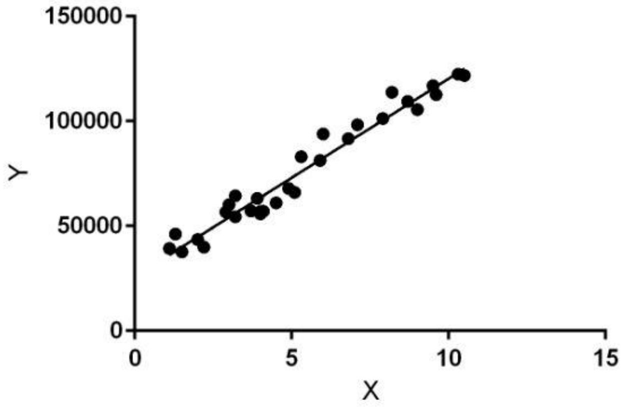
## 1. Explain the linear regression algorithm in detail

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 a machine learning algorithm that finds the best linear-fit relationship on any given data, between independent and dependent variables. It is mostly done by the Sum of Squared Residuals Method.



Hypothesis of Linear Regression is y= mx+C

Where

Y=is the dependent variable

m – is the coefficient of X

x=Independent Variable from which is also known as predictor variable.

C- is the intercept on y axis

If C=0 Then the line goes from the origin

Once we find the best m and C 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.

**Cost Function :**
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.

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

**<u>Gradient Descent</u>:**
To update m and C values in order to reduce Cost function (minimizing RMSE value) and achieving the best fit line the model uses Gradient Descent. It iteratively reduces the value reaching the minimum value.

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

Assumptions about the residuals:

1. Normality assumption: It is assumed that the error terms, $\varepsilon^{(i)}$, 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, $\sigma^2$ . 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.

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

**Correlation Correlation-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 $-1 \leq r \leq +1$. The + and – signs are used for positive linear correlations and negative linear correlations, respectively.

1. **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.
2. **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.
3. **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

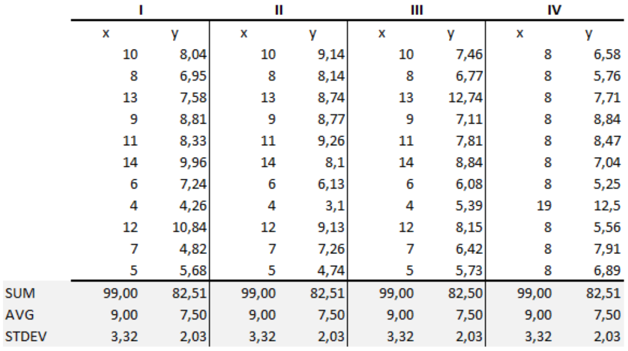
**Coefficient of determination – r^2 or R^2**

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. It is a measure that allows us to determine how certain one can be in making predictions from a certain model/graph.
2. The coefficient of determination is the ratio of the explained variation to the total variation.
3. The coefficient of determination is such that 0 < r 2 < 1, and denotes the strength of the linear association between x and y.

4. The coefficient of determination represents the percent of the data that is the closest to the line of best fit.

## 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. It was 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

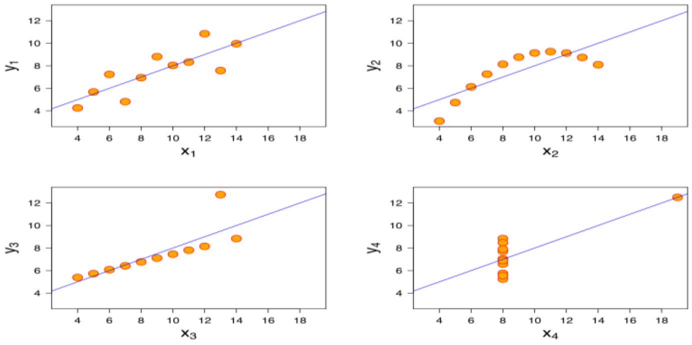


| | I | | II | | III | | IV | |
|---|---|---|---|---|---|---|---|---|
| | x | y | x | y | x | y | x | y |
| | 10 | 8,04 | 10 | 9,14 | 10 | 7,46 | 8 | 6,58 |
| | 8 | 6,95 | 8 | 8,14 | 8 | 6,77 | 8 | 5,76 |
| | 13 | 7,58 | 13 | 8,74 | 13 | 12,74 | 8 | 7,71 |
| | 9 | 8,81 | 9 | 8,77 | 9 | 7,11 | 8 | 8,84 |
| | 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 Iappears 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?

