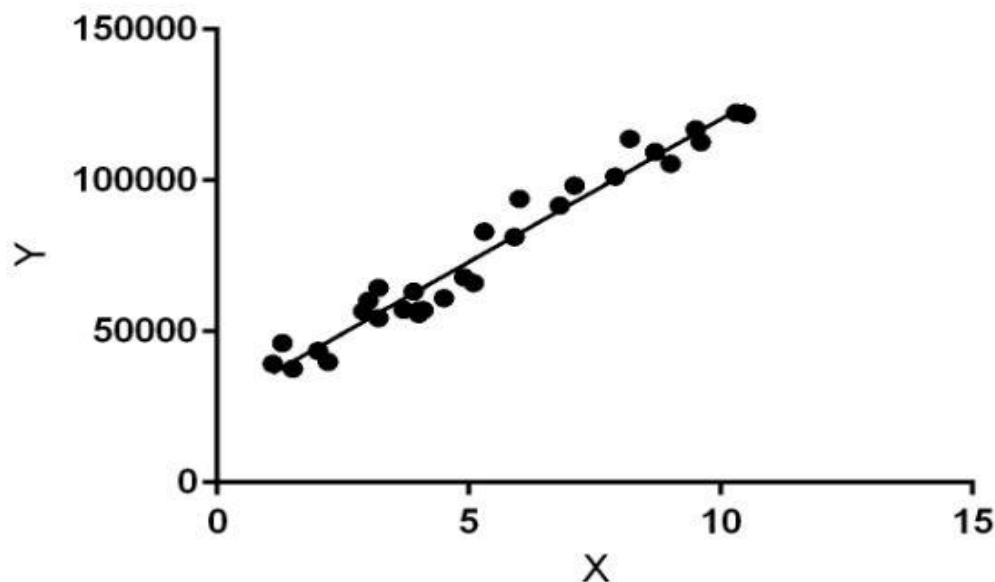


Subjective Questions for Linear Regression

Q: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. 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.

There are two main types:

1) Simple regression : Simple linear regression uses traditional slope-intercept form, where mm and bb are the variables our algorithm will try to “learn” to produce the most accurate predictions. xx represents our input data and yy represents our prediction.

