- **Explain the linear regression algorithm in detail.**

Linear Regression is a machine learning algorithm that is supposed to find best-fit-line relationship among any given data between independent (predictor) and dependent (target) variables.

In simple terms, it means linear regression is a method of finding the best straight-line fitting to the given data, i.e., finding the best linear relationship between the independent and dependent variables.

Linear Regression is one of the most known and well understood algorithm in statistics and machine learning.

There are two types of Linear Regression. Simple Linear Regression and Multiple Linear Regression. Simple Linear Regression Deals with one independent variable while Multiple Linear Regression deals with more than one independent variables.

- **What are the assumptions of linear regression regarding residuals?**

Following are assumptions about linear regression residuals:

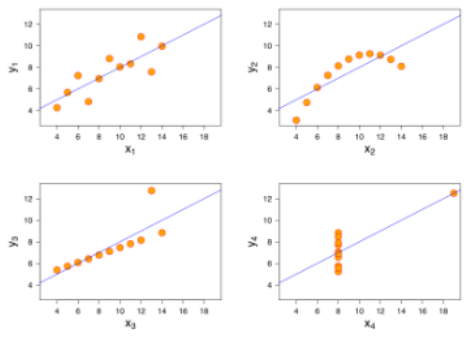
- Error terms are normally distributed
- Error terms are normally distributed around zero with mean value of zero.
- Error terms have constant variance. It is also known as homoscedasticity.
- The error terms are independent of each other, i.e., their pair-wise covariance is zero.

- **What is the coefficient of correlation and the coefficient of determination?**

Coefficient of correlation (r) shows the degree of relationship between two variables x and y. It ranges from -1 to 1. 1 means that both variables are moving up in positive direction. -1 means both variables are opposite to each other. One goes up and other goes down, in perfect negative way. 0 mean that there is no relationship between the two variables.

The coefficient of determination (R^2 or coeff) is used to assess how well a model explains and predicts future outcomes. Higher the better. It is always between 0 and 1. It can never be negative – since it is a squared value. The coefficient of determination, also commonly known as "R-squared," is used as a guideline to measure the accuracy of the model.

- **Explain the Anscombe's quartet in detail.**



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 (sum, mean, standard deviation and correlation coefficient). But things change completely when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

While the descriptive statistics of the above graphs are same but it is regression is different and fooling if not properly considered. The first one seems to be doing a decent job, the second one clearly shows that linear regression can only model linear relationships and is incapable of handling any other kind of data. The third and fourth images showcase the linear regression model's sensitivity to outliers. Had the outlier not been present, we could have got a great line fitted through the data points. So, we should never run a regression without having a good look at our data and consider the following shortcomings of linear regression:

a. It is sensitive to outliers.
b. It models linear relationships only.
c. A few assumptions are required to make the inference.

- **What is Pearson's R?**

The Pearson' R or correlation coefficient is designed for linear relationship which is a measure of the strength of the linear relationship between two variables. If the relationship between the variables is not linear, then the correlation coefficient does not adequately represent the strength of the relationship between the variables. So Pearson's R is a good measure for finding relationship in non-linear relationship. Instead we use Spearman's R.

- **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

Scaling is a method used to normalize the range of independent numeric variables or features of data.

Normalized scaling also called as min-max scaling is a method of rescaling data in the range of 0 and 1. The formula for min-max scaling is:

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

Standardization or Z-score Normalization rescales data to have a mean (μ) of 0 and standard deviation (σ) of 1 (unit variance). The formula is:

$$X_{changed} = \frac{X - \mu}{\sigma}$$

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

Variance inflation factor (VIF) is used to check the presence of multicollinearity in a data set which is calculated using the following formula.

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

When the value of VIF is “inf” or infinite it means the R^2 =1 (1/(1-1)=1/0=infinite). It happens when two intendent variables are multicollinear.

- **What is the Gauss-Markov theorem?**

The Gauss-Markov theorem states that OLS is Best Linear Unbiased Estimators (BLUE).

