Bike Sharing Assignment using Linear regression

Problem Statement:

A bike-sharing system is a service in which bikes are made available for shared use to individuals. A US bike-sharing provider **BoomBikes** has recently suffered considerable dips in their revenues due to the ongoing Corona pandemic. The company is finding it very difficult to sustain in the current market scenario. So, it has decided to come up with a mindful business plan to be able to accelerate its revenue as soon as the ongoing lockdown comes to an end, and the economy restores to a healthy state.

So company decide to understand the factors on which the demand for these shared bikes depends. Specifically, they want to understand the factors affecting the demand for these shared bikes in the American market.

- ✓ Which variables are significant in predicting the demand for shared bikes?
- ✓ How well those variables describe the bike demands?

Based on various meteorological surveys and people's styles, the service provider firm has gathered a large dataset on daily bike demands across the American market based on some factors.

Business Goal:

We are required to model the demand for shared bikes with the available independent variables. It will be used by the management to understand how exactly the demands vary with different features. They can accordingly manipulate the business strategy to meet the demand levels and meet the customer's expectations. Further, the model will be a good way for management to understand the demand dynamics of a new market.

Mohamed Khaleelulla

Assignment-based Subjective Questions:

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Following are the Identified categorical variable from dataset, which is infer on the dependent variable count (i.e., cnt)

yr, holiday, workingday, season & weathersit

- Every next year (i.e., from 2018 to 2019) there is an increase in bike rentals.
- Bike rentals is increase in summer & winter and negative correlation with spring.
- Holiday & workingday is playing a role of bike rental demand. When there is a holiday bike rentals demand is reducing.
- Weathersit, is a factor of climatical infer on the customer. Mist is negative correlation.

2. Why is it important to use **drop_first=True** during dummy variable creation?

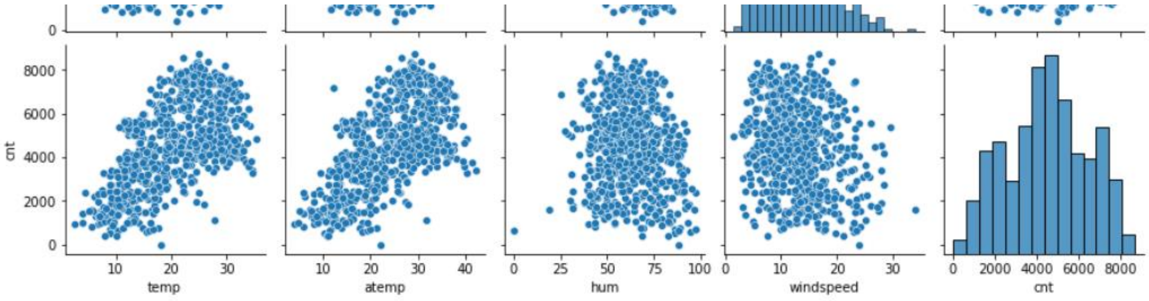
Idea is to reduce the number of variable features taken for building the good linear regression model, as such increase in redundant feature should impact the Adjusted R-Squared value and predict model. So, when you have a categorical variable with 'n' levels, dummy variable creation is to build 'n-1' variables, indicating the levels.

For a variable, season (1: spring, 2: summer, 3: fall, 4: winter) would create a dummy variables by dropping first like the following and not necessary in all the cases.

