ASSIGNMENT BASED SUBJECTIVE QUESTIONS

1) Categorical variables require special attention in regression analysis because, unlike dichotomous or continuous variables, they cannot by entered into the regression equation just as they are. So, hence we create dummy variables for the categorical columns and it helps in plotting the regression plot and interpreting the correlation with the target variable.

2) It depends on the model. If we don't drop the first column then the dummy variables will be correlated. This may affect some models adversely and the effect is stronger when the cardinality is smaller.

   For example, iterative models may have trouble converging and lists of variable importance may be distorted. If we have a small number of dummies, then could remove the first dummy.

   For example, if we have a variable gender, we don't need both a male and female as dummy variables. Just one will be fine. If male=1 then the person is a male and if male=0 then the person is female. However, if we have a category with hundreds of values, then we would not drop the first column. That will make it easier for the model to see all the categories quickly during learning.

3) Looking at the pair plot among the numerical variables, Registered Column has the highest correlation with the target variable (count).
4) The Linear Regression has five key assumptions:
   - Linear relationship
   - Multivariate normality
   - No or little multi-collinearity
   - No auto-correlation
   - Homoscedasticity

   First, 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 two cases, where no and little linearity is present.

   Secondly, the 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, e.g., the Kolmogorov-Smirnov test. When the data is not normally distributed a non-linear transformation (e.g., log-transformation) might fix this issue.

Thirdly, linear regression assumes that there is little or no multi-collinearity in the data. Multi-collinearity occurs when the independent variables are too highly correlated with each other.

Multi-collinearity may be tested with three central criteria:

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

b) 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 multi-collinearity in the data and with $T < 0.01$ there certainly is.

c) 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 multi-collinearity may be present; with $VIF > 10$ there is certainly multi-collinearity among the variables.
If multi-collinearity is found in the data, centering the data (that is deducting the mean of the variable from each score) might help to solve the problem. However, the simplest way to address the problem is to remove independent variables with high VIF values.
Fourth, 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.

d) Condition Index – the condition index is calculated using a factor analysis on the independent variables. Values of 10-30 indicate a mediocre multi-collinearity in the linear regression variables, values > 30 indicate strong multi-collinearity.
Multi-collinearity is found in the data centering the data, that is deducting the mean score might help to solve the problem. Other alternatives to tackle the problems is conducting a factor analysis and rotating the factors to insure independence of the factors in the linear regression analysis.

Fourthly, 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. In other words when the value of $y(x+1)$ is not independent from the value of $y(x)$. While a scatterplot allows you to check for autocorrelations, you can test the linear regression model for autocorrelation with the Durbin-Watson test. Durbin-Watson's d tests the null hypothesis that the residuals are not linearly auto-correlated. While d can assume values between 0 and 4, values around 2 indicate no autocorrelation. As a rule of thumb values of $1.5 < d < 2.5$ show that there is no auto-correlation in the data. However, the Durbin-Watson test only analyses linear autocorrelation and only between direct neighbors, which are first order effects.