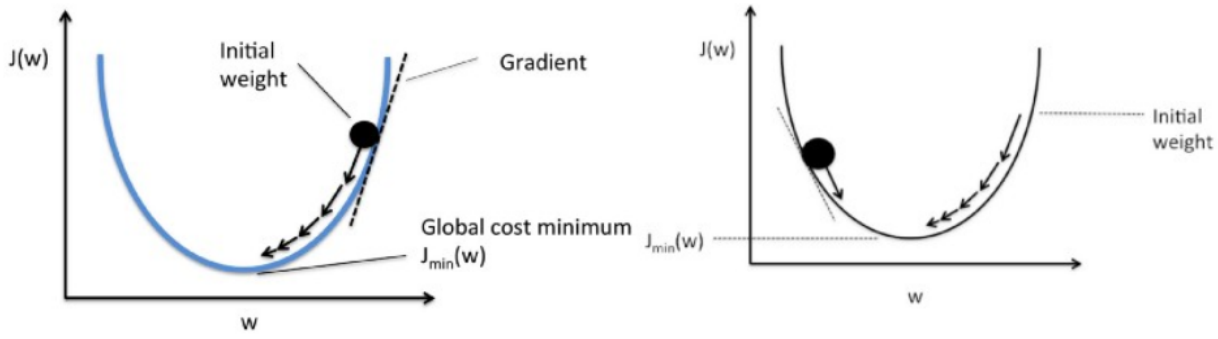
In this context, the definition of “best” refers to the minimum variance or the narrowest sampling distribution. More specifically, when your model satisfies the assumptions, OLS coefficient estimates follow the tightest possible sampling distribution of unbiased estimates compared to other linear estimation methods.

In simple terms, the Gauss-Markov theorem means that if the data generally have about the same variance at each point on x , then the "ordinary least squares" estimate of the would give you the best linear unbiased estimate of how data varies.

- **Explain the gradient descent algorithm in detail.**

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

Gradient descent is best used when the parameters cannot be calculated analytically (e.g. using linear algebra) and must be searched for by an optimization algorithm.



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:

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

The procedure starts off with initial values for the coefficient or coefficients for the function that could be 0.0 or a small random value.

coefficient = 0.0

Then the cost of the coefficients is evaluated by plugging them into the function and calculating the cost.

cost = f(coefficient) or cost = evaluate(f(coefficient))

The derivative of the cost is calculated. The derivative is a concept from calculus and refers to the slope of the function at a given point.

delta = derivative(cost)

Now that we know from the derivative which direction is downhill, we can now update the coefficient values. A learning rate parameter (alpha) must be specified that controls how much the coefficients can change on each update.

coefficient = coefficient – (alpha * delta)

This process is repeated until the cost of the coefficients (cost) is 0.0 or close enough to zero to be good enough.

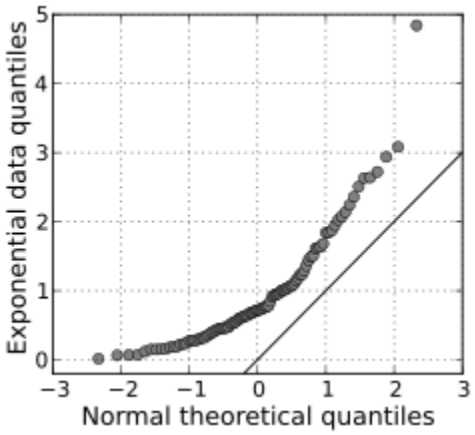
- **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**

A Q-Q (Quantile-Quantile) plots is a plot of two quantiles against each other which is used to determine if two data sets come from populations with a common distribution.

A 45-degree reference line is plotted to see the distribution. 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.

Q-Q plot is important for the following reason:

- The sample sizes do not need to be equal.
- Many distributional aspects can be simultaneously tested. 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.



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