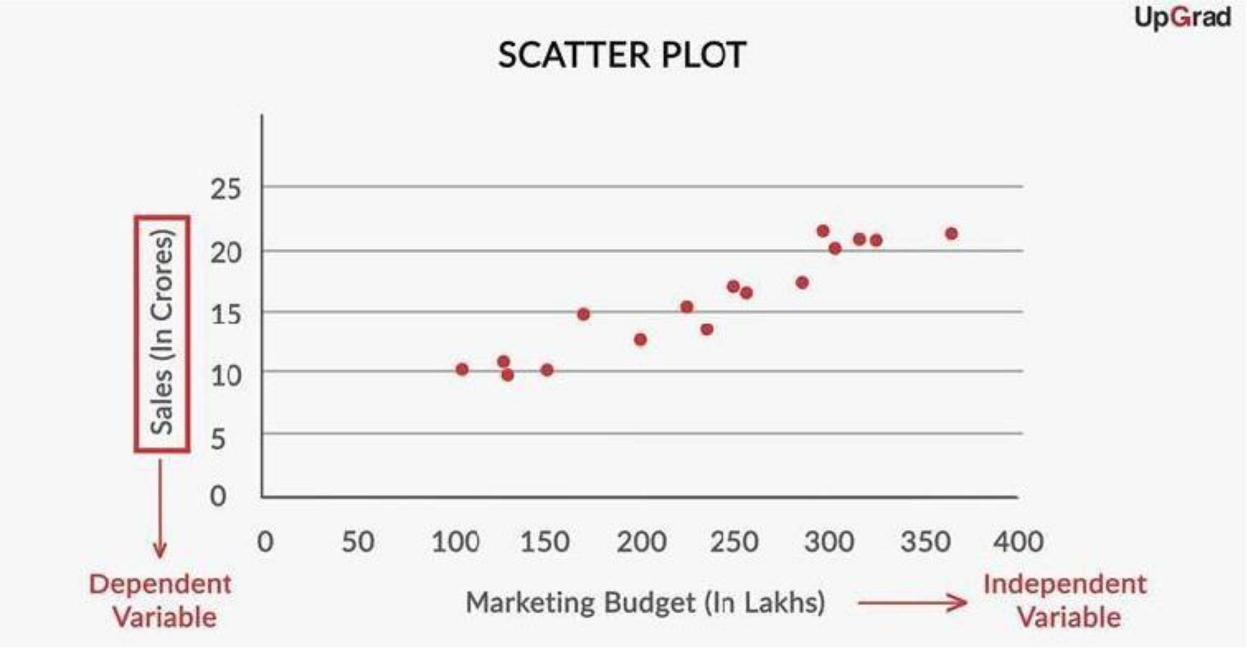
1. Explain the linear regression algorithm in detail

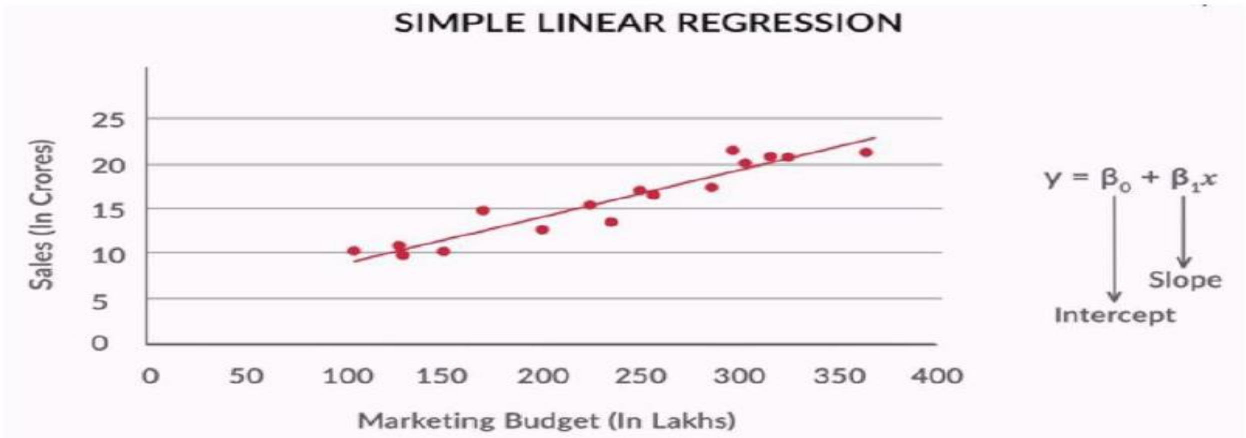
Answer:
The linear regression concepts use the prediction of future it is a form of predictive modelling technique which tells us the relationship between the dependent and independent variables.

These are two types of linear regression
● Simple linear regression
● Multiple linear regression

Simple linear regression:
The regression model is the simple linear regression which explains the relationship between a dependent variable and one independent variable using a straight line. The straight line is plotted on the scatter plot of these two points.



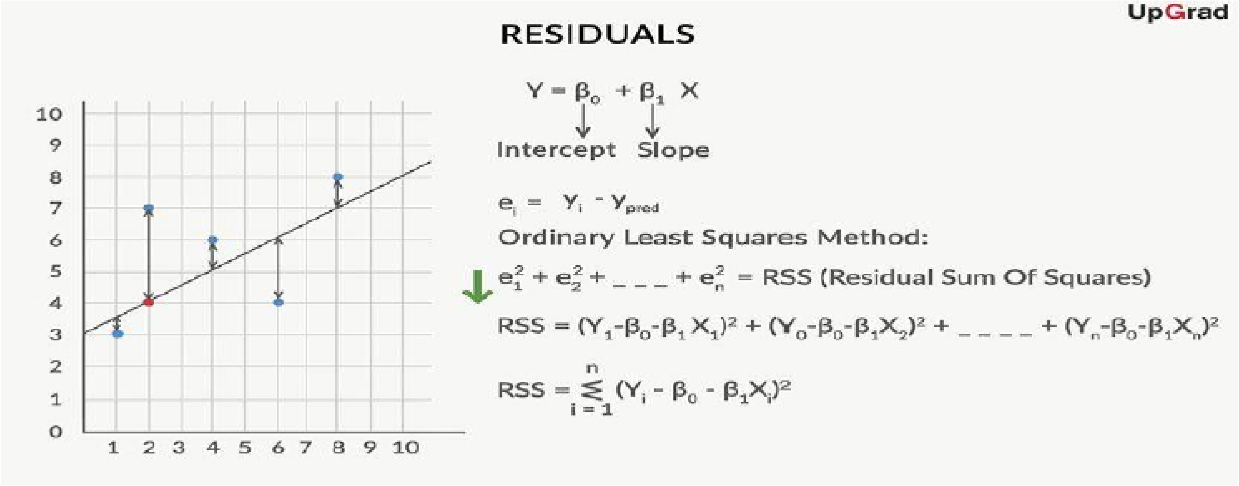
The standard equation of the regression line is given by the following expression: Y = $\beta_0 + \beta_1 X$



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

Answer:

The expression of RSS (Residual Sum of Squares) which is equal to the sum of squares of the residual for each data point in the plot. Residuals for any data point is found by subtracting predicted value of dependent variable from actual value of dependent variable



RESIDUALS

$Y = \beta_0 + \beta_1 X$

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)

$RSS = (Y_1-\beta_0-\beta_1 X_1)^2 + (Y_0-\beta_0-\beta_1 X_2)^2 + \_\_\_\_ + (Y_n-\beta_0-\beta_1 X_n)^2$

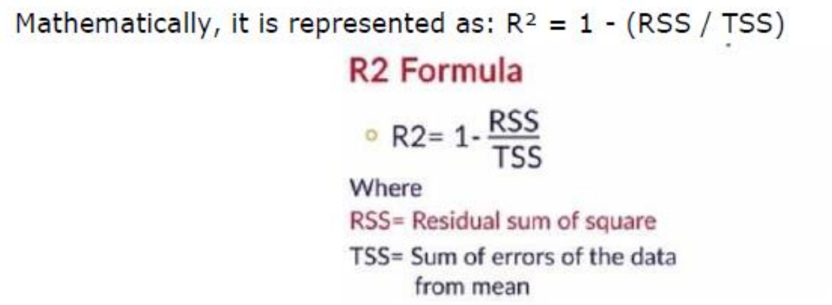
$$RSS = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

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

Answer:

<u>The coefficient of determination:</u>

An alternative way of checking the accuracy of the model, which is R2 statistics. R2 is a number that explains what portion of the given data variation is explained by the developed model. It always takes a value between 0 & 1. 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. Overall, the higher the R-squared, the better the model fits your data.



Mathematically, it is represented as: $R^2 = 1 - (RSS / TSS)$

R2 Formula

○ $R2 = 1 - \frac{RSS}{TSS}$

Where

RSS= Residual sum of square

TSS= Sum of errors of the data from mean

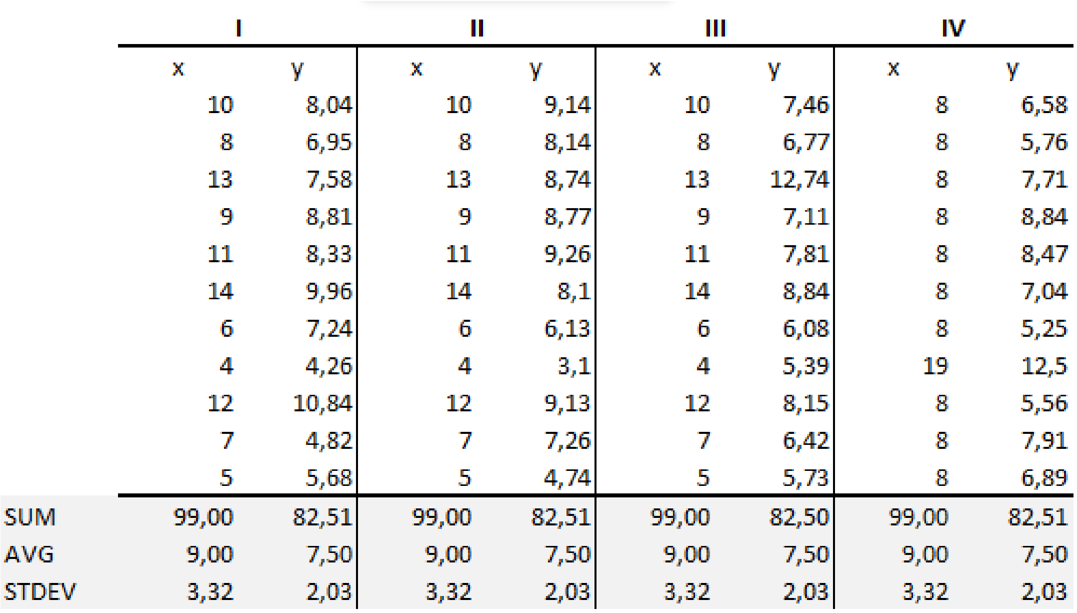
The coefficient of correlation:

It is the degree of relationship between two variables 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 world can be argued to have a correlation value. If they are not cor-related then the correlation value can still be computed which would be 0. The correlation value always lies between -1 and 1.

4. Explain the Anscombe’s quartet in detail.

Answer:

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 we must emphasize COMPLETELY, when they are graphed. Each graph said a different story irrespective of their similar summary statistics.



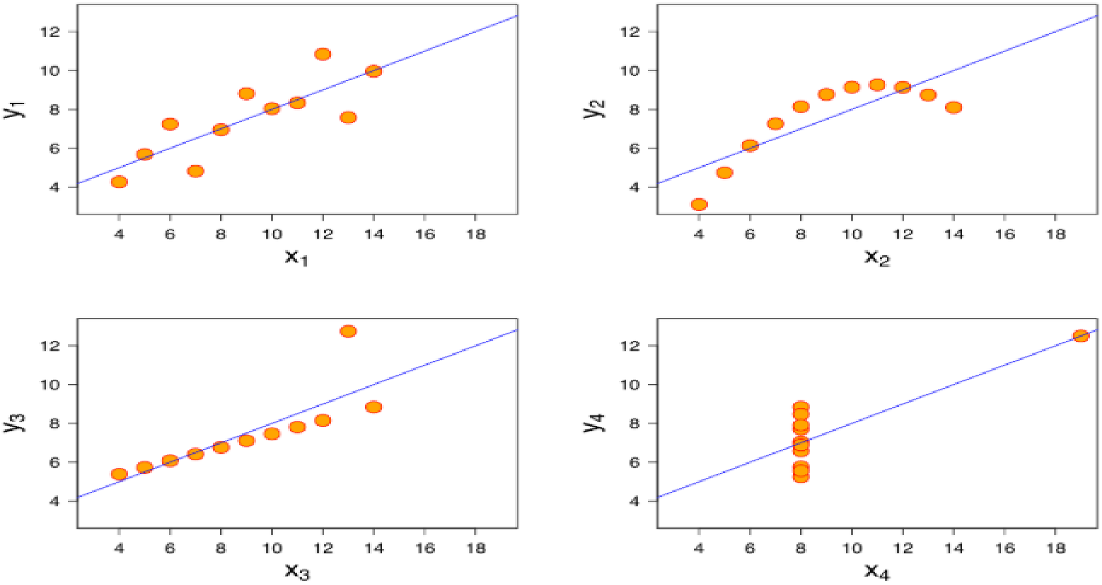
	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	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 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. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

5. What is Pearson's R?

Answer:

The Pearson's correlation coefficient varies between -1 and +1 where

r = 1 means the data is perfectly linear with a positive slope
r = -1 means the data is perfectly linear with a negative slope
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?

Answer:

It is important aspect to consider is feature scaling. When you have a lot of independent variables in a model, a lot of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret. Therefore, we need to scale features because of two reasons

1. Ease of interpretation.
2. Faster convergence for gradient descent methods

We can scale the features using two very popular method:

A. **Standardizing:** The variables are scaled in such a way that their mean is zero and standard deviation is one.

$$x = \frac{x - mean(x)}{sd(x)}$$

B. **Min Max Scaling**: The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.

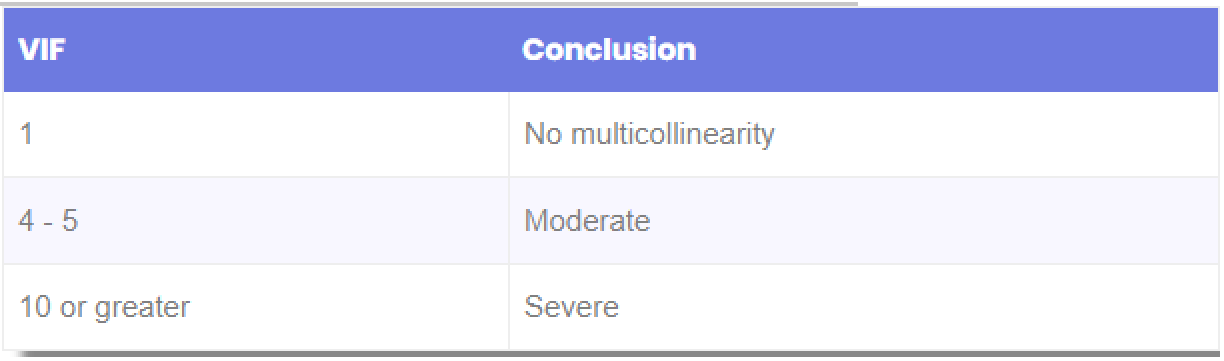
$$x = \frac{x - min(x)}{max(x) - min(x)}$$

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F statistic, p-values, R-square, etc.

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

Answer:

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.



VIF	Conclusion
1	No multicollinearity
4 - 5	Moderate
10 or greater	Severe

8. What is the Gauss-Markov theorem?

Answer:

The Gauss Markov theorem say 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:

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 are not perfectly correlated with each other.
4. Exogeneity: the regressors are not correlated with the error term.
5. Homoscedasticity: no matter what the values of our regressors might be, the error of the variance is constant.

9. Explain the gradient descent algorithm in detail.

Answer:
Gradient descent is an optimization algorithm that is used to find the values of parameters of a function (f) that minimizes a cost function (cost).
Gradient descent is best used when the parameters cannot be calculated analytically and must be searched for by an optimization algorithm.

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)

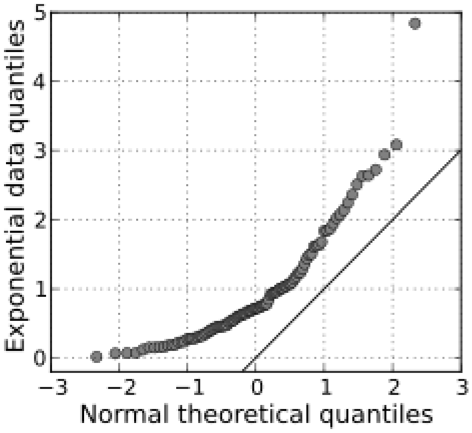
This process is repeated until the cost of the coefficients (cost) is 0.0 or close enough to zero to be good enough.

we can see how simple gradient descent is. It does require you to know the gradient of your cost function or the function you are optimizing, but besides that, it's very straightforward.

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

Answer:

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 45degree 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 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.