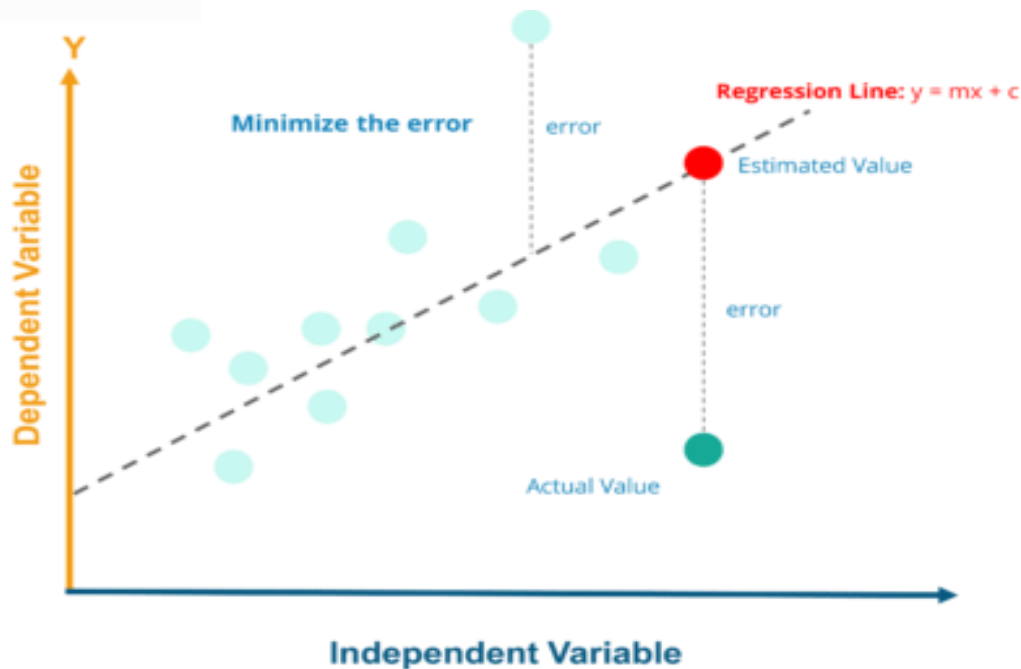
$$y=mx+by=mx+b$$

2) Multiple linear regression : A more complex, multi-variable linear equation might look like this, where w represents the coefficients, or weights, our model will try to learn.

$$f(x,y,z)=w_1x+w_2y+w_3z$$

The variables x, y, z represent the attributes, or distinct pieces of information, we have about each observation. For sales predictions, these attributes might include a company's advertising spend on radio, TV, and newspapers.

$$\text{Sales} = w_1 \text{Radio} + w_2 \text{TV} + w_3 \text{News}$$



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

Assumptions about the Linear Regression residuals:

- 1) Normality assumption:** It is assumed that the error terms, $\epsilon^{(i)}$, are normally distributed.
- 2) 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.

3) Constant variance assumption: It is assumed that the residual terms have the same (but unknown) variance, σ^2 . This assumption is also known as the assumption of homogeneity or homoscedasticity.

4) Independent error assumption: It is assumed that the residual terms are independent of each other, i.e., their pair-wise covariance is zero.

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

Coefficient of correlation is “R” value which is given in the summary table in the Regression output. In other words Coefficient of Determination is the square of Coefficient of Correlation.

R square or coeff. 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 perfectly opposites. One goes up and other goes down, in perfect negative way. 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.

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

In “Numerical calculations are exact, but graphs are rough”.

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 I 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

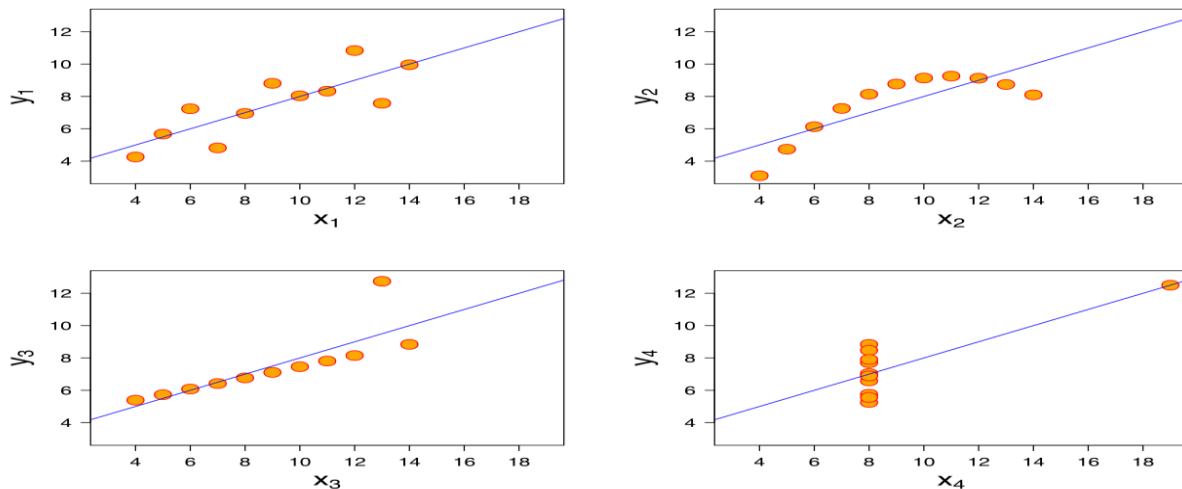
Quartet's Summary Stats

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.

Q.5 : What is Pearson's R?




Correlation is a technique for investigating the relationship between two quantitative, continuous variables, for example, age and blood pressure. Pearson's correlation coefficient (r) is a **measure of the strength of the association** between the two variables.

