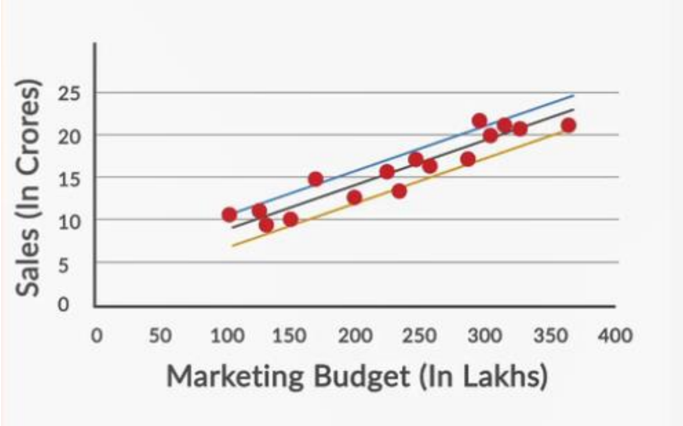
1. Explain the linear regression algorithm in detail.

In simple terms, 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.

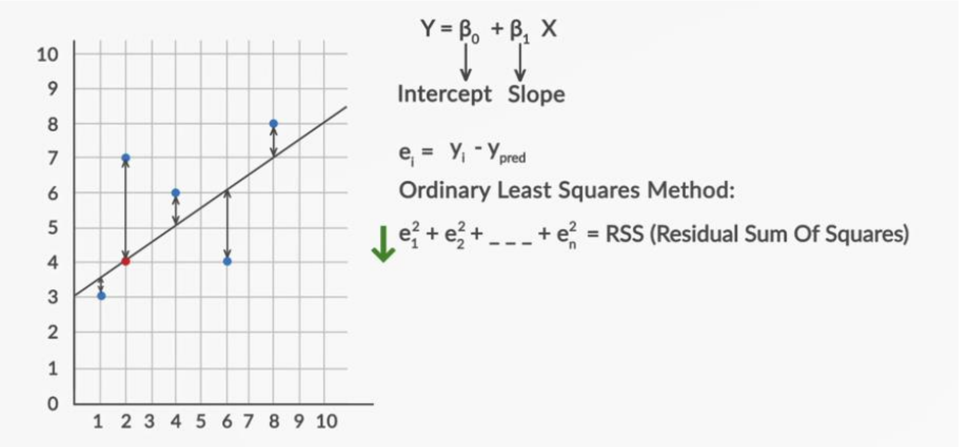
In technical terms, linear regression is a machine learning algorithm that **finds the best linear-fit relationship on any given data, between independent and dependent variables**.

As we have seen in the below graph, we can fit different lines with different m and c value.
Straight line Equation: $y = mx + c$



Our Objective is to fit the best line and reduce the error (difference between the Actual and predicted value. The best-fit line is obtained by minimizing a quantity called Residual Sum of Squares (RSS). RSS is the cost Function for this Problem Statement. We can minimize the residuals by two methods.

1. Differentiation
2. Gradient Descent


10
9
8
7
6
5
4
3
2
1
0
1 2 3 4 5 6 7 8 9 10
$Y = \beta_0 + \beta_1 X$
Intercept
Slope
$e_i = y_i - y_{pred}$
Ordinary Least Squares Method:
$e_1^2 + e_2^2 + \_\_\_ + e_n^2$ = RSS (Residual Sum Of Squares)

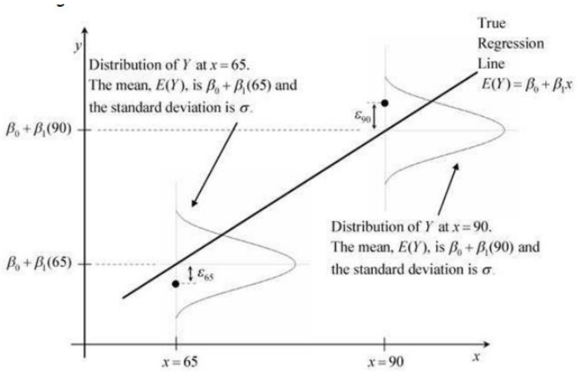
We have to make few assumptions like x and y has linear relationship and some assumptions with respect to Residuals to get the Ordinary Least Squares (OLS) estimators as the Best Linear Unbiased Estimators (BLUE).

