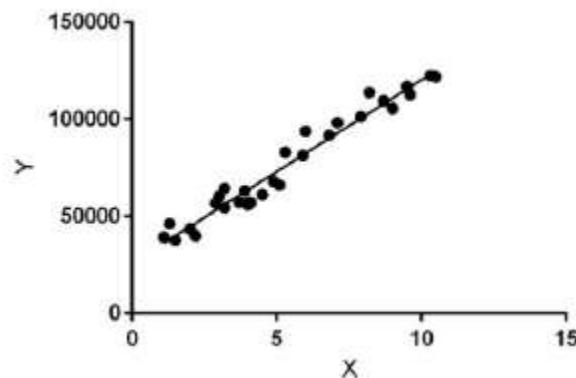


Developer Name :: Sachin Kumar
Email ID :: sachinkumar.ml10@iiitb.net
Roll Number :: DML1950087
Date :: 05/10/2019
Version :: 0.7

1. Explain the linear regression algorithm in detail.

Ans.

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.

$$y = \beta_0 + \beta_1 x$$

\downarrow \downarrow
 Intercept Slope

Once we find the best B0 and B1 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.

How to update B0 and B1 values to get the best fit line ?

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 B0 and B1 values, to reach the best value that minimize the error between predicted y value (pred) and true y value (y).

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

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

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

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

Ans: The assumptions of linear regression regarding residuals are:

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.

Explanations:

If the residuals are not normally distributed, their randomness is lost, which implies that the model is not able to explain the relation in the data.

Also, the mean of the residuals should be zero.

$$Y(i) = \beta_0 + \beta_1 x(i) + \varepsilon(i)$$

This is the assumed linear model, where ε is the residual term.

$$\begin{aligned} E(Y) &= E(\beta_0 + \beta_1 x(i) + \varepsilon(i)) \\ &= E(\beta_0 + \beta_1 x(i) + \varepsilon(i)) \end{aligned}$$

If the expectation(mean) of residuals, $E(\varepsilon(i))$, is zero, the expectations of the target variable and the model become the same, which is one of the targets of the model.

The residuals (also known as error terms) should be independent. This means that there is no correlation between the residuals and the predicted values, or among the residuals themselves. If some correlation is present, it implies that there is some relation that the regression model is not able to identify.

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

Ans:

Correlation Coefficient, r:

1. The quantity r , called the linear correlation coefficient, measures the strength and the direction of a linear relationship between two variables. The linear correlation coefficient is sometimes referred to as the Pearson product moment correlation coefficient in honor of its developer Karl Pearson.

2. The mathematical formula for computing r is:

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2} \sqrt{n(\sum y^2) - (\sum y)^2}}$$

where n is the number of pairs of data.

3. The value of r is such that $-1 < r < +1$. The $+$ and $-$ signs are used for positive linear correlations and negative linear correlations, respectively.
4. **Positive correlation:** If x and y have a strong positive linear correlation, r is close to $+1$. An r value of exactly $+1$ indicates a perfect positive fit. Positive values indicate a relationship between x and y variables such that as values for x increases, values for y also increase.
5. **Negative correlation:** If x and y have a strong negative linear correlation, r is close to -1 . An r value of exactly -1 indicates a perfect negative fit. Negative values indicate a relationship between x and y such that as values for x increase, values for y decrease.
6. **No correlation:** If there is no linear correlation or a weak linear correlation, r is close to 0 . A value near zero means that there is a random, nonlinear relationship between the two variables.
7. **A perfect correlation** of ± 1 occurs only when the data points all lie exactly on a straight line. If $r = +1$, the slope of this line is positive. If $r = -1$, the slope of this line is negative.
8. A correlation greater than 0.8 is generally described as strong, whereas a correlation less than 0.5 is generally described as weak. These values can vary based upon the "type" of data being examined. A study utilizing scientific data may require a stronger correlation than a study using social science data.

