1. Explain the linear regression algorithm in detail.
2. What are the assumptions of linear regression regarding residuals?
3. What is the coefficient of correlation and the coefficient of determination?
4. Explain the Anscombe’s quartet in detail.
5. What is Pearson’s R?
6. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
7. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
8. What is the Gauss-Markov theorem?
9. Explain the gradient descent algorithm in detail.
10. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans 1: Linear regression is one of the most commonly used ML algorithms for predictive analysis.

The overall idea of regression is to examine two things:

(1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable?

(2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable?

These regression estimates are in turn used to explain the relationship between one dependent variable with one or more independent variables.

The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = b0 + b1*x, where y = estimated dependent variable score, b0 = constant, b1 = regression coefficient, and x = score on the independent variable.

Three major uses for regression analysis are :

(1) determining the strength of predictors,

(2) forecasting an effect, and

(3) trend forecasting.

First, the regression might be used to identify the strength of the effect that the independent variable(s) have on a dependent variable. Typical questions are what is the strength of relationship between dose and effect, sales and marketing spending, or age and income.

Second, it can be used to forecast effects or impact of changes. That is, the regression analysis helps us to understand how much the dependent variable changes with a change in one or more independent variables. A typical question is, “how much additional sales income do I get for each additional $1000 spent on marketing?”

Third, regression analysis predicts trends and future values. The regression analysis can be used to get point estimates. A typical question is, “what will the price of gold be in 6 months?”

There are several types of linear regression analyses available to researchers.

- Simple linear regression
  1 dependent variable (interval or ratio), 1 independent variable (interval or ratio or dichotomous)

- Multiple linear regression
  1 dependent variable (interval or ratio) , 2+ independent variables (interval or ratio or dichotomous)

- Logistic regression
  1 dependent variable (dichotomous), 2+ independent variable(s) (interval or ratio or dichotomous)

- Ordinal regression
  1 dependent variable (ordinal), 1+ independent variable(s) (nominal or dichotomous)

- Multinominal regression
  1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio or dichotomous)

- Discriminant analysis
  1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio)

When selecting the model for the analysis, an important consideration is model fitting. Adding independent variables to a linear regression model will always increase the explained variance of the model (typically expressed as R^2). However, overfitting can occur by adding too many variables to the model, which reduces model generalizability.

Ans 2:

1. There should be a linear and additive relationship between dependent (response) variable and independent (predictor) variable(s). A linear relationship suggests that a change in response Y due to one unit change in X^1 is constant, regardless of the value of X^1. An additive relationship suggests that the effect of X^1 on Y is independent of other variables.

2. There should be no correlation between the residual (error) terms. Absence of this phenomenon is known as Autocorrelation.

3. The independent variables should not be correlated. Absence of this phenomenon is known as multicollinearity.

4. The error terms must have constant variance. This phenomenon is known as homoskedasticity. The presence of non-constant variance is referred to heteroskedasticity.

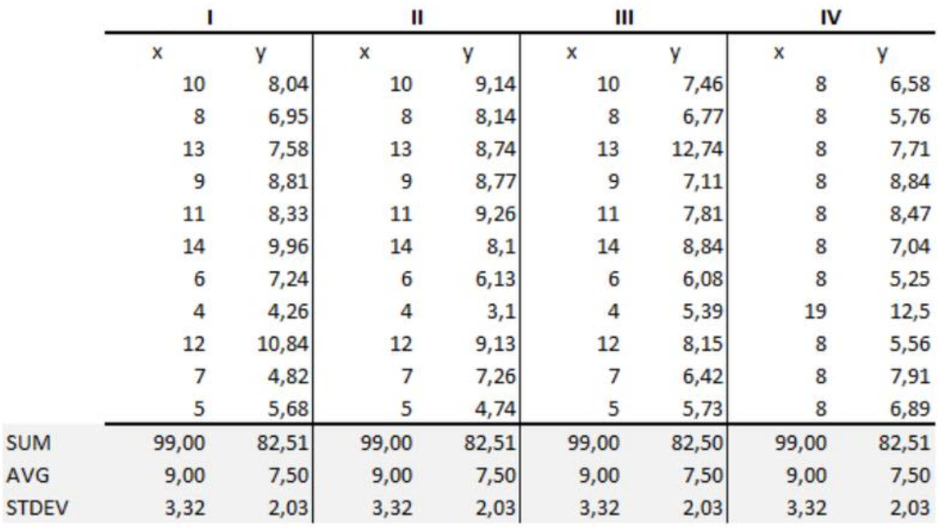
5. The error terms must be normally distributed.