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

Taking a more statistical view:

- Linear regression, at each X, finds the best estimate for Y but At each X, there is a distribution on the values of Y

Model predicts a single value, therefore there is a distribution of error terms at each of these values as can be seen from the figure below.



y
Distribution of Y at $x = 65$.
The mean, $E(Y)$, is $\beta_0 + \beta_1(65)$ and
the standard deviation is σ.
True
Regression
Line
$E(Y) = \beta_0 + \beta_1 x$
$\beta_0 + \beta_1(90)$
ε_{90}
$\beta_0 + \beta_1(65)$
ε_{65}
Distribution of Y at $x = 90$.
The mean, $E(Y)$, is $\beta_0 + \beta_1(90)$ and
the standard deviation is σ.
$x = 65$
$x = 90$
x

Fig 8 - Normal Distribution of Error Terms

Assumptions about the residuals:

- **Normality assumption:** It is assumed that the error terms, ε(i), are normally distributed.
  - 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.

- **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.

  $$Y^{(i)i} = \beta_0 + \beta_1 X^{(i)} + \varepsilon^{(i)}$$

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

  $$E(Y) = E(\beta_0 + \beta_1 x^{(i)} + \varepsilon^{(i)})$$
  $$= E(\beta_0 + \beta_1 x^{(i)} + \varepsilon^{(i)})$$

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

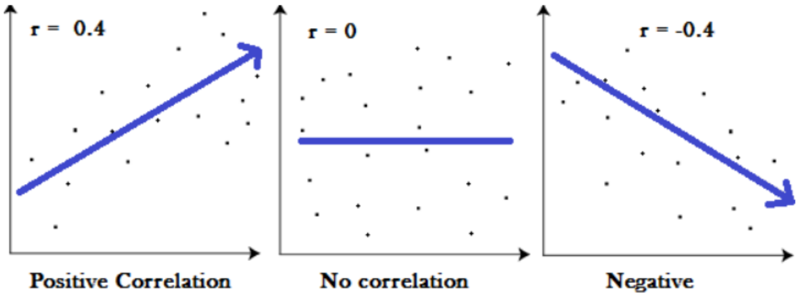
- **Constant variance assumption:** It is assumed that the residual terms have the same (but unknown) variance, σ^2. This assumption is also known as the assumption of homogeneity or homoscedasticity.
  - Generally, non-constant variance arises in presence of outliers or extreme leverage values. Look like, these values get too much weight, thereby disproportionately influences the model's performance. When this phenomenon occurs, the confidence interval for out of sample prediction tends to be unrealistically wide or narrow

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

Correlation coefficient , r :

- The quantity ***r***, called the *linear correlation coefficient*, measures the **strength and the direction of a linear relationship between two variables**. The value of r is such that **$-1 < r < +1$.**



Positive Correlation No correlation Negative

Positive Correlation: Positive r value indicates Values for x increases, values for y also increase.
Negative Correlation: Negative r value indicates Values for x increases, values for y decrease and vice versa.
No correlation: r value is zero.

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

In general term, it provides a measure of how well actual outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model, i.e. expected outcomes. Overall, the higher the R-squared, the better the model fits your data.
Mathematically, it is represented as: **$R^2 = 1 - (RSS / TSS)$**

Where RSS is the Residual Sum of Squares, TSS is Sum of errors of data from mean

- The coefficient of determination is the ratio of the **explained variation to the total variation.**
- The coefficient of determination is a measure of **how well the regression line represent 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.
- The coefficient of determination is such that **$0 < r^2 < 1$**, and denotes the strength of the linear association between x and y. 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.

There are several definitions of R^2 that are only sometimes equivalent.

Case 1: In **simple linear regression** where r^2 is used instead of R^2. When an intercept is included, then r^2 is simply the square of the sample correlation coefficient (i.e., r) between the observed outcomes and the observed predictor values.

Case 2: In Multiple Regression If additional Regressors are included, R^2 is the square of the coefficient of multiple correlation. In both such cases, the coefficient of determination normally ranges from 0 to 1.

