1. Explain the linear regression algorithm in detail.

Regression: Regression analysis is a form of prediction modelling techniques which investigates relationship between dependent and independent variables
Major uses of regression analysis are

- Determining strength of predictors
- Forecasting an effect and
- Trend forecasting

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. Furthermore, the linear relationship can be positive or negative in nature as explained below

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.

Hypothesis function for Linear Regression:

$$Y = \beta_0 + \beta_1 X$$

While training the model we are given:

x: input training data (univariate – one input variable(parameter))
y: labels to data (supervised learning)
When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best $\theta1$ and $\theta2$ values.
β_0: intercept
β_1: coefficient of x
Once we find the best β_0 and β_1 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.

Updating β_0 and β_1 **values to get the best fit line:**

Cost Function (J):
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. So, it is very important to update the $\theta1$ and $\theta2$ values, to reach the best value that minimize the error between predicted y value (predicted) and true y value (y).

$$minimize\frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

Cost function(J) of Linear Regression is the **Root Mean Squared Error (RMSE)** between predicted y value (predicted) and true y value (y).
Gradient Descent**:**
To update β_0 and β_1 values in order to reduce Cost function (minimizing RMSE value) and achieving the best fit line the model uses Gradient Descent. The idea is to start with random β_0 and β_1 values and then iteratively updating the values, reaching minimum cost.

Linear regression models can be classified into two types depending upon the number of independent variables:

- Simple linear regression: This is used when the number of independent variables is 1.
- Multiple linear regression: This is used when the number of independent variables is more than 1.

The strength of a linear regression model is mainly explained by R^2, where $R^2 = 1 - (RSS/TSS)$.

- RSS: Residual sum of squares
- TSS: Total sum of square

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

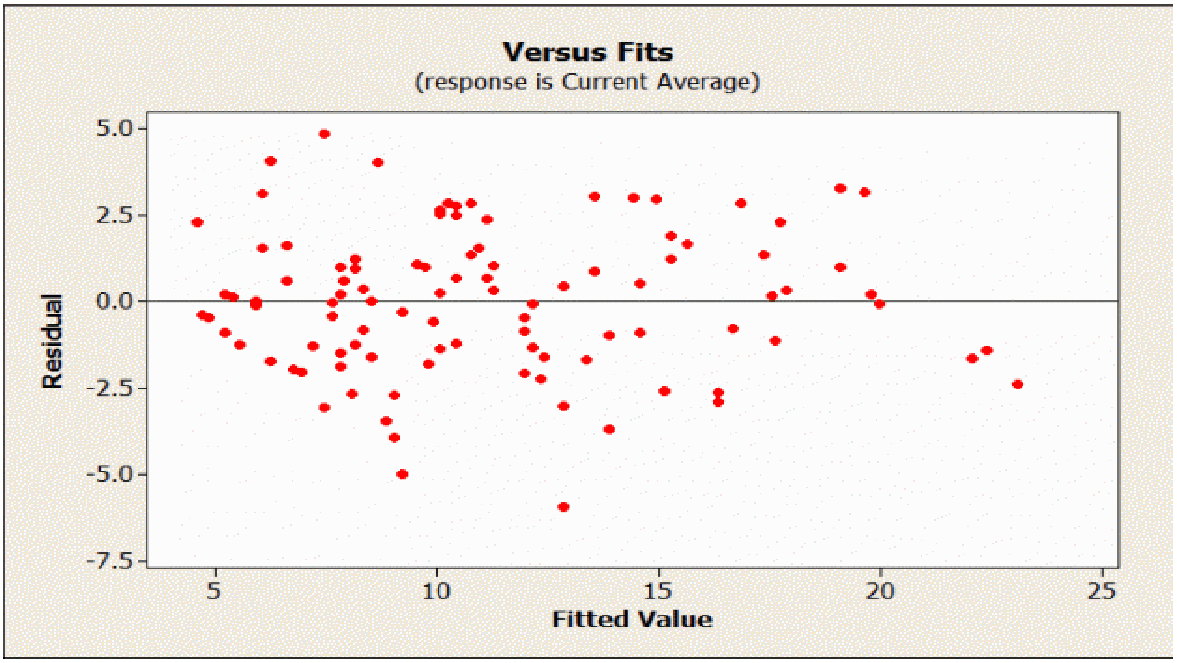
We make a few assumptions when we use linear regression to model the relationship between a response and a predictor. These assumptions are essentially conditions that should be met before we draw inferences regarding the model estimates or before we use a model to make prediction.

