1. Explain the linear regression algorithm in detail.

   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.

   The linear equation assigns one scale factor to each input value or column, called a coefficient and represented by the capital Greek letter Beta (B). One additional coefficient is also added, giving the line an additional degree of freedom (e.g. moving up and down on a two-dimensional plot) and is often called the intercept or the bias coefficient.

   For example, in a simple regression problem (a single x and a single y), the form of the model would be:
   y = B0 + B1*x
   In higher dimensions when we have more than one input (x), the line is called a plane or a hyper-plane. The representation therefore is the form of the equation and the specific values used for the coefficients (e.g. B0 and B1 in the above example).

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

ANS:
A scatter plot of residual values vs predicted values is a goodway to check forhomoscedasticity. There should be no clear pattern in the distribution and if there is a specific pattern,the data is heteroscedastic.

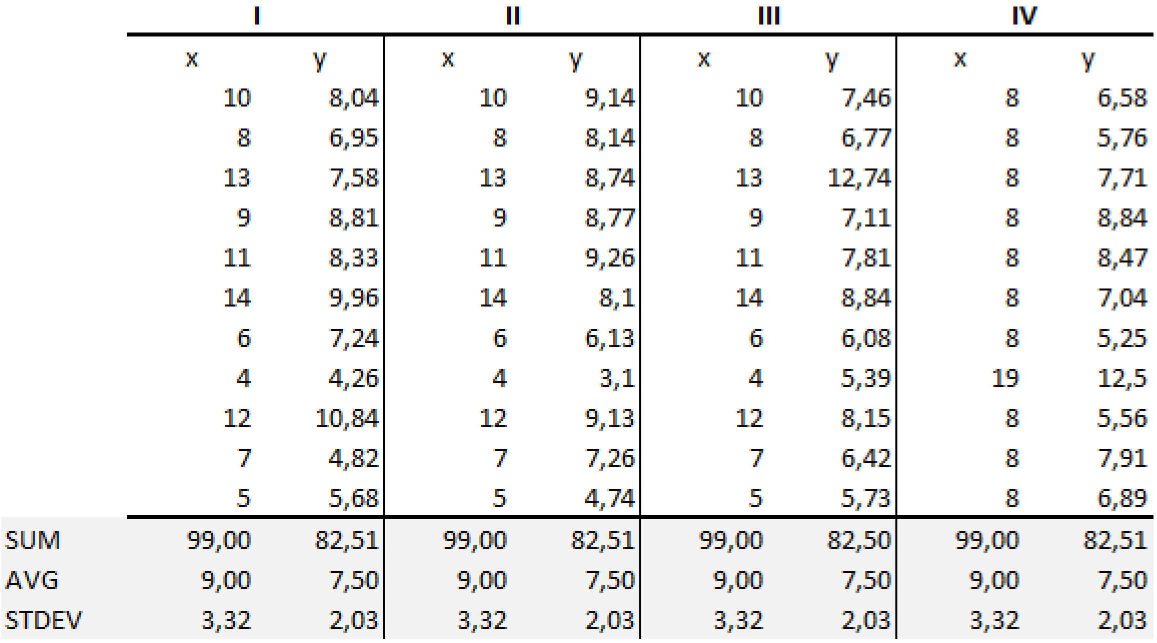
3. What is the coefficient of correlation and the coefficient of determination?

Coefficient of correlation is “R” value which is given in the summary table in the Regression output. R square is also called coefficient of determination. Multiply R times R to get the R square value. In other words Coefficient of Determination is the square of Coefficeint of Correlation.
R square or coeff. of determination shows percentage variation in y which is explained by all the x variables together. Higher the better. It is always between 0 and 1. It can never be negative – since it is a squared value.

Coefficient of Correlation: is the degree of relationship between two variables say x and y. It can go between -1 and 1. 1 indicates that the two variables are moving in unison. They rise and fall together and have perfect correlation. -1 means that the two variables are in perfect opposites. One goes up and other goes down, in perfect negative way. Any two variables in this universe can be argued to have a correlation value. If they are not correlated then the correlation value can still be computed which would be 0. The correlation value always lies between -1 and 1 (going thru 0 – which means no correlation at all – perfectly not related). Correlation can be rightfully explalined 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 invovled here. Thats why R square is a better term.