The first step in studying the relationship between two continuous variables is to draw a scatter plot of the variables to check for linearity. The correlation coefficient should not be calculated if the relationship is not linear. However, conventionally, the independent (or explanatory) variable is plotted on the x-axis (horizontally) and the dependent (or response) variable is plotted on the y-axis (vertically).

The nearer the scatter of points is to a straight line, the higher the strength of association between the variables. Also, it does not matter what measurement units are used.

Values of Pearson's correlation coefficient :

Pearson's correlation coefficient (r) for continuous (interval level) data ranges from -1 to +1:

$r = -1$		data lie on a perfect straight line with a negative slope
$r = 0$		no linear relationship between the variables
$r = +1$		data lie on a perfect straight line with a positive slope

$r = 1$ means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)

$r = -1$ means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)

$r = 0$ means there is no linear association

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

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

What is Scaling ? : Feature scaling is an important technique in Machine Learning and it is one of the most important steps during the preprocessing of data before creating a model. This can make a difference between a weak model and a strong one. The two most important scaling techniques are Standardization and Normalization.

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.

Why scaling is performed? : Scaling is important in the algorithm where distance between the data points is important. For example, in the dataset containing prices of products; without scaling, SVM might treat 1 USD equivalent to 1 INR though 1 USD = 65 INR.

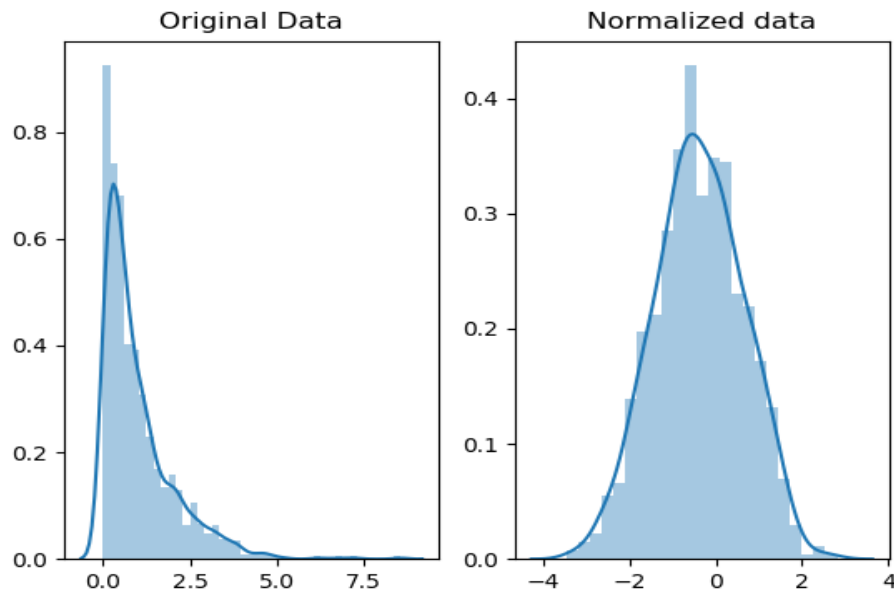
Normalization Scaling :

The point of normalization is to change your observations so that they can be described as a normal distribution.

Normal distribution is a specific statistical distribution where a roughly equal observations fall above and below the mean, the mean and the median are the same, and there are more observations closer to the mean.

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

For normalization, the maximum value you can get after applying the formula is 1, and the minimum value is 0. So all the values will be between 0 and 1.



In scaling, you're changing the range of your data while in normalization you're changing the shape of the distribution of your data.

Standardization Scaling :

Standardization transforms your data such that the resulting distribution has a mean of 0 and a standard deviation of 1.

$$\mathbf{x'} = (\mathbf{x} - \mathbf{x_{mean}}) / \sigma$$

where \mathbf{x} is the original feature vector, $\mathbf{x_{mean}}$ is the mean of that feature vector, and σ is its standard deviation.