Coefficient of Determination, r^2 or R^2 :

1. The coefficient of determination, r^2 , is useful because it gives the proportion of the variance (fluctuation) of one variable that is predictable from the other variable.
2. It is a measure that allows us to determine how certain one can be in making predictions from a certain model/graph.
3. The coefficient of determination is the ratio of the explained variation to the total variation.
4. The coefficient of determination is such that $0 < r^2 < 1$, and denotes the strength of the linear association between x and y .
5. The coefficient of determination represents the percent of the data that is the closest to the line of best fit. For example, if $r = 0.922$, then $r^2 = 0.850$, which means that 85% of the total variation in y can be explained by the linear relationship between x and y (as described by the regression equation). The other 15% of the total variation in y remains unexplained.
6. The coefficient of determination is a measure of how well the regression line represents the data. If the regression line passes exactly through every point on the scatter plot, it would be able to explain all of the variation. The further the line is away from the points, the less it is able to explain.

4. Explain the Anscombe's quartet in detail?

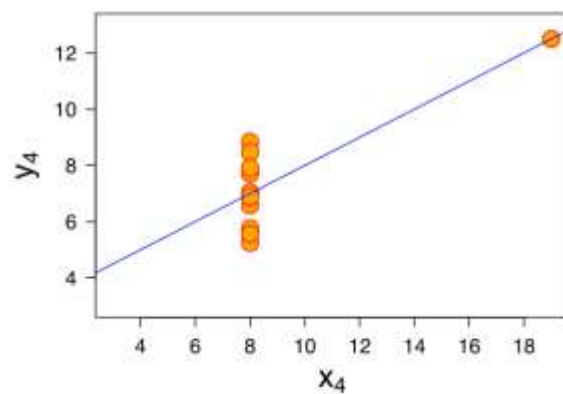
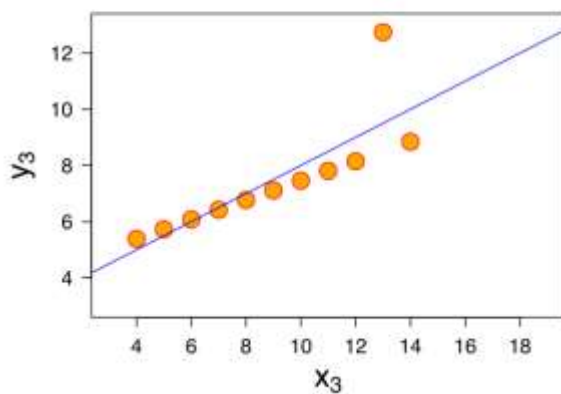
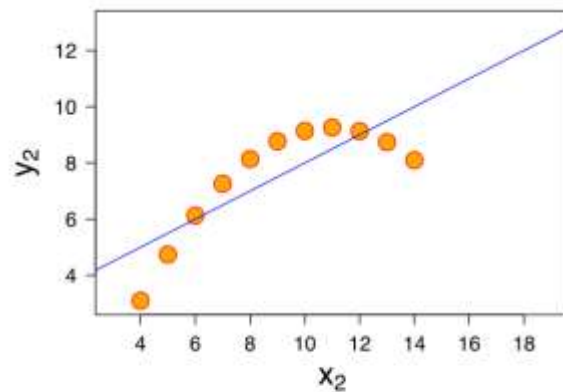
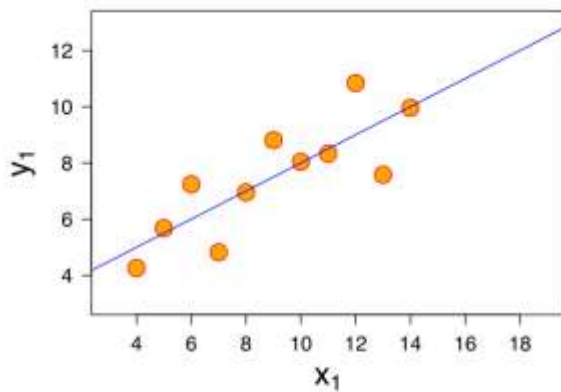
Ans. Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x,y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize **COMPLETELY**, 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

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 appear 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.

5. What is Pearson's R?

Ans. 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.