Ans 3:

Coefficient of correlation: It 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 through 0 – which means no correlation at all – perfectly not related).

Coefficient of determination: It is useful because it gives the proportion of the variance 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/paragraph. It is the ratio of the explained variation to the total variation. It mainly represents the percentage of the data that is closest to the line of the best fit. If the regression line passes exactly throguh 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.

Ans 4: Asncombe's Quartet was developed by Statistician Francis Anscombe. It mainly comprises of 4 fdatasets, each containing 11 (x,y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change **Completely,** when they are plot as graphs. 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

Quartet's Summary Stats

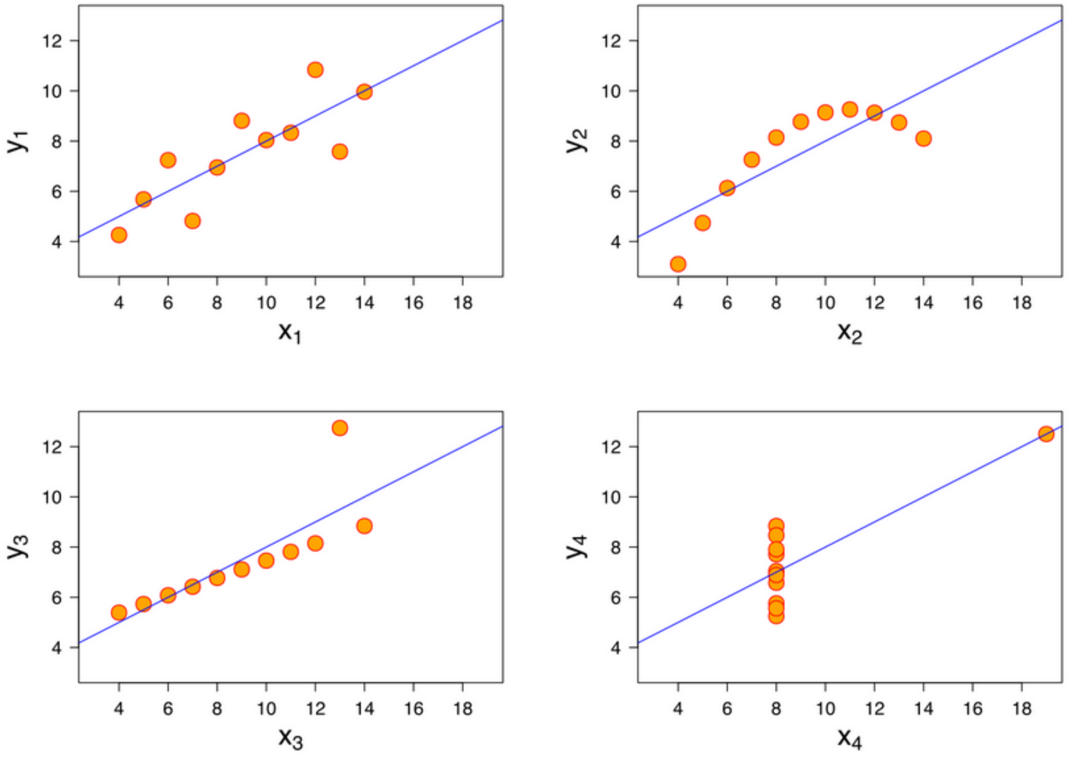
summary statistics show that the means and variances were identical for x and y across the groups:

1. Mean of x is 9 and the mean of y is 7.50 for each dataset.

2. Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset.

3. The correlation coefficient between x and y for each dataset is 0.816.

When we plot these 4 datasets on an x/y coordinate plane, we can notice that they show similar regression lines as well but eacxh dataset depicts a different story:



Ans 5: Pearson's R: It 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.

In a sample it is denoted by r and is by design constrained as follows:

$$-1 \leq r \leq 1$$

Furthermore,

1. Positive values denote positive linear correlation.

2. Negative values denote negative linear correlation.

3. 0 denotes no linear correlation.