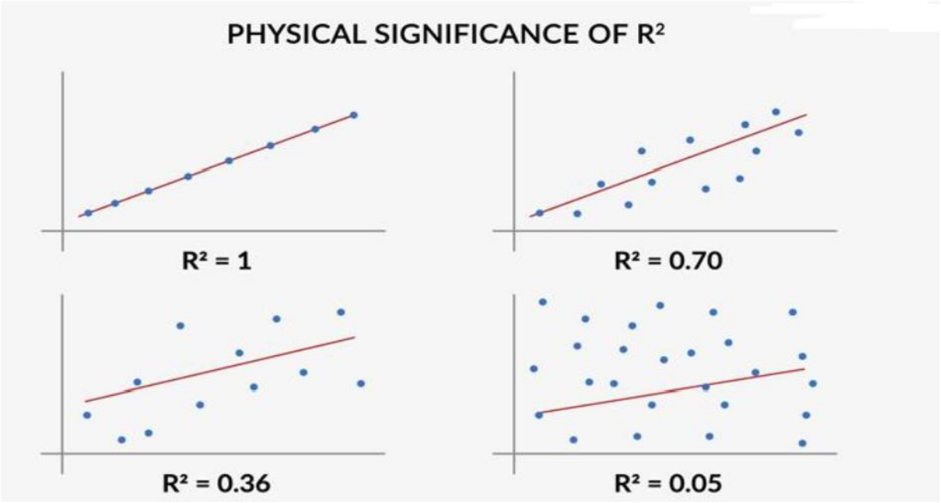
Note: Correlation can be rightfully explained for simple linear regression – because you only have one x and one y variable. For multiple linear regression R is computed, but then it is difficult to explain because we have multiple variables involved here. That's why R square is a better term.

In Graph 1: All the points lie on the line and the R^2 value is a perfect 1

In Graph 2: Some points deviate from the line and the error is represented by the lower R^2 value of 0.70

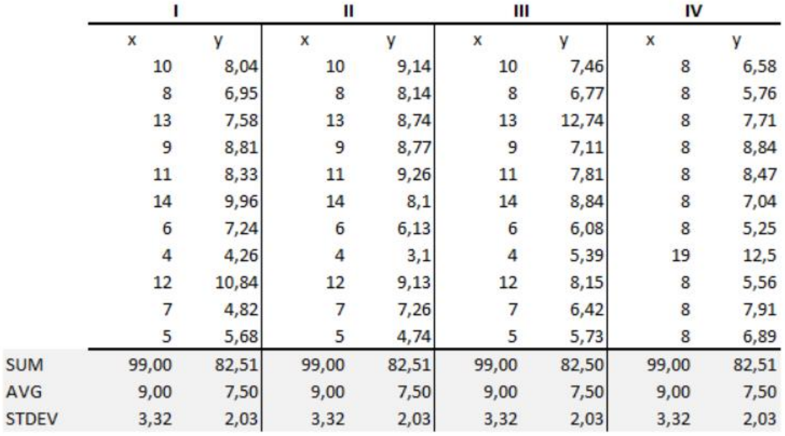
In Graph 3: The deviation further increases and the R^2 value further goes down to 0.36

In Graph 4: The deviation is further higher with a very low R^2 value of 0.05



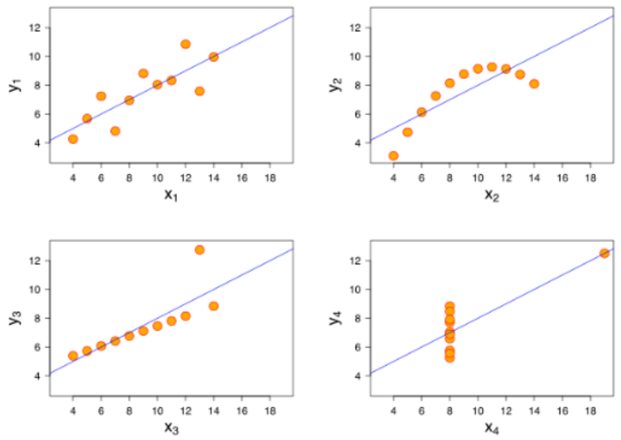
4. Explain the Anscombe’s quartet in detail.