- The true relationship is linear
- Errors are normally distributed
- Homoscedasticity of errors (or, equal variance around the line).
- Independence of the observations

To check regression assumptions, we examine the variability left over after we fit the regression line. We simply graph the residuals and look for any unusual patterns. If a linear model makes sense, the residuals will

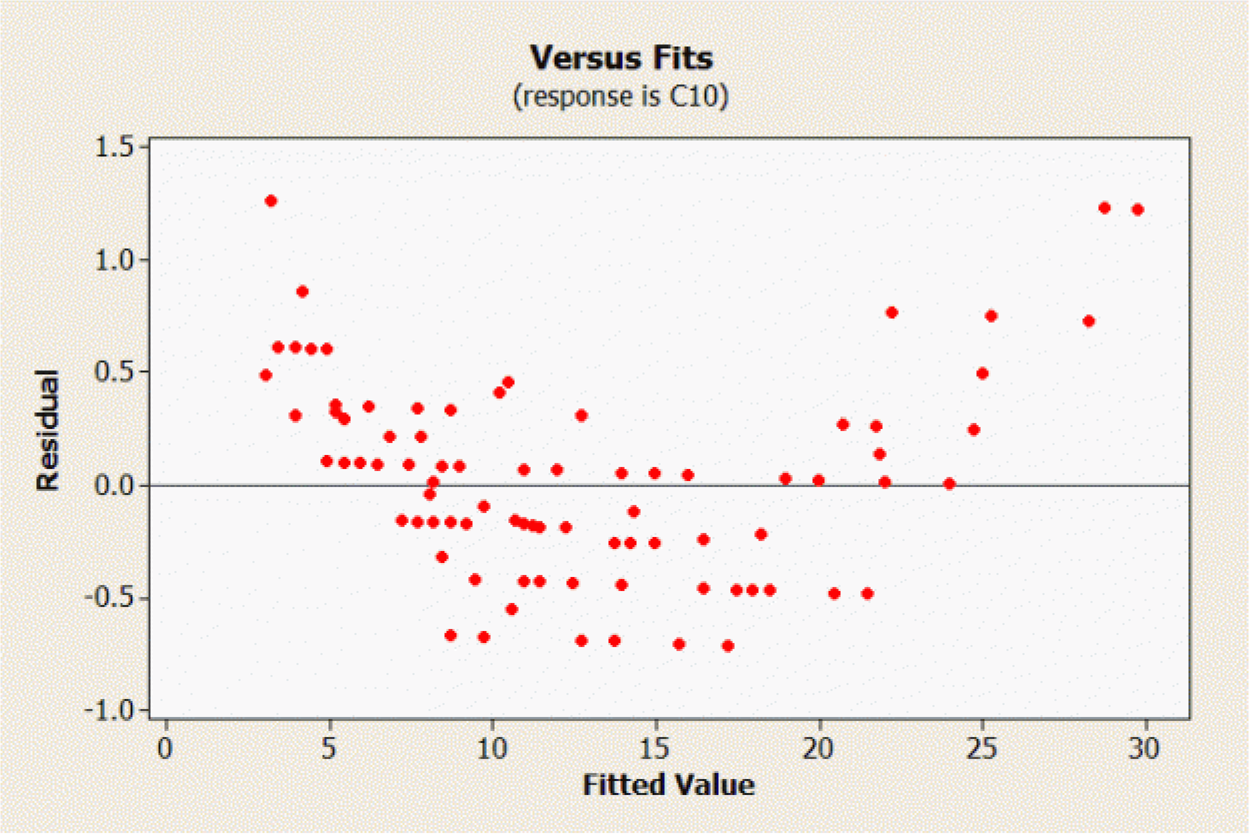
- ✓ have a constant variance
- ✓ be approximately normally distributed (with a mean of zero), and
- ✓ be independent of one another.