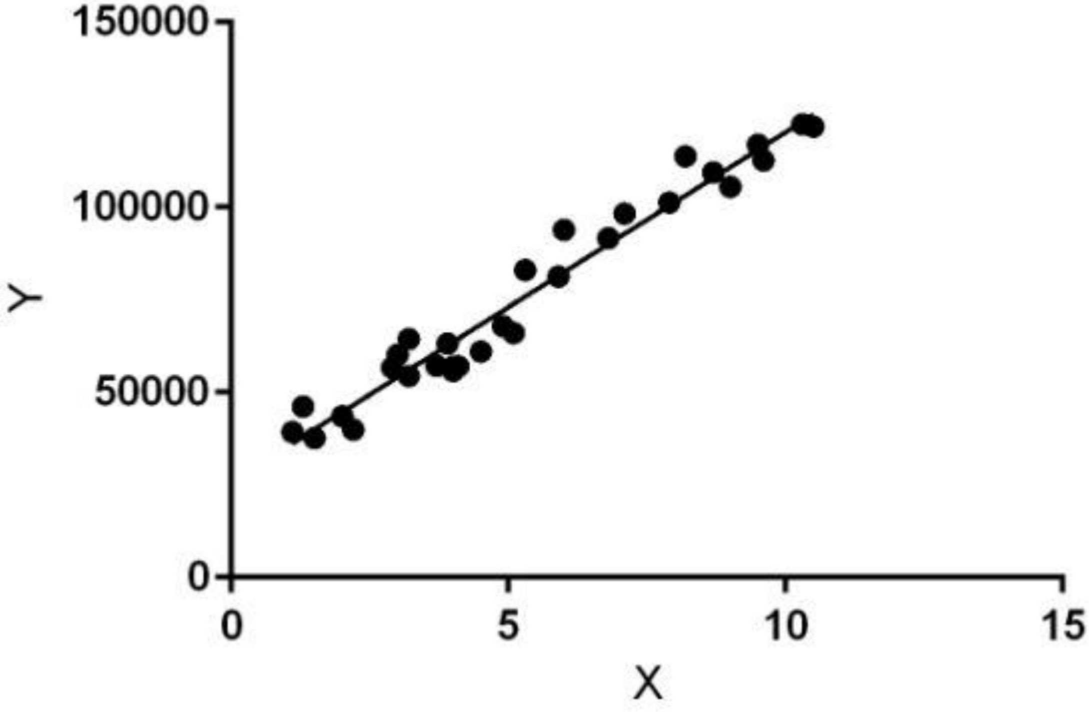
The last assumption of the linear regression analysis is homoscedasticity. The scatter plot is good way to check whether the data are homoscedastic (meaning the residuals are equal across the regression line). The following scatter plots show examples of data that are not homoscedastic. The Goldfeld-Quandt 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.

5) Based on the final model, the top 3 features contributing significantly towards explaining the demand on the shared bikes are-

- Temp
- Year
- September

General Subjective Questions

1) 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 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 :

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

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 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.

θ_1 and θ_2 values get updated to form the best fit line as:

Cost Function (J):

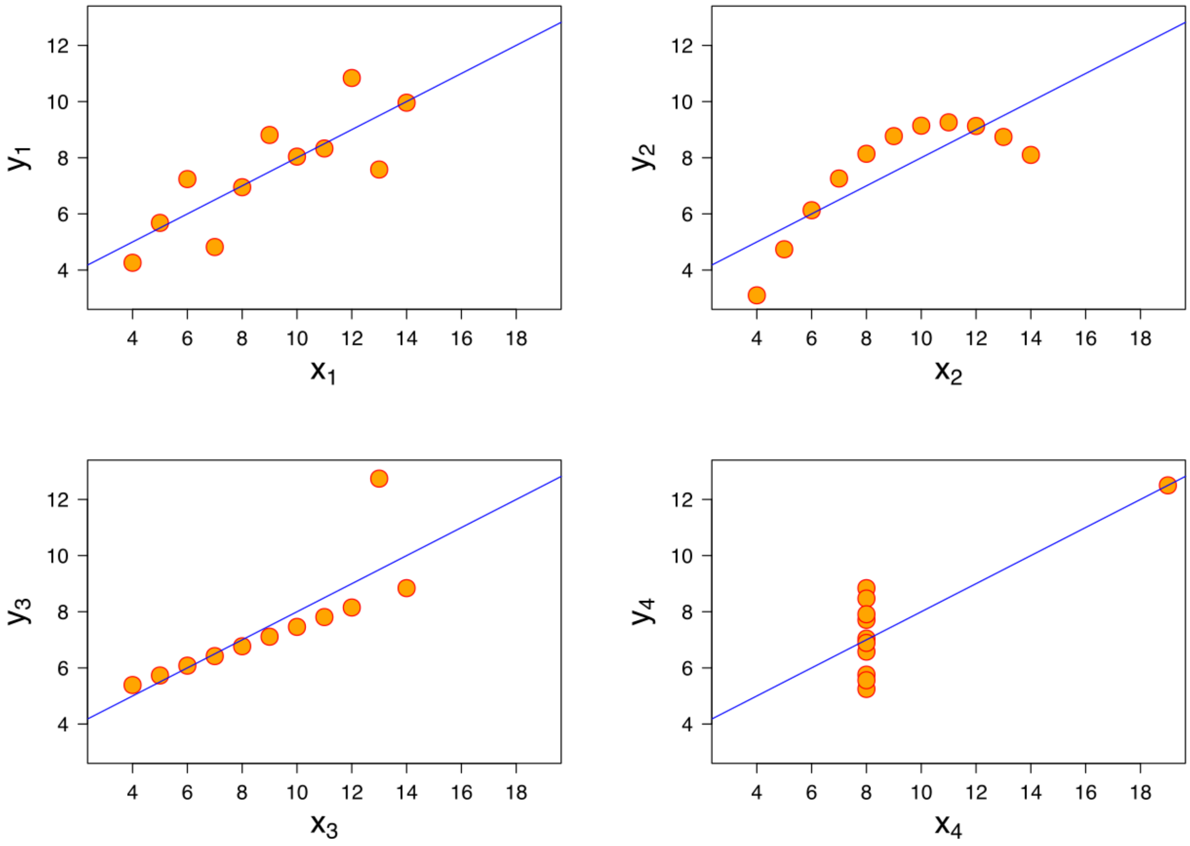
By achieving the best-fit regression line, the model aims to predict y value such that the error difference between predicted value and true value is minimum. So, it is very important to update the θ_1 and θ_2 values, to reach the best value that minimize the error between predicted y value (pred) and true y value (y).