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.



	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 first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated and following the assumption of normality.
- 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). 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.
- Anscombe’s quartet describes both the **importance of graphing data before analyzing it and the effect of outliers** and other influential observations on statistical properties.

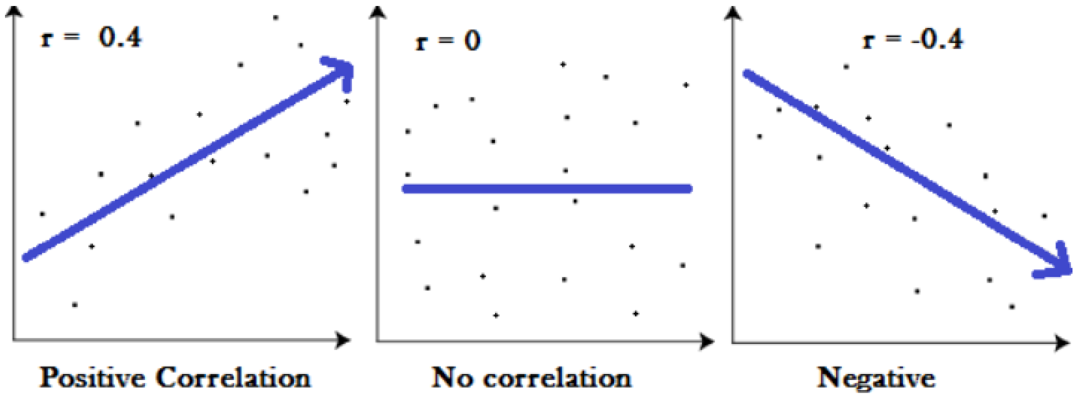
5. What is Pearson's R?

The Pearson's correlation coefficient 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.

The symbol for Pearson's correlation is "ρ" when it is measured in the population and "r" when it is measured in a sample.

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

Where *n* is the number of pairs of data.



- **A correlation coefficient of 1 (Positive Correlation)** means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other. For example, shoe sizes go up in (almost) perfect correlation with foot length.
- **A correlation coefficient of -1 (Negative Correlation) means** that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other. For example, the amount of gas in a tank decreases in (almost) perfect correlation with speed.
- **Zero (No Correlation)** means that for every increase, there isn't a positive or negative increase. The two just aren't related.
- 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.
  - A correlation greater than 0.8 is generally described as *strong*, whereas a correlation less than 0.5 is generally described as *weak*

6. 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 variables** or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step. Scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

For example, many classifiers calculate the distance between two points by the Euclidean distance. If one of the features has a **broad range of values, the distance will be governed by this particular feature**. Therefore, the range of all features should be normalized so that each feature contributes **approximately proportionately** to the final distance.

Another reason why feature scaling is applied is that **gradient descent converges much faster** with feature scaling than without it.

There are two major methods of Scaling:

- Min- Max Scaling
- Standarization

Min-Max Scaling (Normalization)

It is the simplest method and consists in rescaling the range of features to scale the range in **[0, 1].** The general formula for a min-max of [0, 1] is given as:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

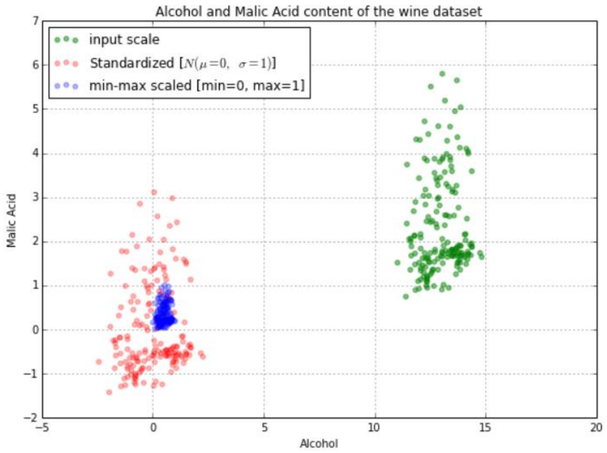
where x is an original value, x' is the normalized value.