The most useful graph for analyzing residuals is a *residual by predicted* plot. This is a graph of each residual value plotted against the corresponding predicted value.If the assumptions are met, the residuals will be randomly scattered around the center line of zero, with no obvious pattern. The residuals will look like an unstructured cloud of points, centered at zero



The points on the plot **above** appear to be randomly scattered around zero, so assuming that the error terms have a mean of zero is reasonable. The vertical width of the scatter doesn't appear to increase or decrease across the fitted values, so we can assume that the variance in the error terms is constant.

What if we did notice a pattern in the plot?



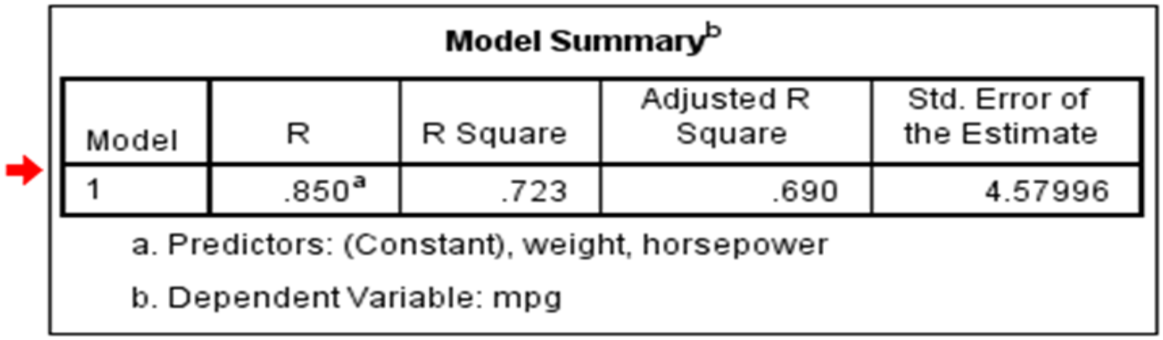
There is definitely a noticeable pattern here! The residuals (error terms) take on positive values with small or large fitted values, and negative values in the middle. The width of the scatter seems consistent, but the points are not randomly scattered around the zero line from left to right. This graph tells us we should not use the regression model that produced these results

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 Coefficient 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.

It is easy to explain the R square in terms of regression. It is not so easy to explain the R in terms of regression.



Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.850[a]	.723	.690	4.57996

a. Predictors: (Constant), weight, horsepower

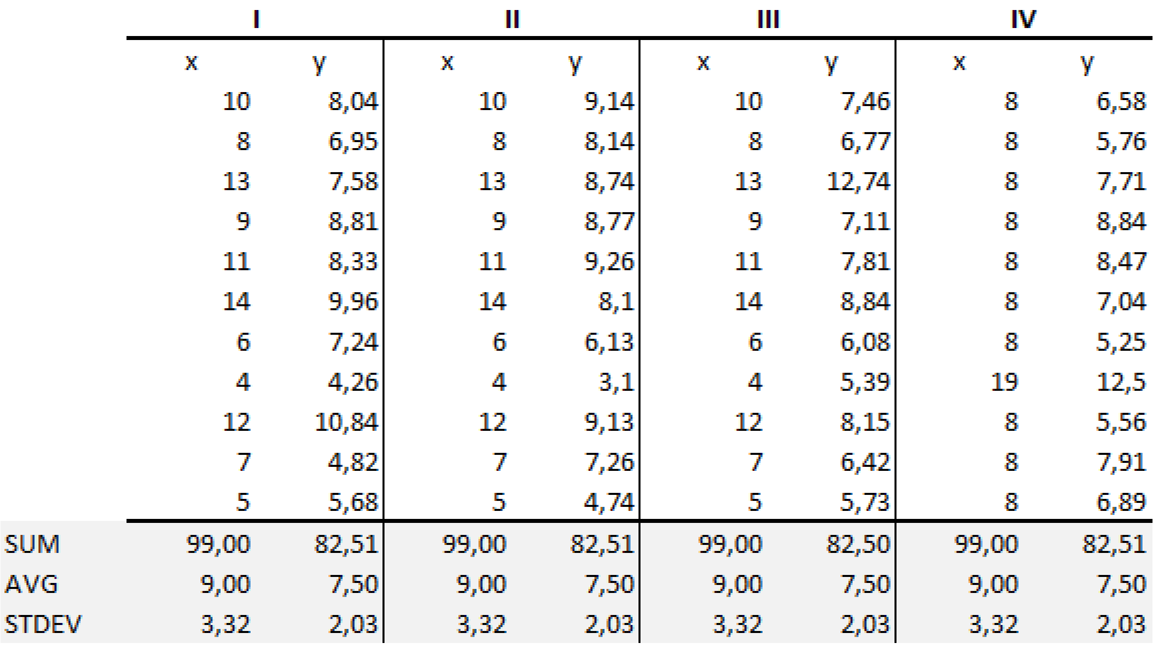
b. Dependent Variable: mpg

Coefficient of Correlation is the R value i.e. .850 (or 85%). Coefficient of Determination is the R square value i.e. .723 (or 72.3%). R square is simply square of R i.e. R times R.

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 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. You can explain R square for both simple linear regressions and for multiple linear regressions.

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, when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.



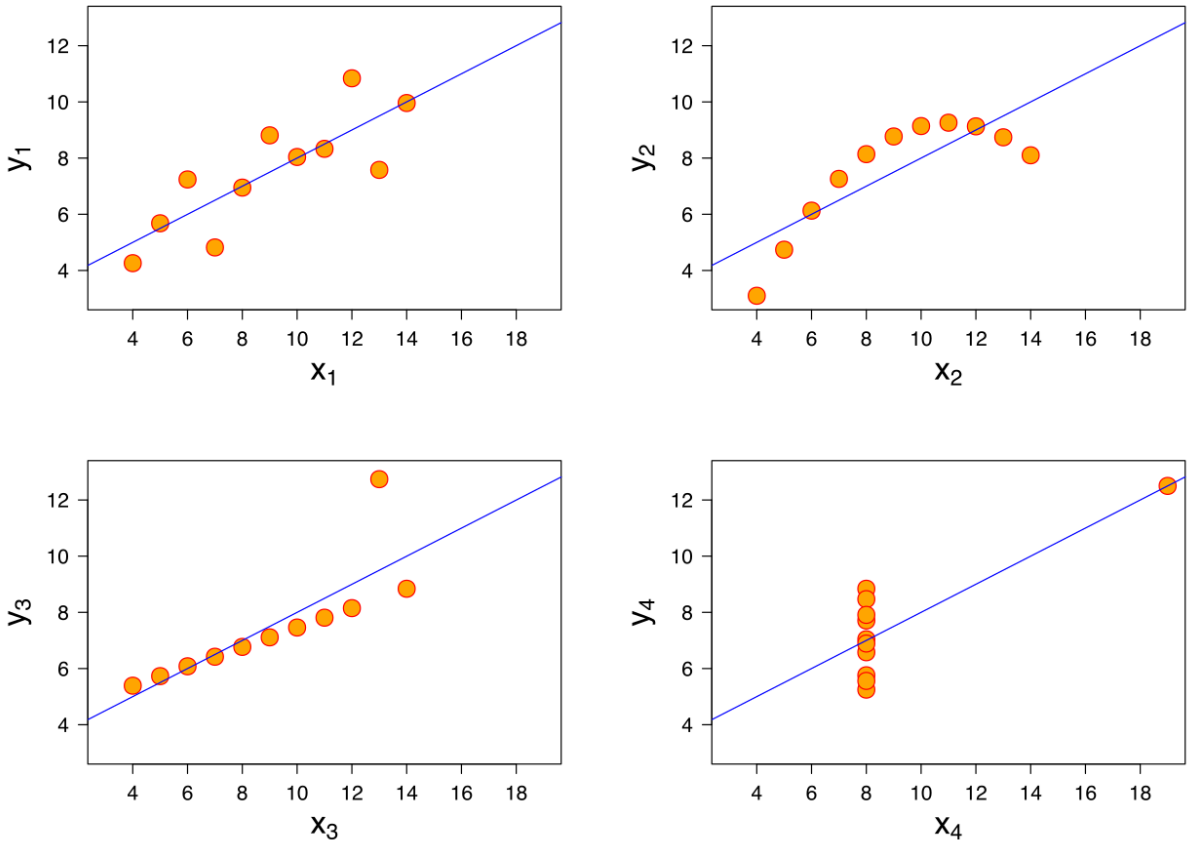
	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	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