- 000 will correspond to Fall
- 100 will correspond to Spring
- 010 will correspond to Summer
- 001 will correspond to Winter

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

temp & atemp is the highest correlation with target variable cnt. i.e., 0.63 coefficients from heatmap

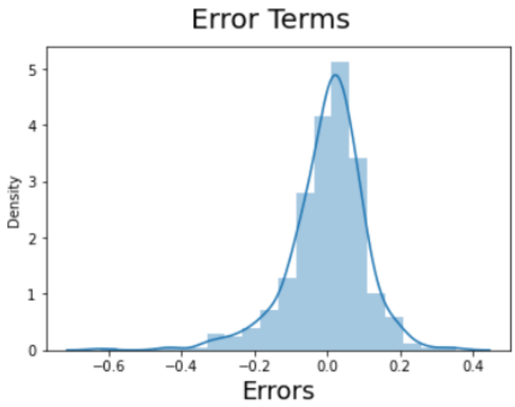


4. How did you validate the assumptions of Linear Regression after building the model on the training set?

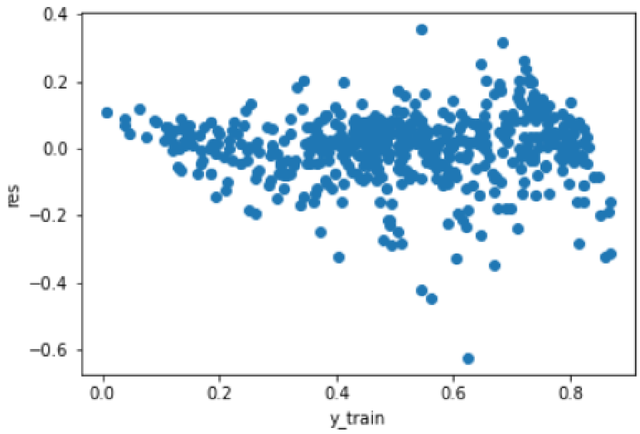
Assumptions of Linear regression are

- Model has a linear relationship between the independent and dependent variable, i.e., X&y. in our case other predicting variables and dependent cnt.

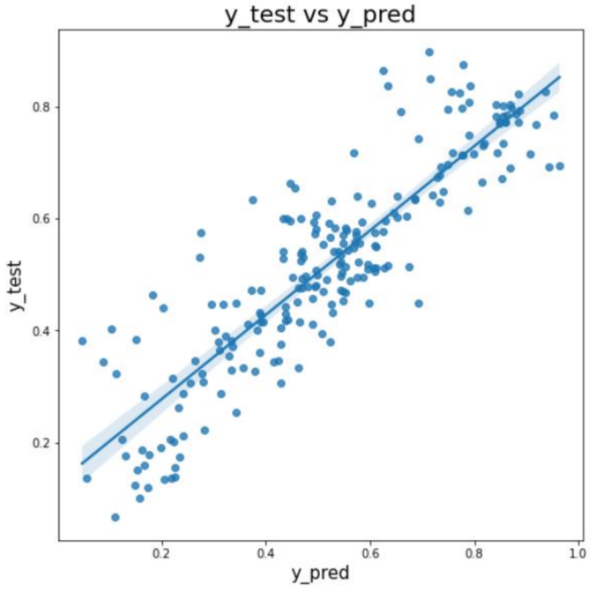
- Error term are normally distributed.



- Error term are independent of each other.