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

VIF provides information about the variance in the estimated regression coefficient increased due to collinearity.

$$\mathbf{VIF} = 1 / (1 - R^2)$$

If R-square is 1 then VIF will be infinite. $R^2 = 1$ comes when datapoints are perfectly fitted in line of equation. So, it means one independent variable is collinear with other. Such variable needs to be eliminated.

Q.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.

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.

Purpose of the Assumptions :

Checking how well our data matches these assumptions is an important part of estimating regression coefficients. When you know where these conditions are violated, you may be able to plan ways to change your experiment setup to help your situation fit the ideal Gauss Markov situation more closely.

In practice, the Gauss Markov assumptions are **rarely all met perfectly**

The Gauss-Markov Assumptions In Algebra :

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

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

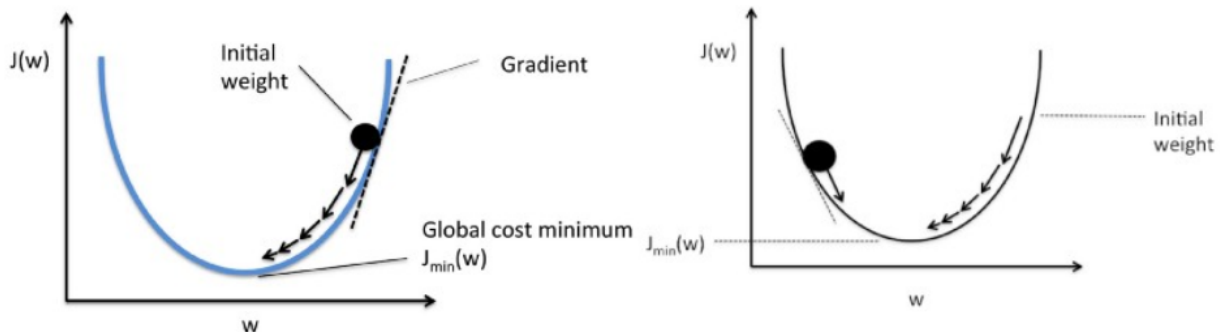
The first of these assumptions can be read as “The expected value of the error term is zero.”.

The second assumption is collinearity, the third is exogeneity, and the fourth is homoscedasticity.

Q.9 : Explain the gradient descent algorithm in detail.

Gradient descent is an optimisation algorithm. In linear regression, it is used to optimise the cost function and find the values of the β s (estimators) corresponding to the optimised value of the cost function.

Gradient descent works like a ball rolling down a graph (ignoring the inertia). The ball moves along the direction of the greatest gradient and comes to rest at the flat surface (minima).



Gradient Descent

Mathematically, the aim of gradient descent for linear regression is to find the solution of $\text{ArgMin } J(\theta_0, \theta_1)$, where $J(\theta_0, \theta_1)$ is the cost function of the linear regression. It is given by:

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$$

Here, h is the linear hypothesis model, $h = \Theta_0 + \Theta_1 x$, y is the true output, and m is the number of datapoints in the training set.

Gradient descent starts with a random solution, and then, based on the direction of the gradient, the solution is updated to the new value, where the cost function has a lower value.

The update is:

Repeat until convergence:

$$\theta_j = \theta_j - \alpha \frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)}) \cdot x_j^{(i)} \text{ for } j = 1, 2, \dots, n$$

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

What is a Q-Q Plot? :

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q-Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q-Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

How to Make a Q-Q Plot :

Sample question: Do the following values come from a normal distribution?
7.19, 6.31, 5.89, 4.5, 3.77, 4.25, 5.19, 5.79, 6.79.