Quartet’s Summary Stats

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 appear 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.
Anscombe's Quartet shows that multiple data sets with many similar statistical properties can still be vastly different from one another when graphed.
Anscombe's Quartet warns of the dangers of outliers in data sets.
Anscombe's Quartet reminds us that graphing data prior to analysis is good practice, outliers should be removed when analyzing data, and statistics about a data set do not fully depict the data set in its entirety.

5. What is Pearson’s R?

- **Correlation coefficients** are used in statistics to measure how strong a relationship is between two variables. There are several types of correlation coefficient: Pearson’s correlation (also called Pearson’s *R*) is a **correlation coefficient** commonly used in linear regression.
- Correlation between sets of data is a measure of how well they are related. The most common measure of correlation in stats is the Pearson Correlation. The full name is the **Pearson Product Moment Correlation (PPMC)**. It shows the linear relationship between two sets of data.
- In simple terms, it answers the question, *can a line be drawn in a graph to represent the data?* Two letters are used to represent the Pearson correlation: Greek letter rho (ρ) for a population and the letter “r” for a sample.
- Potential problems with Pearson correlation is that The PPMC is not able to tell the difference between dependent variables and independent variables
- In addition, the PPMC will not give you any information about the slope of the line; it only tells you whether there is a relationship.

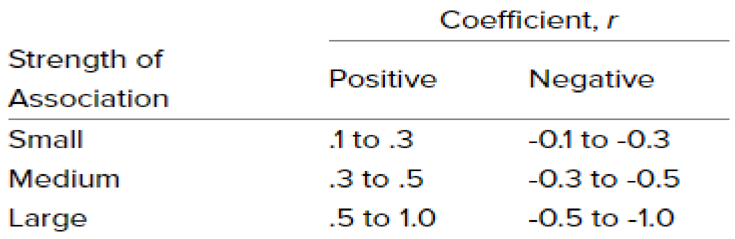
<u>Pearson's assumptions of correlation:</u>

There are five assumptions that are made with respect to Pearson's correlation:

- The variables must be either interval or ratio measurements
- The variables must be approximately normally distributed
- There is a linear relationship between the two variables
- Outliers are either kept to a minimum or are removed entirely.
- There is homoscedasticity of the data.