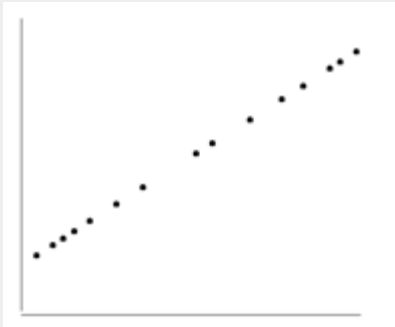
Values of Pearson's correlation coefficient

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

= -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

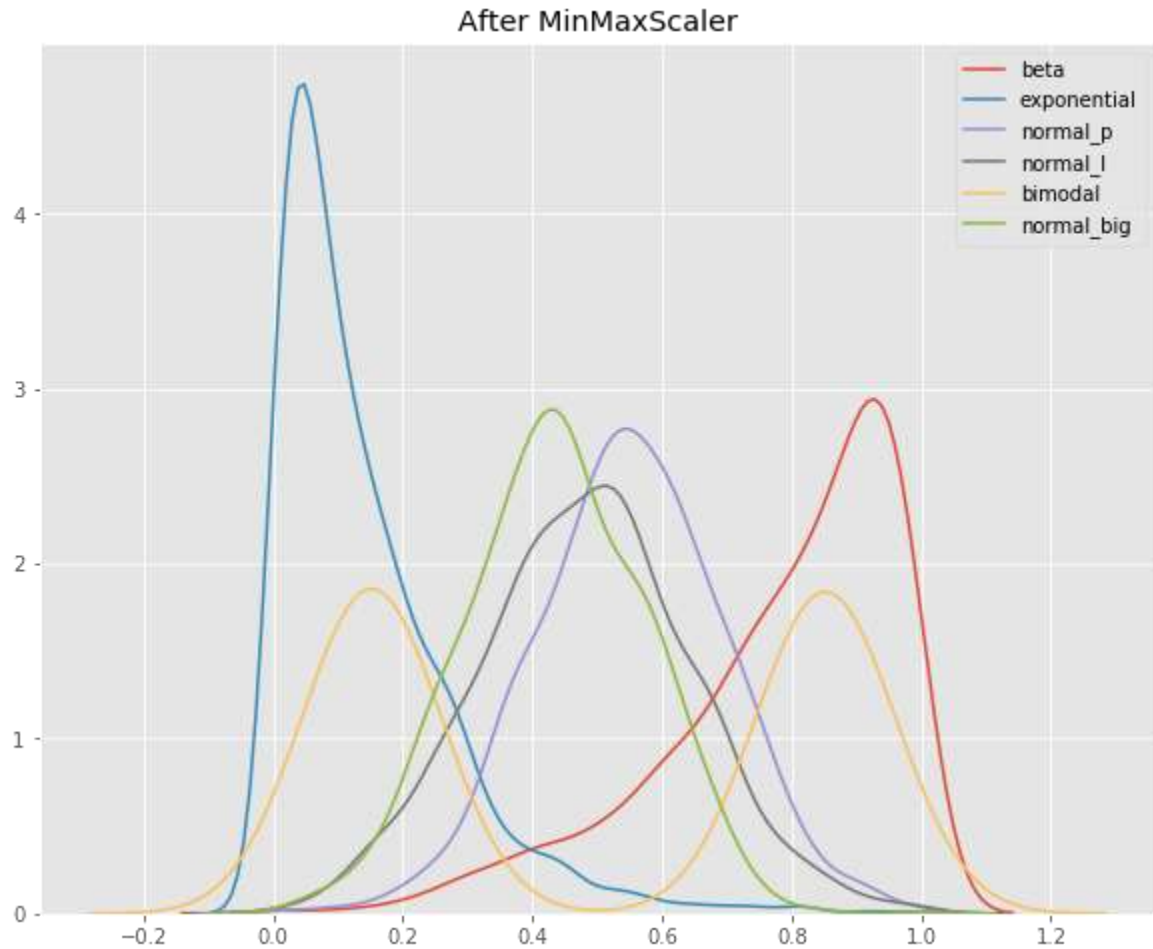
6. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans. 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.

Below is the difference between normalized & standardized scaling:

MinMaxScaler:

For each value in a feature, MinMaxScaler subtracts the minimum value in the feature and then divides by the range. The range is the difference between the original maximum and original minimum. MinMaxScaler preserves the shape of the original distribution. It doesn't meaningfully change the information embedded in the original data. Note that MinMaxScaler doesn't reduce the importance of outliers. The default range for the feature returned by MinMaxScaler is 0 to 1. Here's the kdeplot after MinMaxScaler has been applied.