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

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

Cost function(J) of Linear Regression is the Root Mean Squared Error (RMSE) between predicted y value (pred) and true y value (y).

2) Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and 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 analyzing it and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough".



For all four datasets:

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x :	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y :	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression :	0.67	to 2 decimal places

- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.
- The second graph (top right) is not distributed normally; while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third graph (bottom left), the distribution is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze 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

It is not known how Anscombe created his datasets. Since its publication, several methods to generate similar data sets with identical statistics and dissimilar graphics have been developed.

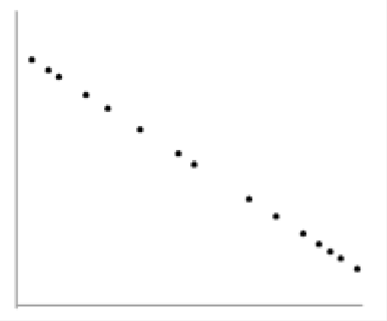
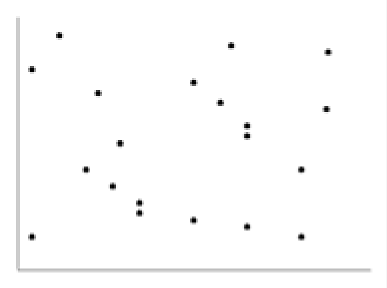
3) Correlation is a technique for investigating the relationship between two quantitative, continuous variables, for example, age and blood pressure. Pearson's correlation coefficient (r) is a **measure of the strength of the association** between the two variables.

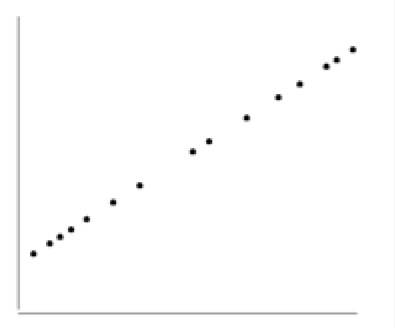
The first step in studying the relationship between two continuous variables is to draw a scatter plot of the variables to check for linearity. The correlation coefficient should not be calculated if the relationship is not linear. For correlation only purposes, it does not really matter on which axis the variables are plotted. 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

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.

4) I was recently working with a dataset that had multiple features spanning varying degrees of magnitude, range, and units. This is a significant obstacle as a few machine learning algorithms are highly sensitive to these features.