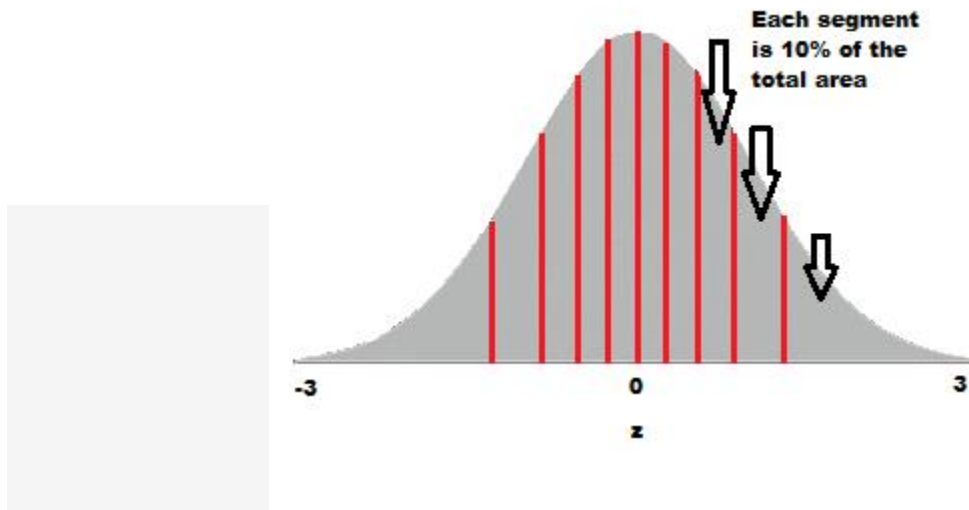
Step 1: Order the items from smallest to largest.

- 3.77

- 4.25
- 4.50
- 5.19
- 5.89
- 5.79
- 6.31
- 6.79
- 7.19

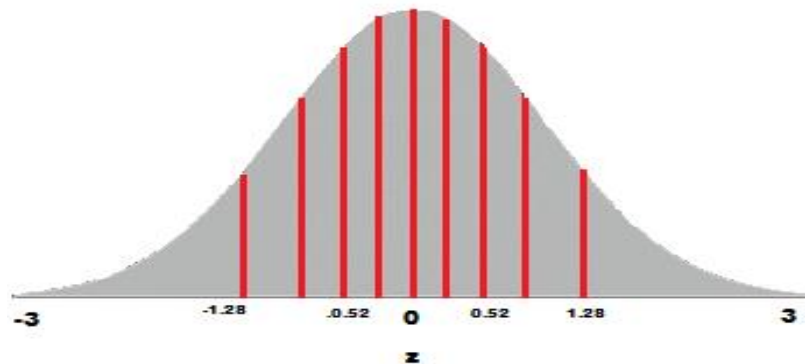
Step 2: Draw a normal distribution curve.

Divide the curve into $n+1$ segments. We have 9 values, so divide the curve into 10 equally-sized areas. For this example, each segment is 10% of the area (because $100\% / 10 = 10\%$).



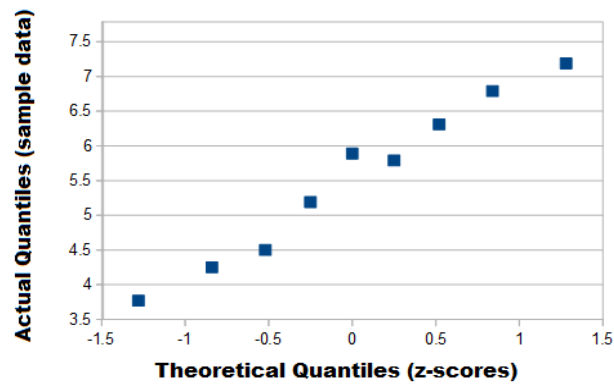
Step 3: Find the z-value (cut-off point) for each segment in Step 3. These segments are *areas*, so refer to a z-table (or use software) to get a z-value for each segment. The z-values are:

- 10% = -1.28
- 20% = -0.84
- 30% = -0.52
- 40% = -0.25
- 50% = 0
- 60% = 0.25
- 70% = 0.52
- 80% = 0.84
- 90% = 1.28
- 100% = 3.0



Step 4:

Plot your data set values (Step 1) against your normal distribution cut-off points (Step 3).
I used Open Office for this chart:



The (almost) straight line on this q q plot indicates the data is approximately normal.