<u>guidelines to interpreting Pearson's correlation coefficient:</u>

The following guidelines have been proposed:



	Coefficient, *r*	
Strength of Association	Positive	Negative
Small	.1 to .3	-0.1 to -0.3
Medium	.3 to .5	-0.3 to -0.5
Large	.5 to 1.0	-0.5 to -1.0

<u>How to Find Pearson’s Correlation Coefficients:</u>

For e.g.:

Sample question: Find the value of the correlation coefficient from the following table:

SUBJECT	AGE X	GLUCOSE LEVEL Y
1	43	99
2	21	65
3	25	79
4	42	75
5	57	87
6	59	81

Step 1: Make a chart. Use the given data, and add three more columns: xy, x2, and y2.

SUBJECT	AGE X	GLUCOSE LEVEL Y	XY	X^2	Y^2
1	43	99			
2	21	65			
3	25	79			
4	42	75			
5	57	87			
6	59	81			

Step 2: Multiply x and y together to fill the xy column. For example, row 1 would be 43 × 99 = 4,257.

SUBJECT	AGE X	GLUCOSE LEVEL Y	XY	X^2	Y^2
1	43	99	4257		
2	21	65	1365		
3	25	79	1975		
4	42	75	3150		
5	57	87	4959		
6	59	81	4779		

Step 3: Take the square of the numbers in the x column and put the result in the x2 column.

SUBJECT	AGE X	GLUCOSE LEVEL Y	XY	X^2	Y^2
1	43	99	4257	1849	
2	21	65	1365	441	
3	25	79	1975	625	
4	42	75	3150	1764	
5	57	87	4959	3249	
6	59	81	4779	3481	

Step 4: Take the square of the numbers in the y column and put the result in the y2 column.

SUBJECT	AGE X	GLUCOSE LEVEL Y	XY	X^2	Y^2
1	43	99	4257	1849	9801
2	21	65	1365	441	4225
3	25	79	1975	625	6241
4	42	75	3150	1764	5625
5	57	87	4959	3249	7569
6	59	81	4779	3481	6561

Step 5: Add up all of the numbers in the columns and put the result at the bottom of the column. The Greek letter sigma (Σ) is a short way of saying “sum of.”

SUBJECT	AGE X	GLUCOSE LEVEL Y	XY	X^2	Y^2
1	43	99	4257	1849	9801
2	21	65	1365	441	4225
3	25	79	1975	625	6241
4	42	75	3150	1764	5625
5	57	87	4959	3249	7569
6	59	81	4779	3481	6561
Σ	247	486	20485	11409	40022

Step 6: Use the following correlation coefficient formula.

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

The answer is: 2868 / 5413.27 = 0.529809
From our table:

- Σx = 247
- Σy = 486
- Σxy = 20,485
- Σx2 = 11,409
- Σy2 = 40,022
- n is the sample size, in our case = 6

The correlation coefficient =

- 6(20,485) – (247 × 486) / [√[[6(11,409) – (2472)] × [6(40,022) – 4862]]]
  = 0.5298

The range of the correlation coefficient is from -1 to 1. Our result is 0.5298 or 52.98%, which means the variables have a moderate positive correlation.

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

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.

Why scaling is performed:

- 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 Euclidian 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 suppress this effect, we need to bring all features to the same level of magnitudes. This can be achieved by scaling.

Techniques to perform Feature Scaling
Consider the two most important ones:

- **Min-Max Normalization**: This technique re-scales a feature or observation value with distribution value between 0 and 1.

$$X_{new} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

- **Standardization**: It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.

$$X_{new} = \frac{X_i - X_{mean}}{\text{Standard Deviation}}$$

Difference between normalized scaling and standardized scaling:

The terms *normalization* and *standardization* are sometimes used interchangeably, but they usually refer to different things.
Normalization usually means to scale a variable to have a values between 0 and 1,
while *standardization* transforms data to have a mean of zero and a standard deviation of 1. This standardization is called a **z-score**, and data points can be standardized with the following formula:

Normalization:

- Assume that in the given population (list of values) is in the range of [x_min, x_max]

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

- Here we subtract min value of the population from all values so that the new range of population will be [0, (x_max - x_min)]
- Now we will divide each element by (x_max - x_min) to convert the range of population into [0, 1]
- Summary: Normalization used where we want to scale down the population to [0, 1]

Standardization:

- Assume the same population again with range [x_min, x_max]
- First calculate standard deviation (x_bar means average of population)

$$\sigma = \sqrt{\frac{\Sigma\ (x - \overline{x})^2}{n}}$$

σ = lower case sigma
Σ = capital sigma
$\overline{x}$ = x bar

- Now standardize the population with following equation (µ = average of population that we computed above):

$$x_{new} = \frac{x - \mu}{\sigma}$$

- Summary: With standardization we can transform the data into a range such that the new population has mean (average) = 0 and standard deviation = 1.

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

Variance inflation fVariance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables**.**

- A variance inflation factor (VIF) provides a measure of multicollinearity among the independent variables in a multiple regression model.
- Detecting multicollinearity is important because while it does not reduce the explanatory power of the model, it does reduce the statistical significance of the independent variables.
- A large VIF on an independent variable indicates a highly collinear relationship to the other variables that should be considered or adjusted for in the structure of the model and selection of independent variables.

Understanding large (infinite) VIF value

- Multicollinearity exists when there is a linear relationship, or correlation, between one or more of the independent variables or inputs
- Multicollinearity creates a problem in the multiple regression because since the inputs are all influencing each other, they are not actually independent, and it is difficult to test how much the combination of the independent variables affects the dependent variable, or outcome, within the regression model.
- Using variance inflation factors helps to identify the severity of any multicollinearity issues so that the model can be adjusted
- Variance inflation factor measures how much the behavior (variance) of an independent variable is influenced, or inflated, by its interaction/correlation with the other independent variables
- When significant multicollinearity issues exist, the **variance inflation factor will be very large** for the variables involved.
- A value of 1 means that the predictor is not correlated with other variables. The higher the value, the greater the correlation of the variable with other variables. Values of more than 4 or 5 are sometimes regarded as being moderate to high, with values of 10 or more being regarded as very high. These numbers are just rules of thumb; in some contexts, a VIF of 2 could be a great problem (e.g., if estimating price elasticity), whereas in straightforward predictive applications very high VIFs may be unproblematic.
- If one variable has a high VIF it means that other variables must also have high VIFs. In the simplest case, two variables will be highly correlated, and each will have the same high VIF.
- Where a VIF is high, it makes it difficult to disentangle the relative importance of predictors in a model, particularly if the standard errors are regarded as being large.
- This is particularly problematic in two scenarios, where:
    - The focus of the model is on making inferences regarding the relative importance of the predictors.
    - The model is to be used to make predictions in a different data set, in which the correlations may be different.

The higher the VIF, the more the standard error is inflated, and the larger the confidence interval and the smaller the chance that a coefficient is determined to be statistically significant

8. What is the Gauss-Markov theorem?

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.

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.

Gauss Markov Assumptions

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.

Purpose of the Assumptions

The Gauss Markov assumptions guarantee the validity of ordinary least squares for estimating regression coefficients.

Checking how well our data matches these assumptions is an important part of estimating regression coefficients. When you know where these conditions are violated, you may be able to plan ways to change your experiment setup to help your situation fit the ideal Gauss Markov situation more closely.

In practice, the Gauss Markov assumptions are rarely all met perfectly, but they are still useful as a benchmark, and because they show us what ‘ideal’ conditions would be. They also allow us to pinpoint problem areas that might cause our estimated regression coefficients to be inaccurate or even unusable.

The Gauss-Markov Assumptions in Algebra

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

yi = xi‘ β + εi

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

- E{εi} = 0, i = 1, . . . , N
- {ε1……εn} and {x1…..,xN} are independent
- cov{εi, εj} = 0, i, j = 1,…., N I ≠ j.
- V{ε1 = σ2, 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.

9. Explain the gradient descent algorithm in detail

Gradient Descent is an optimization algorithm used for minimizing the cost function in various machine learning algorithms. It is basically used for updating the parameters of the learning model.

Types of gradient Descent:

- Batch Gradient Descent: This is a type of gradient descent which processes all the training examples for each iteration of gradient descent. But if the number of training examples is large, then batch gradient descent is computationally very expensive. Hence if the number of training examples is large, then batch gradient descent is not preferred. Instead, we prefer to use stochastic gradient descent or mini-batch gradient descent.
- Stochastic Gradient Descent: This is a type of gradient descent which processes 1 training example per iteration. Hence, the parameters are being updated even after one iteration in which only a single example has been processed. Hence this is quite faster than batch gradient descent. But again, when the number of training examples is large, even then it processes only one example which can be additional overhead for the system as the number of iterations will be quite large.
- Mini Batch gradient descent: This is a type of gradient descent which works faster than both batch gradient descent and stochastic gradient descent. Here b examples where b<m are processed per iteration. So even if the number of training examples is large, it is processed in batches of b training examples in one go. Thus, it works for larger training examples and that too with lesser number of iterations.

Variables used:
Let m be the number of training examples.
Let n be the number of features.
Note: if b == m, then mini batch gradient descent will behave similarly to batch gradient descent.

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)

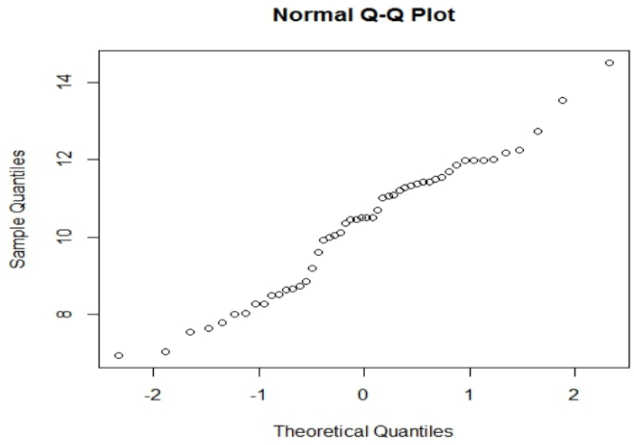
This process is repeated until the cost of the coefficients (cost) is 0.0 or close enough to zero to be good enough.

Summary:

- Optimization is a big part of machine learning.
- Gradient descent is a simple optimization procedure that you can use with many machine learning algorithms.
- Batch gradient descent refers to calculating the derivative from all training data before calculating an update.
- Stochastic gradient descent refers to calculating the derivative from each training data instance and calculating the update immediately.

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

- Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other.
- A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it.
- The purpose of Q-Q plots is to find out if two sets of data come from the same distribution. A 45-degree angle is plotted on the Q-Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.
- The image above shows quantiles from a theoretical normal distribution on the horizontal axis. It's being compared to a set of data on the y-axis.
- This particular type of Q-Q plot is called a **normal quantile-quantile (QQ) plot.** The points are not clustered on the 45-degree line, and in fact follow a curve, suggesting that the sample data is not normally distributed
- 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.



The q-q plot is formed by:

- Vertical axis: Estimated quantiles from data set 1
- Horizontal axis: Estimated quantiles from data set 2
- Both axes are in units of their respective data sets. That is, the actual quantile level is not plotted. For a given point on the q-q plot, we know that the quantile level is the same for both points, but not what that quantile level is.
- If the data sets have the same size, the q-q plot is essentially a plot of sorted data set 1 against sorted data set 2. If the data sets are not of equal size, the quantiles are usually picked to correspond to the sorted values from the smaller data set and then the quantiles for the larger data set are interpolated

The advantages of the q-q plot are:

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

The q-q plot is used to answer the following questions:

- Do two data sets come from populations with a common distribution?
- Do two data sets have common location and scale?
- Do two data sets have similar distributional shapes?
- Do two data sets have similar tail behavior?