I’m sure most of you must have faced this issue in your projects or your learning journey. For example, one feature is entirely in kilograms while the other is in grams, another one is liters, and so on. How can we use these features when they vary so vastly in terms of what they’re presenting?

Here’s the curious thing about feature scaling – it improves (significantly) the performance of some machine learning algorithms and does not work at all for others. What could be the reason behind this quirk?

Also, what’s the difference between normalization and standardization? These are two of the most commonly used feature scaling techniques in machine learning but a level of ambiguity exists in their understanding. When should you use which technique?

I will answer these questions and more in this article on feature scaling. We will also implement feature scaling in Python to give you a practice understanding of how it works for different machine learning algorithms.

What is Normalization?

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.

Here’s the formula for normalization:

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

Here, Xmax and Xmin 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

What is Standardization?

Standardization is another scaling technique where the values are centered around 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.

Here’s the formula for standardization:

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

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.

Now, the big question in your mind must be when should we use normalization and when should we use standardization? Let’s find out!

The Big Question – Normalize or Standardize?

Normalization vs. standardization is an eternal question among machine learning newcomers. Let me elaborate on the answer in this section.

- Normalization is good to use when you know that the distribution of your 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 Neighbors and Neural Networks.
- 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 you have outliers in your data, they will not be affected by standardization.

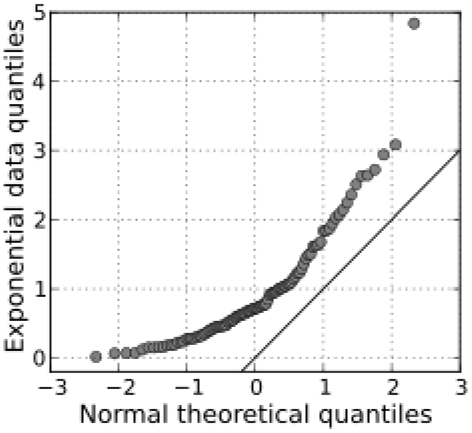
However, at the end of the day, the choice of using normalization or standardization will depend on your problem and the machine learning algorithm you are using. There is no hard and fast rule to tell you when to normalize or standardize your data. **You can always start by fitting your model to raw, normalized and standardized data and compare the performance for best results.**

It is a good practice to fit the scaler on the training data and then use it to transform the testing data. This would avoid any data leakage during the model testing process. Also, the scaling of target values is generally not required.

5) In general one starts with the selection of all variables, and proceeds by repeatedly deselecting variables showing a high **VIF**. ... An **infinite VIF value** indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an **infinite VIF** as well).

6) **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.



A Q Q plot showing the 45 degree reference line. Image: skbkekas|Wikimedia Commons.

The image above shows quantiles from a theoretical normal distribution on the horizontal axis. It's being compared to a set of data on the y-axis. This particular type of Q Q plot is called a **normal quantile-quantile (QQ) plot.** The points are not clustered on the 45 degree line, and in fact follow a curve, suggesting that the sample data is not normally distributed.