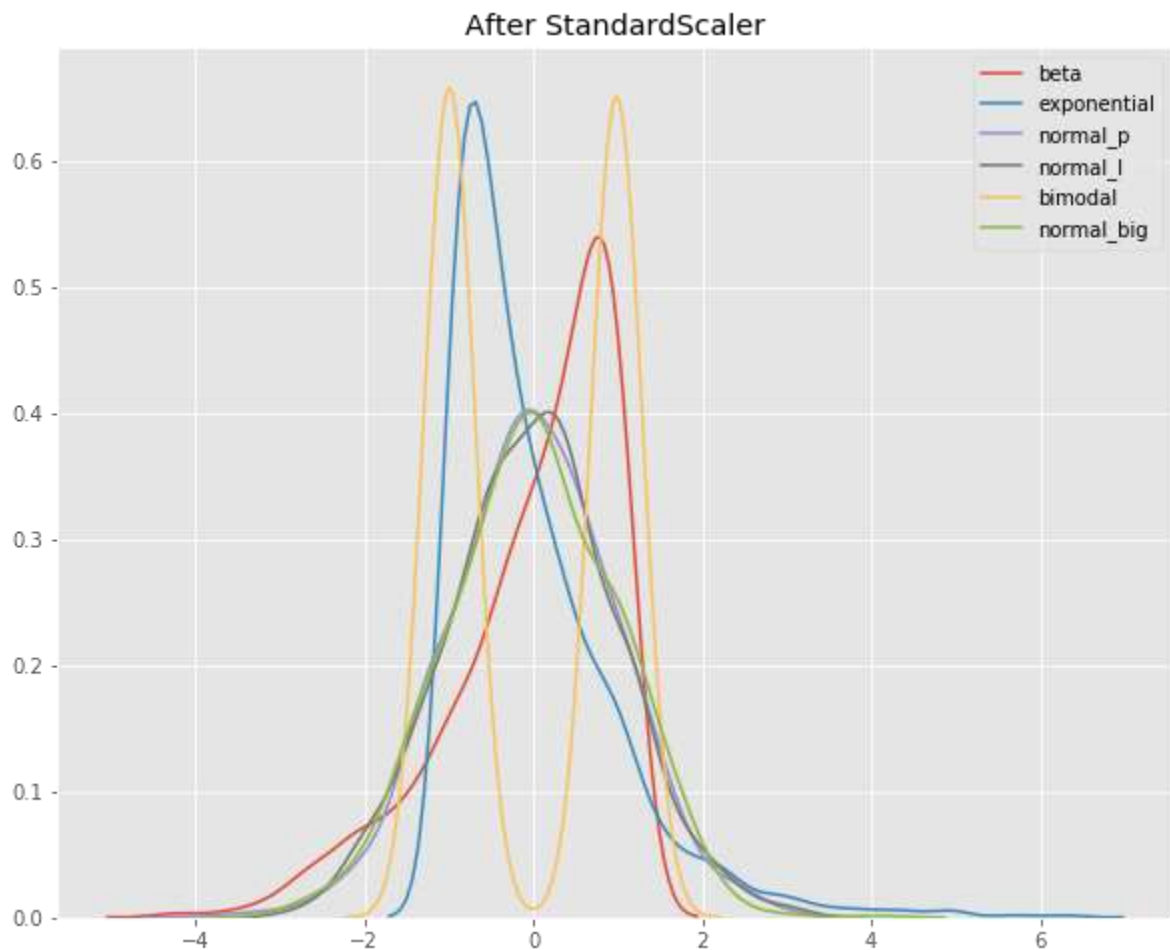
Notice how the features are all on the same relative scale. The relative spaces between each feature's values have been maintained. MinMaxScaler is a good place to start unless you know you want your feature to have a normal distribution or want outliers to have reduced influence.

StandardAero:

It standardizes a feature by subtracting the mean and then scaling to unit variance. Unit variance means dividing all the values by the standard deviation. Standard Scaler does not meet the strict definition of *scale* I introduced earlier. Standard Scaler results in a distribution with a standard deviation equal to 1. The variance is equal to 1 also, because

variance = standard deviation squared. And 1 squared = 1.

Standard Scaler makes the mean of the distribution 0. About 68% of the values will lie between -1 and 1.