4. The closer the value is to 1 or -1, the stronger the linear correlation.

It Is given by,

$$\rho_{X,Y} = \frac{\mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y]}{\sqrt{\mathrm{E}[X^2] - [\mathrm{E}[X]]^2}\,\sqrt{\mathrm{E}[Y^2] - [\mathrm{E}[Y]]^2}}.$$

Ans 6: Scaling a step of Data Pre Processing which is applied to independent variables or features of data. It basically helps to normalise **the** data within a particular range. Sometimes, it also helps in speeding up **the** calculations in an algorithm.

It is mainly done beause Most of the times, your dataset will contain features highly varying in magnitudes, units and range. But since, most of the machine learning algorithms use Eucledian distance between two data points in their computations, this is a problem.If left alone, these algorithms only take in the magnitude of features neglecting the units. The results would vary greatly between different units, 5kg and 5000gms. The features with high magnitudes will weigh in a lot more in the distance calculations than features with low magnitudes. To supress this effect, we need to bring all features to the same level of magnitudes. This can be achieved by scaling.

1. Standardized Scaling :

Standardisation replaces the values by their Z scores This redistributes the features with their mean **μ = 0** and standard deviation **σ =1** `.sklearn.preprocessing.scale` helps us implementing standardisation in python.

2. Normalized Scaling:

This distribution will have values between -1 and 1with $\mu=0$.Standardisation and Mean Normalization can be used for algorithms that assumes zero centric data like Principal Component Analysis(PCA).

Ans 7: If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect correlation, then VIF = infinity. 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. This would mean that that standard error of this coefficient is inflated by a factor of 2 (square root of variance is the standard deviation). The standard error of the coefficient determines the confidence interval of the model coefficients. If the standard error is large, then the confidence intervals may be large, and the model coefficient may come out to be non-significant due to the presence of multicollinearity. A general rule of thumb is that if VIF > 10 then there is multicollinearity. Note that this is a rough rule of thumb, in some cases we might choose to live with high VIF values if it does not affect our model results such as when we are fitting a quadratic or cubic model or depending on the sample size a large value of VIF may not necessarily indicate a poor model.

Ans 8: The **Gauss Markov theorem** tells us that if a certain set of assumptions are met, the ordinary least squares estimate for regression coefficients gives you the *best linear unbiased estimate (BLUE)* possible.

There are five Gauss Markov assumptions (also called *conditions*):

1. **Linearity**: the parameters we are estimating using the OLS method must be themselves linear.
2. **Random**: our data must have been randomly sampled from the population.
3. **Non-Collinearity**: the regressors being calculated aren’t perfectly correlated with each other.
4. **Exogeneity**: the regressors aren’t correlated with the error term.
5. **Homoscedasticity**: no matter what the values of our regressors might be, the error of the variance is constant.

We can summarize the Gauss-Markov Assumptions succinctly in algebra, by saying that a linear regression model represented by

$y_i = x_i' \beta + \varepsilon_i$

and generated by the ordinary least squares estimate is the best linear unbiased estimate (BLUE) possible if

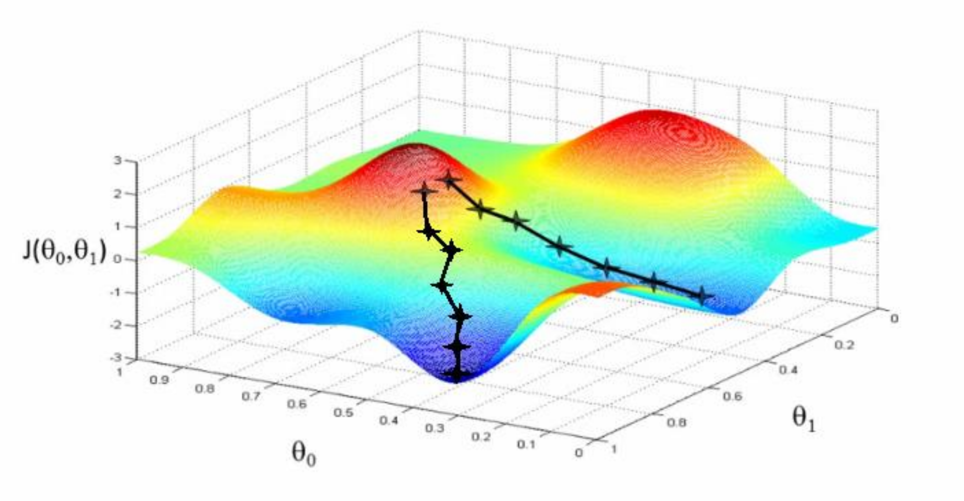
- $E\{\varepsilon_i\} = 0, i = 1, . . . , N$
- $\{\varepsilon_1 …… \varepsilon_n\}$ and $\{x_1 ….., x_N\}$ are independent
- $\text{cov}\{\varepsilon_i, \varepsilon_j\} = 0, i, j = 1,…., N \; I \neq j.$
- $V\{\varepsilon_1 = \sigma^2, i= 1, ….N$