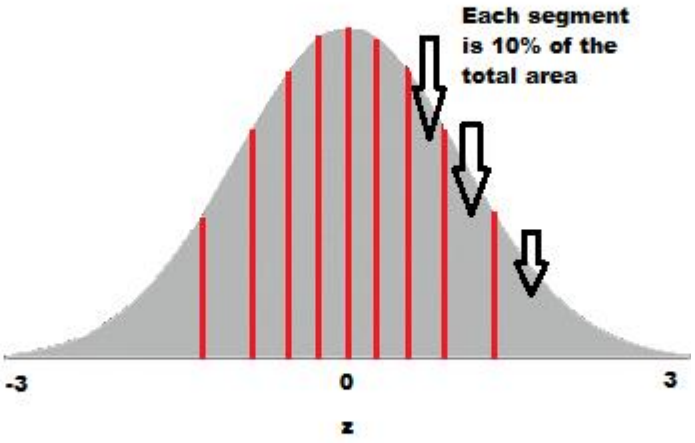
How to Make a Q Q Plot

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

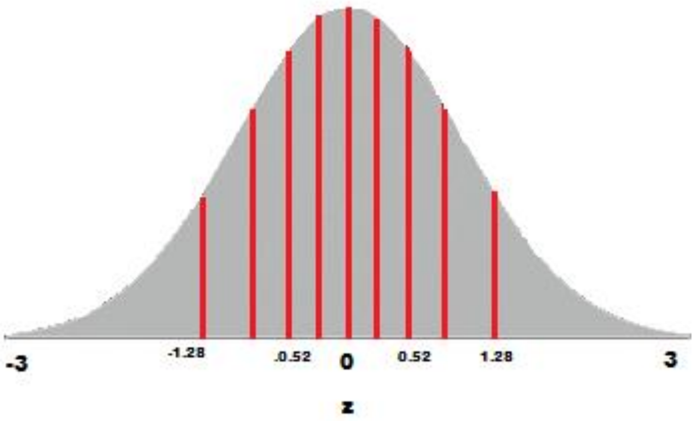
Step 2: **Draw a normal distribution curve.** Divide the curve into n+1 segments. We have 9 values, so divide the curve into 10 equally-sized areas. For this example, each segment is 10% of the area (because 100% / 10 = 10%).



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:

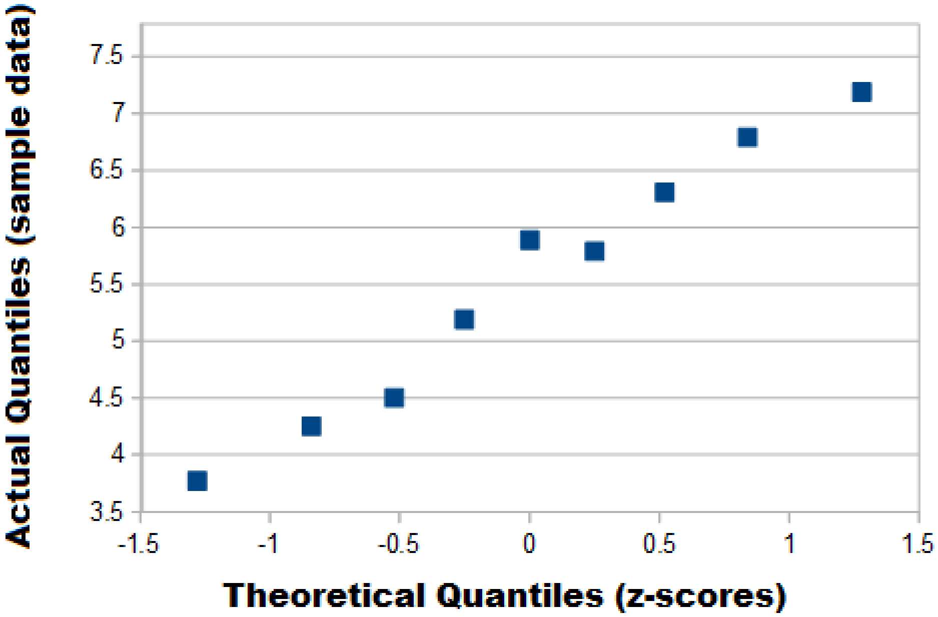
- 10% = -1.28
- 20% = -0.84
- 30% = -0.52
- 40% = -0.25
- 50% = 0

- 60% = 0.25
- 70% = 0.52
- 80% = 0.84
- 90% = 1.28
- 100% = 3.0



A few of the z-values plotted on the graph.

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.

Note: This example used the standard normal distribution, but if think your data could have come from a different normal distribution (i.e. one with a different mean and standard deviation) then you could use that instead.

Q Q Plots and the Assumption of Normality

The assumption of normality is an important assumption for many statistical tests; you assume you are sampling from a normally distributed population. The normal Q Q plot is one way to assess normality. However, you don’t have to use the normal distribution as a comparison for your data; you can use any continuous distribution as a comparison (for example a Weibull distribution or a uniform distribution), as long as you can calculate the quantiles. In fact, a common procedure is to test out several different distributions with the Q Q plot to see if one fits your data well.