In the plot above, you can see that all four distributions have a mean close to zero and unit variance. The values are on a similar scale, but the range is larger than after Min-Max Scaler. Deep learning algorithms often call for zero mean and unit variance. Regression-type algorithms also benefit from normally distributed data with small sample sizes. Standard Scaler does distort the relative distances between the feature values, so it's generally my second choice in this family of transformations.

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

Ans.

An infinite VIF will be returned for two variables that are exactly collinear, variables that are exactly

the same or linear transformations of each other. Perhaps check your variables for exact collinearity,

using the pairs function or something similar.

In VIF, each feature is regression against all other features. If R^2 is more which means this feature is

correlated with other features. [0] $VIF = 1 / (1 - R^2)$ When R^2 reaches 1, VIF reaches infinity.

8. What is the Gauss-Markov theorem?

Ans.

A theorem that proves that if the error terms in a multiple regression have the same variance and are

uncorrelated, then the estimators of the parameters in the model produced by least squares

estimation is better (in the sense of having lower dispersion about the mean) than any other unbiased

linear estimator.

This is pretty much considered the “big boy” reason least squares fitting can be considered a good implementation of linear regression.

Suppose you are building a model of the form:

$$y(i) = B \cdot x(i) + e(i)$$

where B is a vector (to be inferred), i is an index that runs over the available data (say 1 through n), $x(i)$ is a per-example vector of features, and $y(i)$ is the scalar quantity to be modeled. Only $x(i)$ and $y(i)$ are observed. The $e(i)$ term is the unmodeled component of $y(i)$ and you typically hope that the $e(i)$ can be thought of unknowable effects, individual variation, ignorable errors, residuals, or noise. How

weak/strong assumptions you put on the $e(i)$ (and other quantities) depends on what you know, what you are trying to do, and which theorems you need to meet the pre-conditions of. The Gauss-Markov theorem assures a good estimate of B under weak assumptions.

To apply the Gauss-Markov theorem the Wikipedia says you must assume your data has the following properties:

- $E[e(i)] = 0$ (lack of structural errors, needed to avoid bias)
- $V[e(i)] = c$ (equal variance, one form of homoscedasticity)
- $\text{cov}[e(i), e(j)] = 0$ for $i \neq j$ (non-correlation of errors)

Where expectation ($E[]$), variance ($V[]$), and covariance ($\text{cov}[]$)

9. Explain the gradient descent algorithm in detail.

Ans.

Gradient Descent is an optimization algorithm used for minimizing the cost function in various machine learning algorithms. It is basically used for updating the parameters of the learning model.

Types of gradient Descent:

- **Batch Gradient Descent:** This is a type of gradient descent which processes all the training examples for
 - each iteration of gradient descent. But if the number of training examples is large, then batch gradient
 - descent is computationally very expensive. Hence if the number of training examples is large, then batch
 - gradient descent is not preferred. Instead, we prefer to use stochastic gradient descent or mini-batch
 - gradient descent.
 - Let $h\theta(x)$ be the hypothesis for linear regression. Then, the cost function is given by:
 - Let Σ represents the sum of all training examples from $i=1$ to m .

- $J_{\text{train}}(\theta) = (1/2m) \sum (h_{\theta}(x(i)) - y(i))^2$
 - Repeat {
 - $\theta_j = \theta_j - (\text{learning rate}/m) * \sum (h_{\theta}(x(i)) - y(i))x_j$
 - (i)
 - For every $j = 0 \dots n$
 - }
- Where $x_j(i)$ Represents the j th feature of the i th training example. So if m is very large, then the derivative
- term fails to converge at the global minimum.
- **Stochastic Gradient Descent:** This is a type of gradient descent which processes 1 training example per
 - iteration. Hence, the parameters are being updated even after one iteration in which only a single
 - example has been processed. Hence this is quite faster than batch gradient descent. But again, when the
 - number of training examples is large, even then it processes only one example which can be additional
 - overhead for the system as the number of iterations will be quite large.
 - Randomly shuffle the data set so that the parameters can be trained evenly for each type of data.
 - As mentioned above, it takes into consideration one example per iteration.
 - Hence,
 - Let $(x(i), y(i))$ be the training example
 - $\text{Cost}(\theta, (x(i), y(i))) = (1/2) \sum (h_{\theta}(x(i)) - y(i))^2$
 - $J_{\text{train}}(\theta) = (1/m) \sum \text{Cost}(\theta, (x(i), y(i)))$
 - Repeat {
 - For $i=1$ to m {
 - $\theta_j = \theta_j - (\text{learning rate}) * \sum (h_{\theta}(x(i)) - y(i))x_j$
 - (i)
 - For every $j = 0 \dots n$
 - }
 - }