The first of these assumptions can be read as “The expected value of the error term is zero.”. The second assumption is collinearity, the third is exogeneity, and the fourth is homoscedasticity.

Ans 9: To explain Gradient Descent I’ll use the classic mountaineering example.

Suppose you are at the top of a mountain, and you have to reach a lake which is at the lowest point of the mountain (a.k.a valley). A twist is that you are blindfolded and you have zero visibility to see where you are headed. So, what approach will you take to reach the lake? The best way is to check the ground near you and observe where the land tends to descend. This will give an idea in what direction you should take your first step. If you follow the descending path, it is very likely you would reach the lake.

To represent this graphically, notice the below graph.



Let us now map this scenario in mathematical terms.

Suppose we want to find out the best parameters (θ1) and (θ2) for our learning algorithm. Similar to the analogy above, we see we find similar mountains and valleys when we plot our “cost space”. Cost space is nothing but how our algorithm would perform when we choose a particular value for a parameter.

So on the y-axis, we have the cost J(θ) against our parameters θ1 and θ2 on x-axis and z-axis respectively. Here, hills are represented by red region, which have high cost, and valleys are represented by blue region, which have low cost.

Now there are many types of gradient descent algorithms. They can be classified by two methods mainly:

- **On the basis of data ingestion**
    1. Full Batch Gradient Descent Algorithm
    2. Stochastic Gradient Descent Algorithm

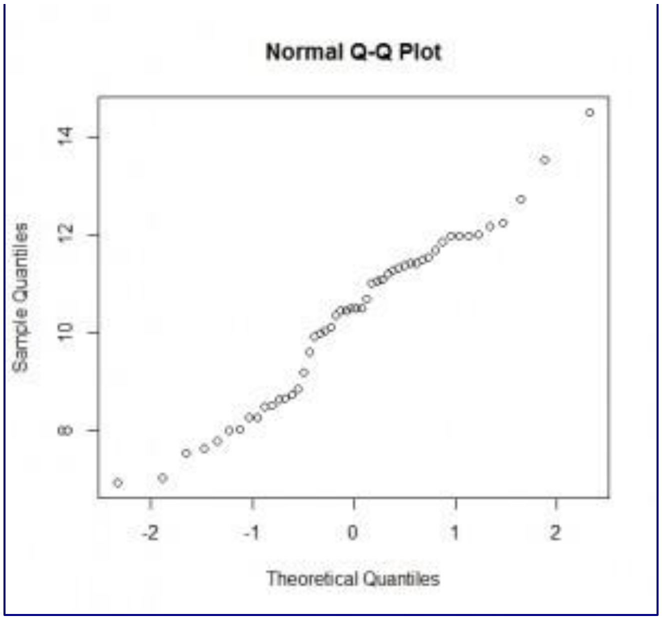
In full batch gradient descent algorithms, you use whole data at once to compute the gradient, whereas in stochastic you take a sample while computing the gradient.

On the basis of differentiation techniques

1. First order Differentiation
2. Second order Differentiation

Gradient descent requires calculation of gradient by differentiation of cost function. We can either use first order differentiation or second order differentiation.

Ans 10: 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.



QQ Plots in Linear Regression: QQ-plots are ubiquitous in statistics. Most people use them in a single, simple way: fit a linear regression model, check if the points lie approximately on the line, and if they don't, your residuals aren't Gaussian and thus your errors aren't either.