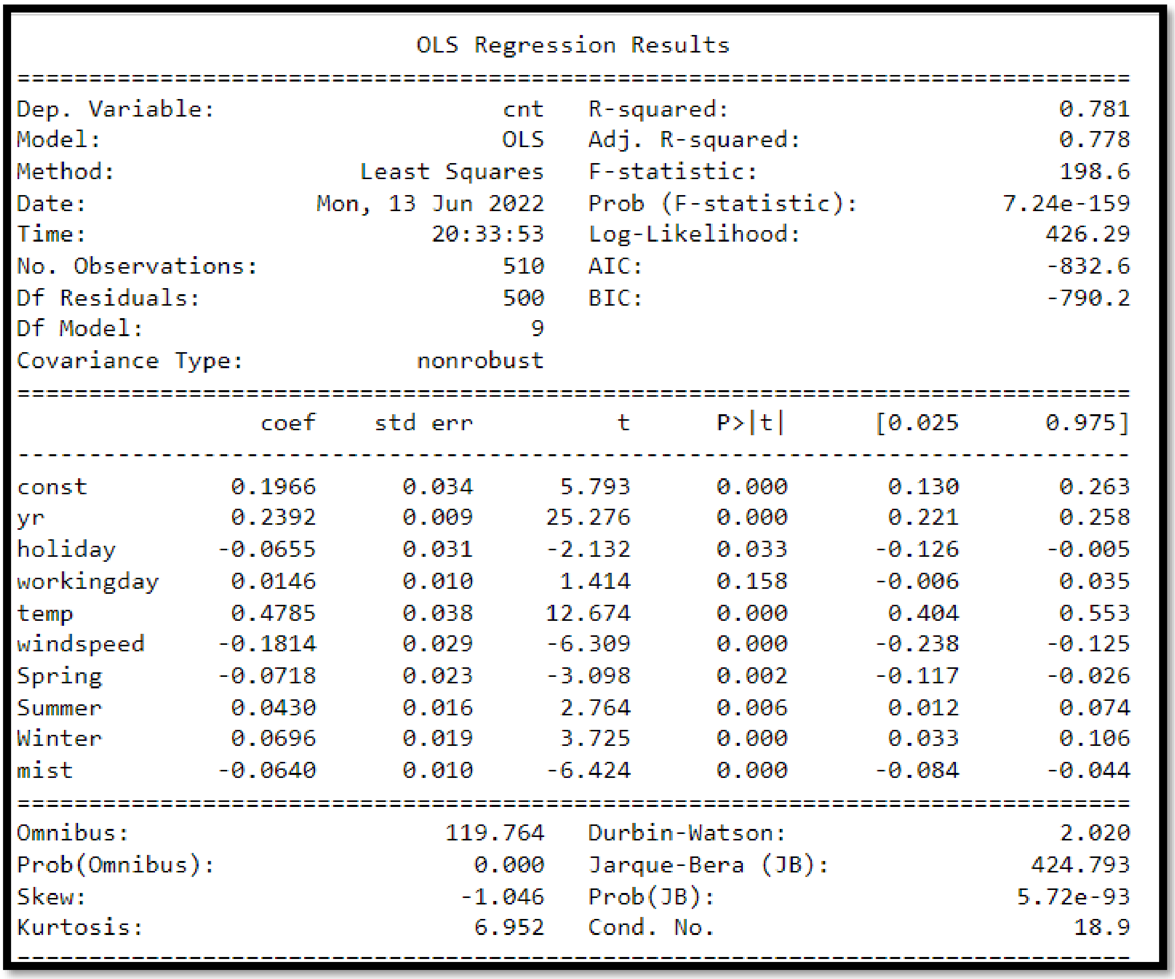
- **Error terms have *constant variance* (homoscedasticity):**



5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Below are the top 3 variables contributing significant toward an increase in demand of shared bikes.

- temp
- yr
- and, winter.