Pearson's R correlation coefficient which is the correlation coefficient . Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

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

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values

**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

S**tandardization** 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:

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

$\mu$ is the mean of the feature values and $\sigma$ is the standard deviation of the feature values. Note that in this case, the values are not restricted to a particular range.

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

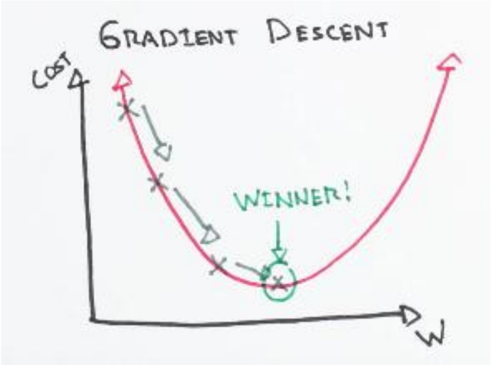
A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.
VIF=1\[1-R^2]
If there is perfect correlation, then **VIF** = **infinity**
VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity

## 8. What is the Gauss-Markov theorem?

The Gauss-Markov theorem states that if your linear regression model satisfies the first six classical assumptions, then ordinary least squares (OLS) regression produces unbiased estimates that have the smallest variance of all possible linear estimators.
According to this theorem the OLS procedure produces unbiased estimates that have the minimum variance. The sampling distributions are centered on the actual population value and are the tightest possible distributions.

## 9. Explain the gradient descent algorithm in detail.

Gradient descent is an optimization algorithm used to find the values of parameters (coefficients) of a function (f) that minimizes a cost function (cost).It is minimized by iteratively moving in the direction of steepest descent as defined by the negative of the gradient. In machine learning, we use gradient descent to update the parameters of our model.
Equation of line is y(p)=β0+β1x
Where β0 is the intercept of the fitted line and β1 is the coefficient for the independent variable x.
To get the optimum betas we need to reduce cost function for all data points.
Which is given as
J(θ0,θ1)=∑Ni=1(yi−yi(p))2
So by using the Gradient Descent optimization algorithm we will move towards the negative of gradient .



To Compute θ1, we saw that the equation will look like this,

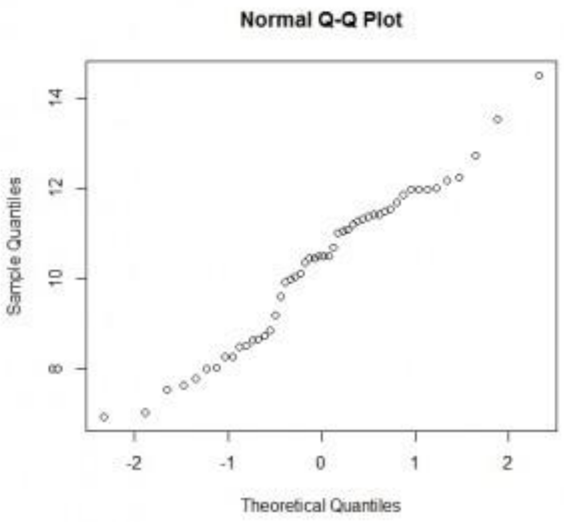
θ1=θ0 −η∂/∂θ*J(θ)

Where η is known as the learning rate, which defines the speed at which we want to move towards negative of the gradient.

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

The Q-Q plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential. For example, if we run a statistical analysis that assumes our dependent variable is Normally distributed, we can use a Normal Q-Q plot to check that assumption.

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. Here's an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles