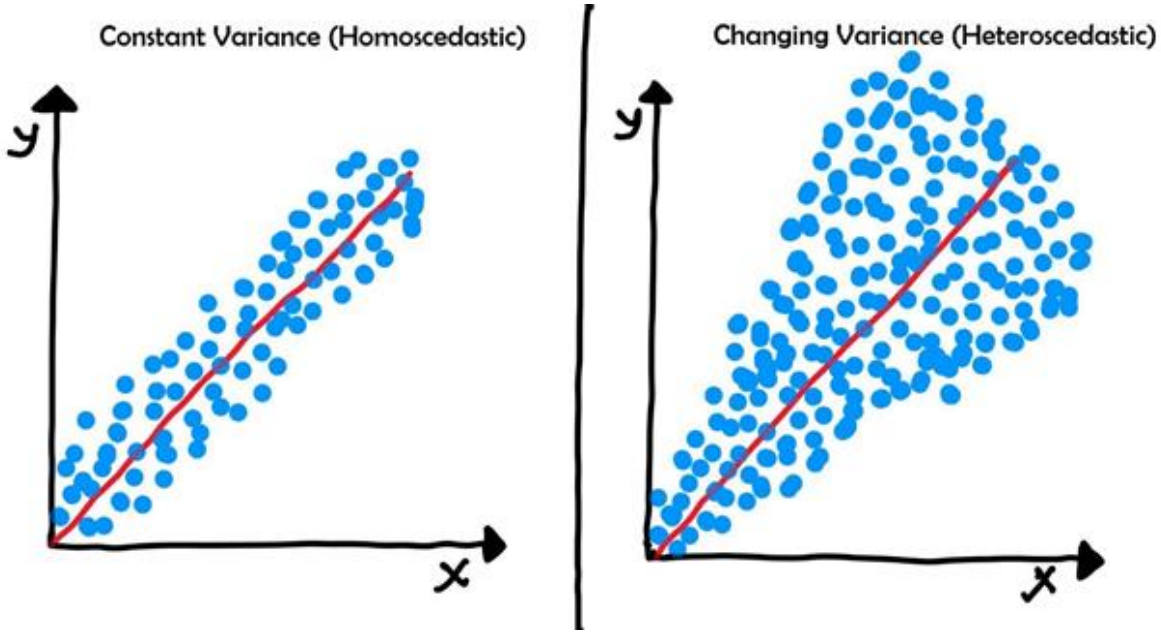
2. Normal distribution of error terms: It represents the assumption of normality. Which exhibits that error terms generally follow a **normal distribution with mean equal to zero** in most cases.



3. Independence of error terms: It explains that the error terms should not be dependent on one another. It means, there should not be any meaningful distribution between independent variable and error term.



4. Constant variance of error terms: This assumption says that the variance should not increase or decrease as the error values change. Also, the variance should not follow any pattern as the error terms change.



Question 3: What is the coefficient of correlation and the coefficient of determination?

Answer: The difference between coefficient of correlation and the coefficient of determination is explained as below:

Coefficient of Correlation: It is the strength of relationship between two variables say, x and y. It always falls in the range between -1 and 1 and is denoted by 'R'. It is of the 2 types:

· **Positive Correlation:** If one value changes, the other also changes in the same direction (increase/decrease).

· **Negative Correlation:** If one value changes, the other changes in the opposite direction.

Coefficient of Determination: It explains proportion of the variance in the dependent variable that is predictable from the

independent variable. It is the square of the correlation of correlation and is denoted by R2. It falls in the range between 0 and 1. It can never be negative since it is a squared value. It also denotes the goodness of best fit line and usually we get it in the summary table in the Regression output.

Question 4: Explain the Anscombe's quartet in detail.

Answer: Anscombe's Quartet was developed by statistician **Francis Anscombe**. This is a method which keeps four datasets, each containing eleven (x, y) pairs. The important thing to note about these datasets is that they share the same descriptive statistics. Each graph tells a different story irrespective of their similar summary statistics. Below is the glimpse of the statistics of the 4 datasets:

	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

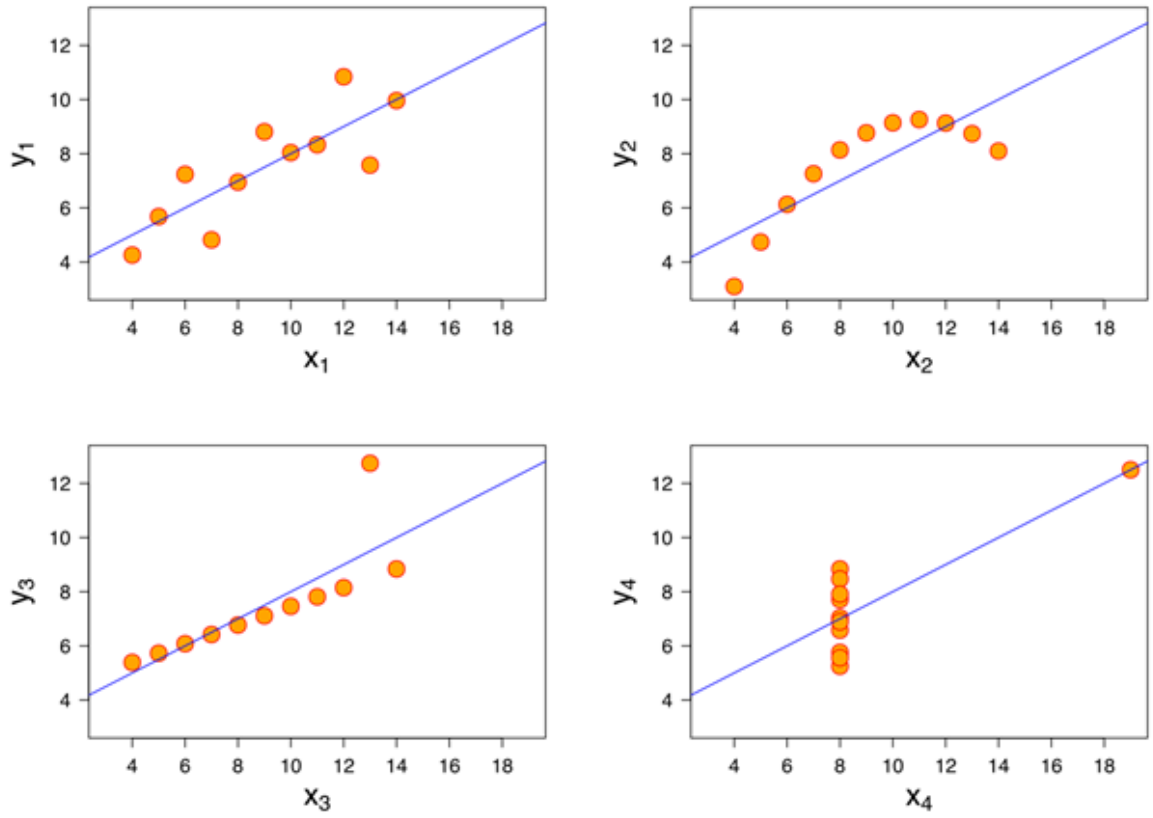
The summary statistics show that the means and the variances were identical for x and y across the groups:

· 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.

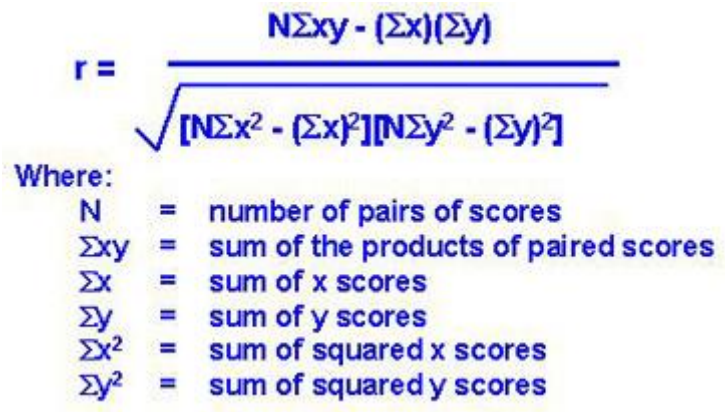
Additionally, Anscombe's Quartet warns of the dangers of outliers in data sets.

Question 5: What is Pearson's R?

Answer: Pearson's R was developed by **Karl Pearson** and it is a correlation coefficient which is a measure of the strength of a linear association between two variables and it is denoted by 'r'. it has a value between +1 and −1, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation.

Mathematically, Pearson's correlation coefficient is denoted as the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name.

Formula:



Example:

· Statistically significant relationship between age and height.

· Relationship between temperature and ice cream sales.

· Relationship among job satisfaction, productivity, and income.

· Which two variables have the strongest co-relation between age, height, weight, size of family and family income.

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

Answer: Scaling is the process to normalize the data within a particular range. Many times, in our dataset we see that multiple variables are in different ranges. So, scaling is required to bring them all in a single range.

The two most discussed scaling methods are **Normalization** and **Standardization**. Normalization typically scales the values into a range of [0,1]. Standardization typically scales data to have a mean of 0 and a standard deviation of 1 (unit variance).

Formula of Normalized scaling:

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

Formula of Standardized scaling:

$$x = \frac{x - mean(x)}{sd(x)}$$

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

Answer: The value of VIF is calculated by the below formula:

$$VIF_i = \frac{1}{1 - R_i^2}$$

Where, 'i' refers to the ith variable.

If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. It denotes perfect correlation in variables.

Question 8: What is the Gauss-Markov theorem?

Answer: This theorem was named after **Carl Friedrich Gauss** and **Andrey Markov.** In statistics, the Gauss–Markov theorem states that in a linear regression model in which the errors are uncorrelated, have equal variances and expectated value of zero, the best linear unbiased estimator (BLUE) of the coefficients is given by the ordinary least squares (OLS) estimator, provided it exists.

Here "best" refers to the minimum variance or the narrowest sampling distribution.

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

Gauss Markov Assumptions:

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

I. **Linearity**: the parameters we are estimating using the OLS method must be themselves linear.

II. **Random**: our data must have been randomly sampled from the population.

III. **Non-Collinearity**: the regressors being calculated aren't perfectly correlated with each other.

IV. **Exogeneity**: the regressors aren't correlated with the error term.

V. **Homoscedasticity**: no matter what the values of our regressors might be, the error of the variance is constant.

The notation for the model of a population is the following:
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$

Question 9: Explain the gradient descent algorithm in detail.

Answer: Gradient descent is an optimization algorithm used to find the values of the parameters (coefficients) of a function (f) that minimizes a given cost function (cost).

The equation for the line that's fit the data, is given as:

$y(p)=\beta 0+\beta 1x$

Where β0 is the intercept of the fitted line and β1 is the coefficient for the independent variable x.

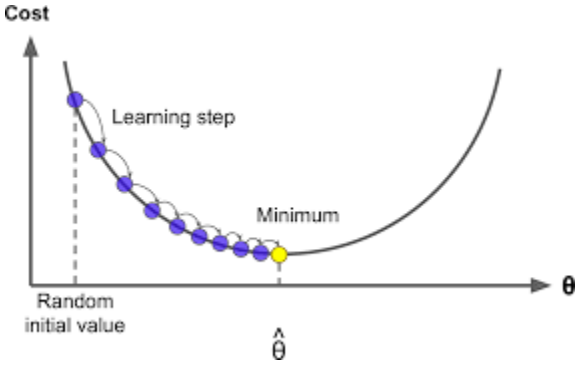
The main challenge is how to find β0 and β1. To find the optimum betas, we need to reduce the cost function for all data points, which is given as:

$J(\theta 0,\theta 1)=\sum Ni=1(yi-yi(p))2$

So, Gradient descent is an iterative form solution of order one. So to compute optimal thetas, we need to apply Gradient Descent to the Cost function, which is given as follows:

$$\frac{\partial}{\partial \theta} J(\theta)$$

So, starting at the top of the mountain, we take our first step downhill in the direction specified by the negative gradient. Next, we recalculate the negative gradient (passing in the coordinates of our new point) 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.

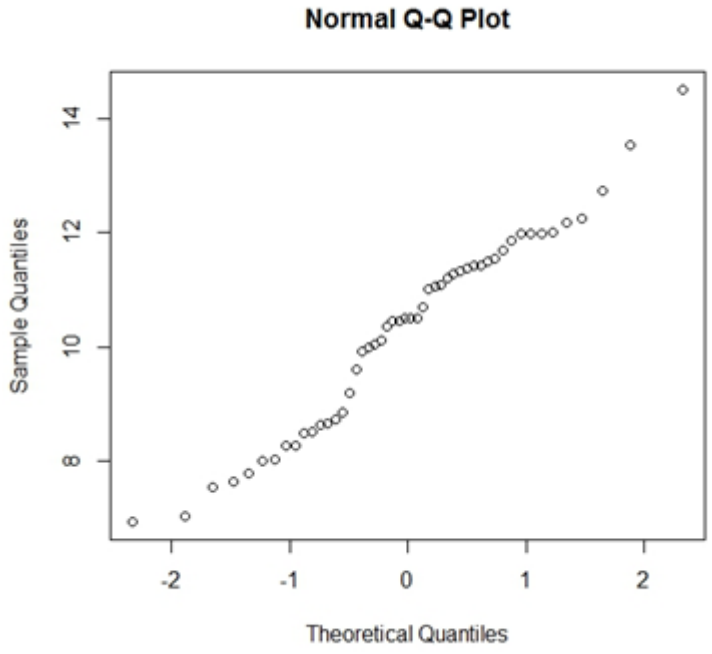


Gradient descent depends on the Learning Rate. With a high learning rate, we can cover more ground each step, but we risk overshooting the lowest point since the slope of the hill is constantly changing. With a very low learning rate, we can confidently move in the direction of the negative gradient since we are recalculating it so frequently. A low learning rate is more precise, but calculating the gradient is time-consuming, so it will take us a very long time to get to the bottom.

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

Answer: The Q-Q plot or quantile-quantile plot is a graphical technique for determining if two data sets come from populations with a common distribution.

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.



Use of Q-Q plot in Linear Regression: The Q-Q plot is used to see if the points lie approximately on the line. If they don't, it means, our residuals aren't Gaussian (Normal) and thus, our errors are also not Gaussian.

Importance of Q-Q plot: Below are the points:

I. The sample sizes do not need to be equal.

II. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers.

III. The q-q plot can provide more insight into the nature of the difference than analytical methods.