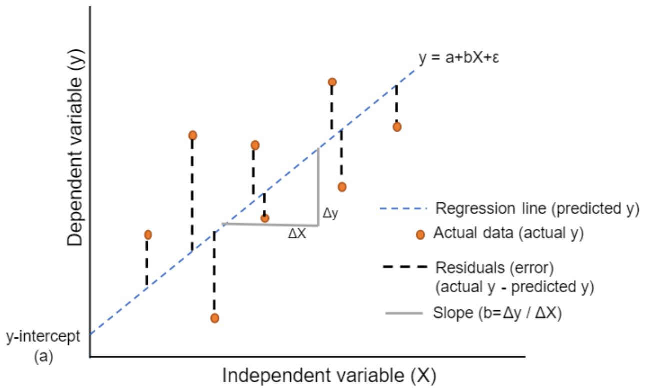


```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.781
Model:                            OLS   Adj. R-squared:                  0.778
Method:                 Least Squares   F-statistic:                     198.6
Date:                Mon, 13 Jun 2022   Prob (F-statistic):          7.24e-159
Time:                        20:33:53   Log-Likelihood:                 426.29
No. Observations:                 510   AIC:                            -832.6
Df Residuals:                     500   BIC:                            -790.2
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.1966      0.034      5.793      0.000       0.130       0.263
yr             0.2392      0.009     25.276      0.000       0.221       0.258
holiday       -0.0655      0.031     -2.132      0.033      -0.126      -0.005
workingday     0.0146      0.010      1.414      0.158      -0.006       0.035
temp           0.4785      0.038     12.674      0.000       0.404       0.553
windspeed     -0.1814      0.029     -6.309      0.000      -0.238      -0.125
Spring        -0.0718      0.023     -3.098      0.002      -0.117      -0.026
Summer         0.0430      0.016      2.764      0.006       0.012       0.074
Winter         0.0696      0.019      3.725      0.000       0.033       0.106
mist          -0.0640      0.010     -6.424      0.000      -0.084      -0.044
==============================================================================
Omnibus:                      119.764   Durbin-Watson:                   2.020
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              424.793
Skew:                          -1.046   Prob(JB):                     5.72e-93
Kurtosis:                       6.952   Cond. No.                         18.9
==============================================================================
```

General Subjective Questions:

1. Explain the linear regression algorithm in detail.
   Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as **sales, salary, age, product price,** etc. Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

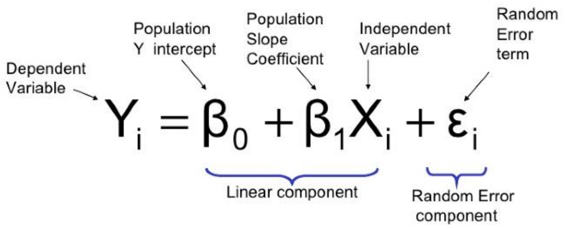
The linear regression model provides a sloped straight line representing the relationship between the variables. Consider the below image:



Linear regression can be further divided into two types of the algorithm:

Simple Linear Regression:

If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.



Multiple Linear regression:

If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

$$\boxed{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ \dots + \beta_i X_i}$$

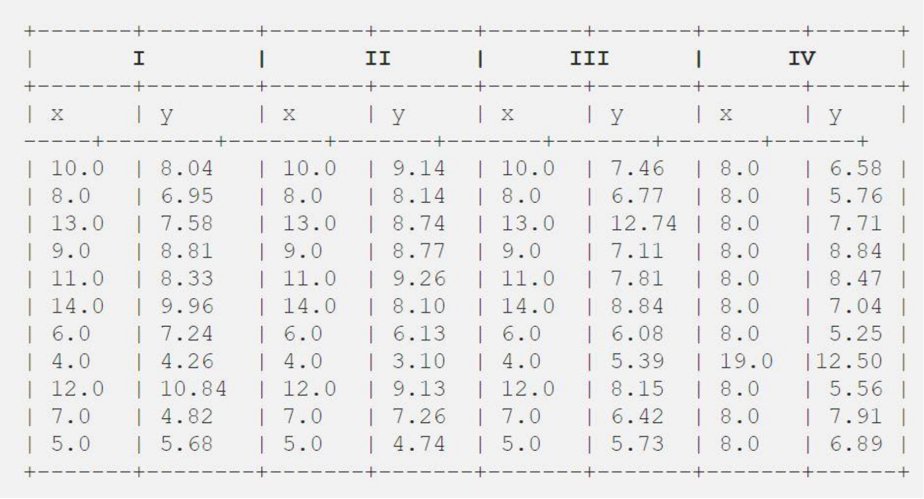
Y : Dependent variable
β_0 : Intercept
β_i : Slope for X_i
X = Independent variable

2. 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 analyzing it and the effect of outliers on statistical properties.