- **Mini Batch gradient descent:** This is a type of gradient descent which works faster than both batch

- gradient descent and stochastic gradient descent. Here b examples where $b < m$ are processed per

iteration. So even if the number of training examples is large, it is processed in batches of b training

examples in one go. Thus, it works for larger training examples and that too with lesser number of iterations.

Say b be the no of examples in one batch, where $b < m$. Assume $b = 10$, $m = 100$; **Note:** However we can

adjust the batch size. It is generally kept as power of 2. The reason behind it is because some hardware

such as GPUs achieve better run time with common batch sizes such as power of 2.

Repeat {

For $i=1, 11, 21, \dots, 91$

Let Σ be the summation from i to $i+9$ represented by k .

$\theta_j = \theta_j - (\text{learning rate/size of } (b)) * \Sigma (h_{\theta}(x(k)) - y(k))x_j(k)$

For every $j = 0 \dots n$

}

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

Ans.

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.

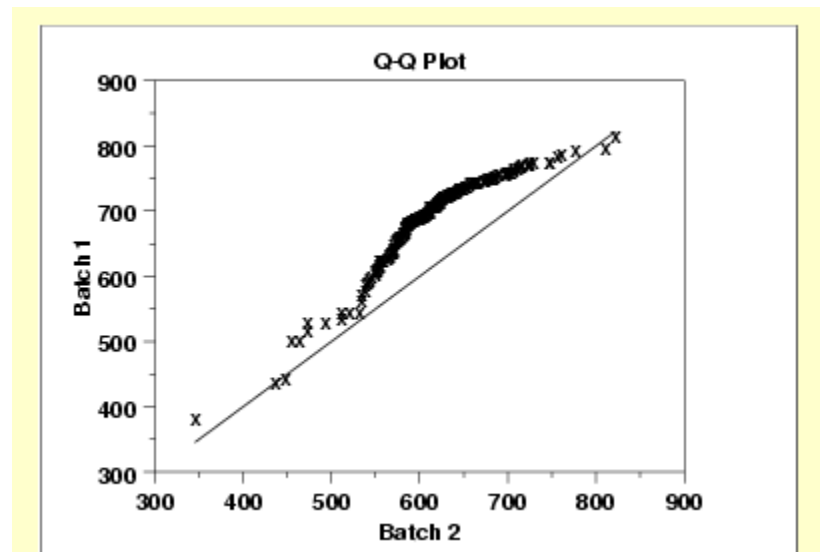
A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the

evidence for the conclusion that the two data sets have come from populations with different distributions.

The advantages of the q-q plot are:

1. The sample sizes do not need to be equal.
2. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot. For example, if the two data sets come from populations whose distributions differ only by a shift in location, the points should lie along a straight line that is displaced either up or down from the 45-degree reference line.

The q-q plot is similar to a probability plot. For a probability plot, the quantiles for one of the data samples are replaced with the quantiles of a theoretical distribution.



This q-q plot shows that

1. These 2 batches do not appear to have come from populations with a common distribution.
2. The batch 1 values are significantly higher than the corresponding batch 2 values.
3. The differences are increasing from values 525 to 625. Then the values for the 2 batches get closer again.

The q-q plot is formed by:

- Vertical axis: Estimated quantiles from data set 1
- Horizontal axis: Estimated quantiles from data set 2

Both axes are in units of their respective data sets. That is, the actual quantile level is not plotted. For a given point on the q-q plot, we know that the quantile level is the same for both points, but not what that quantile level actually is.

If the data sets have the same size, the q-q plot is essentially a plot of sorted data set 1 against sorted data set 2. If the data sets are not of equal size, the quantiles are usually picked to correspond to the sorted values from the smaller data set and then the quantiles for the larger data set are interpolated.