4. 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. 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

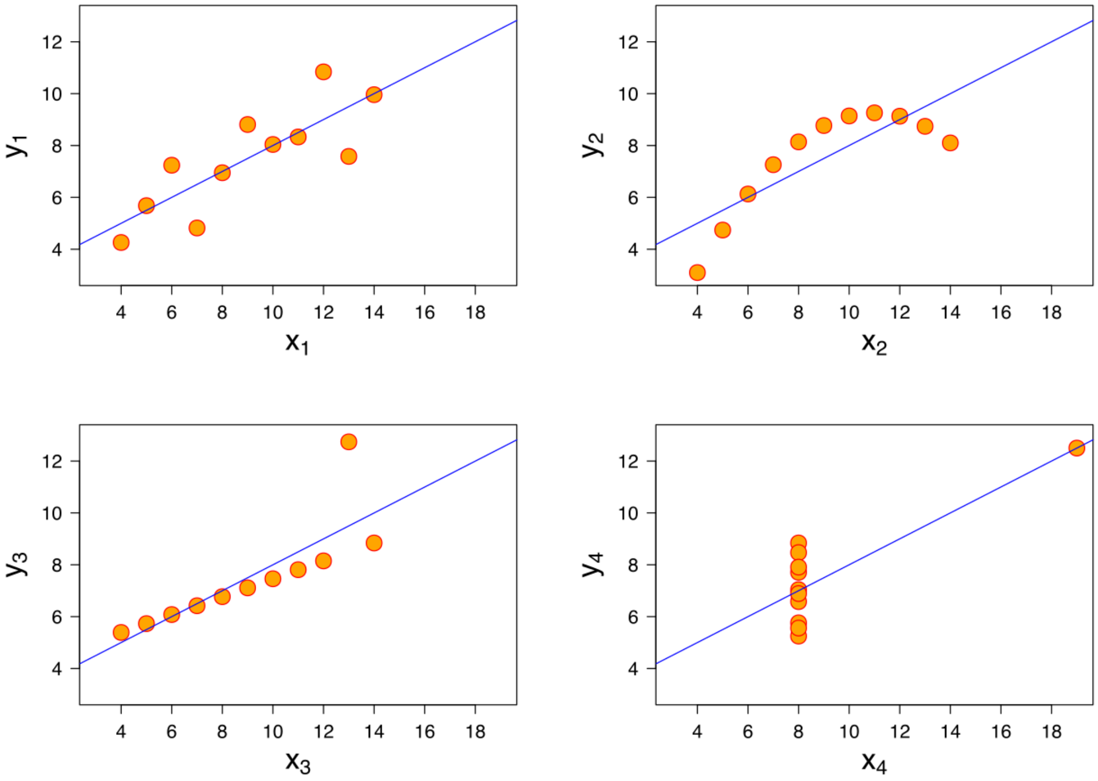
The summary statistics show that the means and the variances were identical for x and y across the groups :

•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 appears 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 correlation coefficient is the test statistics that measures the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship.

The Pearson's correlation coefficient varies between -1 and +1 where:
$r = 1$ means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
$r = -1$ means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
$r = 0$ means there is no linear association
$r > 0 < 5$ means there is a weak association
$r > 5 < 8$ means there is a moderate association
$r > 8$ means there is a strong association

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

Feature 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 preprocessing while using machine learning algorithms.

Scaling
In scaling (also called min-max scaling), you transform the data such that the features are within a specific range e.g. [0, 1].
$x' = x - xmin/xmax - xmin$
where x' is the normalized value.

Scaling is important in the algorithms such as support vector machines (SVM) and k-nearest neighbors (KNN) where distance between the data points is important. For example, in the dataset containing prices of products; without scaling, SVM might treat 1 USD equivalent to 1 INR though 1 USD = 65 INR.