Simple understanding:

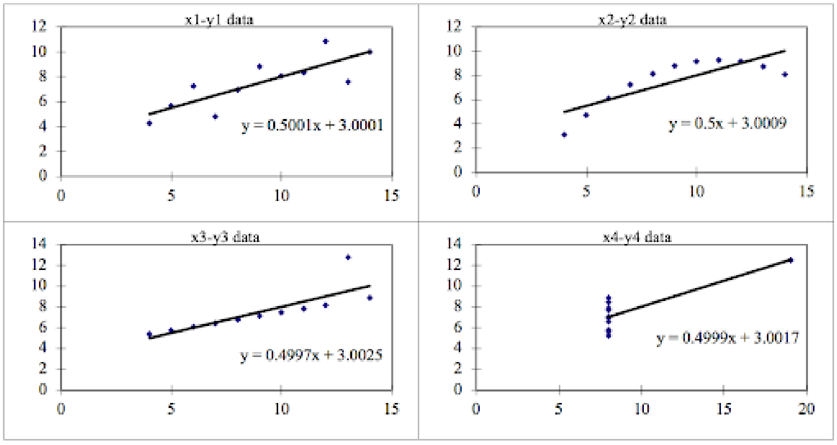
Once Francis John "Frank" Anscombe who was a statistician of great repute found 4 sets of 11 data-points in his dream and requested the council as his last wish to plot those points. Those 4 sets of 11 data-points are given below.



```
+-------+--------+-------+-------+-------+-------+-------+------+
|       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 |
+-------+--------+-------+-------+-------+-------+-------+------+
```

After that, the council analyzed them using only descriptive statistics and found the mean, standard deviation, and correlation between x and y.

Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

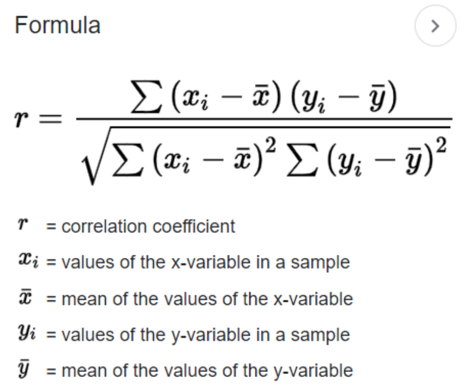


3. What is Pearson’s R?

In Statistics, the Pearson's Correlation Coefficient is also referred to as **Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation**. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.

Whenever we discuss correlation in statistics, it is generally Pearson's correlation coefficient. However, it cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.

Pearson's correlation coefficient is 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. Pearson's Correlation Coefficient is named after Karl Pearson. He formulated the correlation coefficient from a related idea by Francis Galton in the 1880s.



Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

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

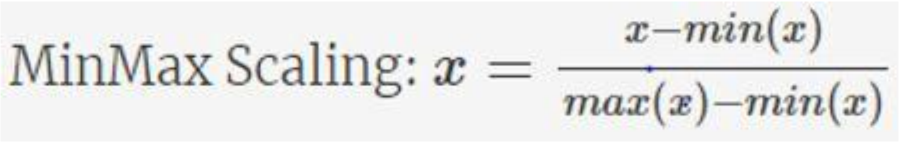
Scaling: It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Scaling is performed on collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Normalization/Min-Max Scaling:

- It brings all of the data in the range of 0 and 1
- sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.



$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:

- Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).
- sklearn.preprocessing.scale helps to implement standardization in python.
- One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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

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

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

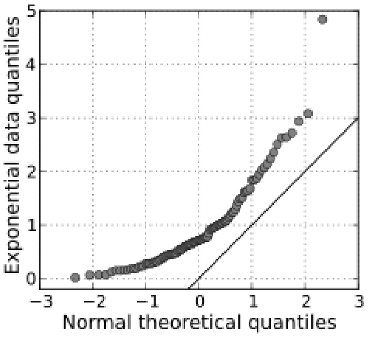
An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well)

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

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:



If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.