Standardization (Z-score Normalization)

Feature standardization makes the values of each feature in the data have **zero-mean** (when subtracting the mean in the numerator) and **unit-variance**.

$$x' = \frac{x - \bar{x}}{\sigma}$$

Where x is the original feature vector, $\bar{x} = \text{average}(x)$ is the mean of that feature vector, and σ is its standard deviation.



One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers. Scaling clusters all the data very close together. It might cause algorithms such as gradient descent to take longer to converge to the same solution.

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

Multicollinearity occurs when two or more predictor variables in a multiple regression are highly correlated meaning that one can be linearly predicted from the others with a substantial degree of accuracy. You can assess multicollinearity by VIF.

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2}$$

where $R^2{}_i$ is the coefficient of determination of the regression equation

If two Independent Variables are perfectly correlated i.e., correlation is 1 and R^2 is 1 then

VIF = 1/ (1-1) = 1/0 = infinity that is the estimate is as imprecise as it can be.

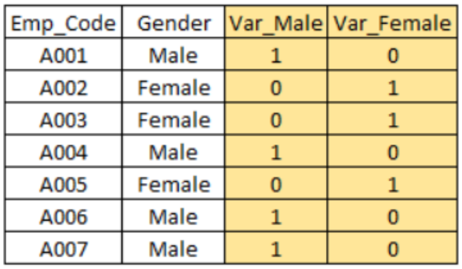
A **High VIF** indicates that the variable under consideration is almost a perfect linear combination of the independent variables already in the equation and that it should not be added to the regression equation.

Examples:

Creating dummy variables for categorical variable:

Let’s take a variable ‘gender’. We can produce two variables, namely, “Var_Male” with values 1 (Male) and 0 (No male) and “Var_Female” with values 1 (Female) and 0 (No Female). We can also create dummy variables for more than two classes of a categorical variables with **n or n-1 dummy variables**.

Two columns for Gender gives the same inference. If we are using both variable for modeling, it gives unprecise estimate. And VIF will be infinity. That’s why we have create n- 1 dummy variable.



Emp_Code	Gender	Var_Male	Var_Female
A001	Male	1	0
A002	Female	0	1
A003	Female	0	1
A004	Male	1	0
A005	Female	0	1
A006	Male	1	0
A007	Male	1	0

8. What is the Gauss-Markov theorem?

The Gauss-Markov theorem states that if your **linear regression** model satisfies the linear regression assumptions, then ordinary least squares (OLS) regression produces unbiased estimates that have the smallest variance of all possible linear estimators. The Gauss-Markov theorem states that OLS is **BLUE** (Best Linear Unbiased Estimator).

The assumptions of linear regression are:

- The assumption about the form of the model: It is assumed that there is a linear relationship between the dependent and independent variables. It is known as the 'linearity assumption'.

Assumptions about the residuals:

- Normality assumption: It is assumed that the error terms, ε(i), are normally distributed.
- 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.
- Constant variance assumption: It is assumed that the residual terms have the same (but unknown) variance, σ2 . This assumption is also known as the assumption of homogeneity or homoscedasticity.
- Independent error assumption: It is assumed that the residual terms are independent of each other, i.e. their pair-wise covariance is zero.

Assumptions about the estimators:

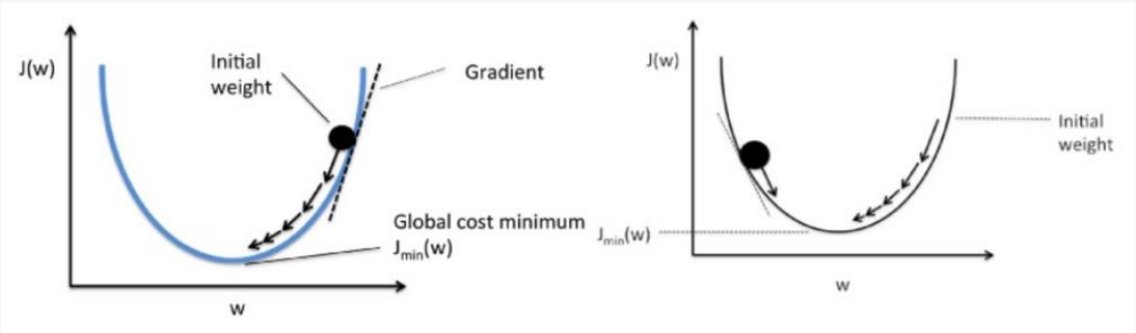
- The independent variables are measured without error.
- The independent variables are linearly independent of each other, i.e. there is no multicollinearity in the data

9. Explain the gradient descent algorithm in detail.

Gradient descent is an optimization algorithm used in Linear regression to **optimize the cost function** and find the values of the βs (estimators) corresponding to the optimized value of the cost function to get best fit line to our data.

Gradient descent is **an iterative form solution** of order one. Though gradient descent looks complicated for a 1D function, it's easier to compute the optimal minima using gradient descent for **higher dimension** 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

Gradient Descent

Mathematically, the aim of gradient descent for linear regression is to find the solution of ArgMin J(θ_0, θ_1), where J(θ_0, θ_1) is the cost function of the linear regression. It is given by —

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

Here, h is the linear hypothesis model, h= $\theta_0 + \theta_1$x,
y is the true output, and m is the number of data points 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

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

A learning rate is used as a scale factor and the coefficients are updated in the direction towards minimizing the error. When using this method, you must select a learning rate (alpha) parameter that determines the size of the improvement step to take on each iteration of the procedure.

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

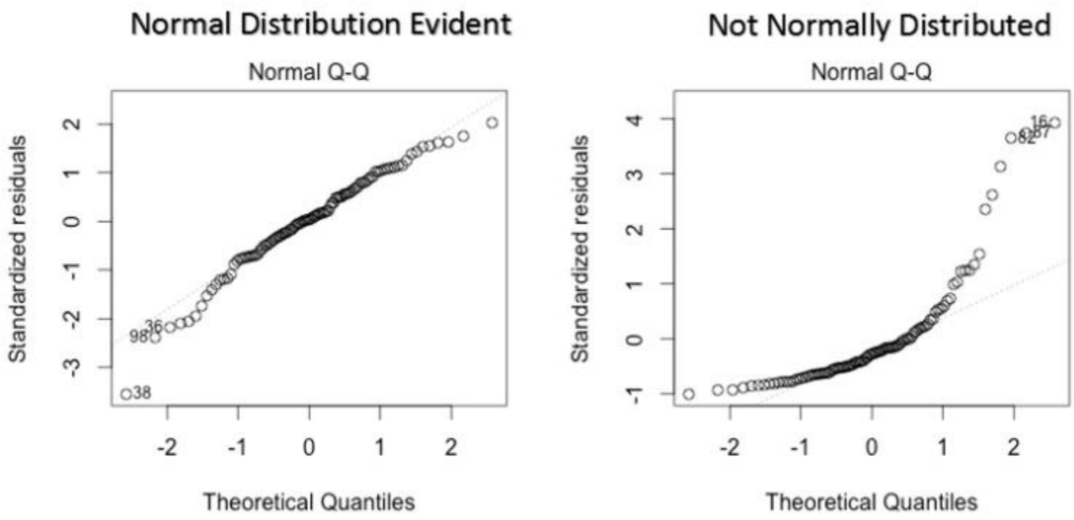
A Q-Q plot stands for a "**quantile-quantile plot**". It is a plot where the axes are **purposely transformed** in order to make a **normal (or Gaussian) distribution appear in a straight line**. In other words, a perfectly normal distribution would exactly follow a line with slope = 1 and intercept = 0.

Therefore, if the plot does not appear to be roughly a straight line, then the underlying distribution is not normal. If it bends up, then there are more "high flyer" values than expected

In Linear Regression,

The Q-Q plot is a scatter plot which helps us validate the **assumption of Linear Regression that whether the distribution of the residual is normal or not.**

They compare the distribution of Error terms to a normal distribution by plotting the quartiles of Error terms against the quartiles of a normal distribution. If Error terms are normally distributed then they should form an approximately straight line.



Solution: If the errors are not normally distributed, non – linear transformation of the variables (response or predictors) can bring improvement in the model.