The point of normalization is to change your observations so that they can be described as a normal distribution.
Normal distribution (Gaussian distribution), also known as the bell curve, is a specific statistical distribution where a roughly equal observations fall above and below the mean, the mean and the median are the same, and there are more observations closer to the mean.
Standardization (also called z-score normalization) transforms your data such that the resulting distribution has a mean of 0 and a standard deviation of 1. It's the definition that we read in the last paragraph.
$x'=x-xmean/\sigma$
where x is the original feature vector, xmean is the mean of that feature vector, and σ is its standard deviation.
The z-score comes from statistics, defined as
$z=x-\mu/\sigma$

Simply called normalization, it's just another way of normalizing data. Note that, it's a different from min-max scaling in numerator, and from z-score normalization in the denominator.
$x'=x-xmeanxmax-xmin$
For normalization, the maximum value you can get after applying the formula is 1, and the minimum value is 0. So all the values will be between 0 and 1.

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

   The user has to select the variables to be included by ticking off the corresponding check boxes. ... 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).

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. The proof for this theorem goes way beyond the scope of this blog post. However, the critical point is that when you satisfy the classical assumptions, you can be confident that you are obtaining the best possible coefficient estimates. The Gauss-Markov theorem does not state that these are just the best possible estimates for the OLS procedure, but the best possible estimates for any linear model estimator. Think about that!

In my post about the classical assumptions of OLS linear regression, I explain those assumptions and how to verify them. In this post, I take a closer look at the nature of OLS estimates. What does the Gauss-Markov theorem mean exactly when it states that OLS estimates are the best estimates when the assumptions hold true?

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.

Regression analysis is like any other inferential methodology. Our goal is to draw a random sample from a population and use it to estimate the properties of that population. In regression analysis, the coefficients in the equation are estimates of the actual population parameters.

The notation for the model of a population is the following:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

The betas (β) represent the population parameter for each term in the model. Epsilon (ε) represents the random error that the model doesn’t explain. Unfortunately, we’ll never know these population values because it is generally impossible to measure the entire population. Instead, we’ll obtain estimates of them using our random sample.

The notation for an estimated model from a random sample is the following:

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_k X_k + e$$

The hats over the betas indicate that these are parameter estimates while e represents the residuals, which are estimates of the random error.

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).
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.
The goal is to continue to try different values for the coefficients, evaluate their cost and select new coefficients that have a slightly better (lower) cost.
Repeating this process enough times will lead to the bottom of the bowl and you will know the values of the coefficients that result in the minimum cost.
Gradient Descent Procedure
The procedure starts off with initial values for the coefficient or coefficients for the function. These could be 0.0 or a small random value.
coefficient = 0.0
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. We need to know the slope so that we know the direction (sign) to move the coefficient values in order to get a lower cost on the next iteration.
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.

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. It’s just a visual check, not an air-tight proof, so it is somewhat subjective. But it allows us to see at-a-glance if our assumption is plausible, and if not, how the assumption is violated and what data points contribute to the violation.

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.

Now what are “quantiles”? These are often referred to as “percentiles”. These are points in your data below which a certain proportion of your data fall. For example, imagine the classic bell-curve standard Normal distribution with a mean of 0. The 0.5 quantile, or 50th percentile, is 0. Half the data lie below 0. That’s the peak of the hump in the curve. The 0.95 quantile, or 95th percentile, is about 1.64. 95 percent of the data lie below 1.64. The following R code generates the quantiles for a standard Normal distribution from 0.01 to 0.99 by increments of 0.01:

```
qnorm(seq(0.01,0.99,0.01))
```

We can also randomly generate data from a standard Normal distribution and then find the quantiles. Here we generate a sample of size 200 and find the quantiles for 0.01 to 0.99 using the quantile function:

```
quantile(rnorm(200),probs = seq(0.01,0.99,0.01))
```

So we see that quantiles are basically just your data sorted in ascending order, with various data points labelled as being the point below which a certain proportion of the data fall. However it’s worth noting there are many ways to calculate quantiles. In fact, the quantile function in R offers 9 different quantile algorithms! See help(quantile)for more information.

Q-Q plots take your sample data, sort it in ascending order, and then plot them versus quantiles calculated from a theoretical distribution. The number of quantiles is selected to match the size of your sample data. While Normal Q-Q Plots are the ones most often used in practice due to so many statistical methods assuming normality, Q-Q Plots can actually be created for any distribution.