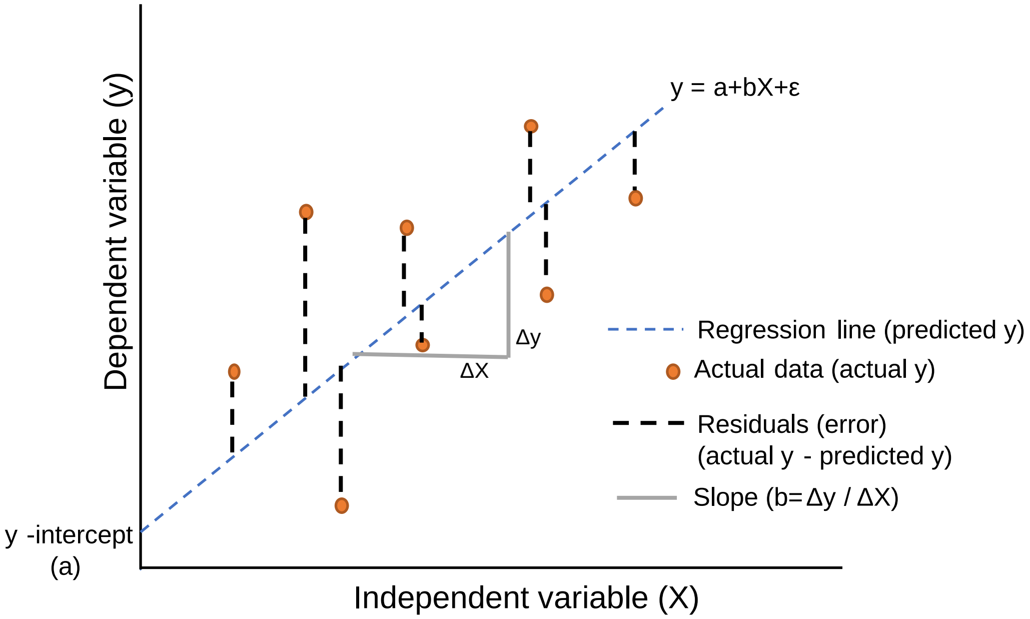
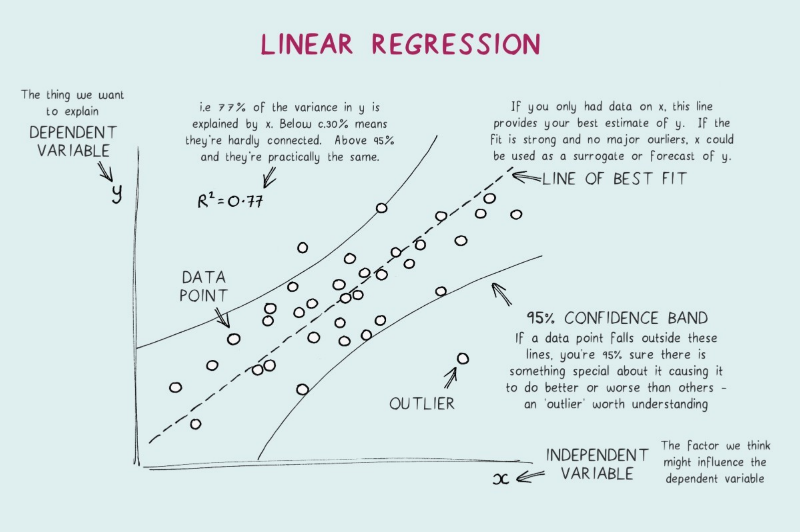
1. Using a graph to illustrate slope and intercept, define basic linear regression.
2. In a graph, explain the terms rise, run, and slope.
3. Use a graph to demonstrate slope, linear positive slope, and linear negative slope, as well as the different conditions that contribute to the slope.
4. Use a graph to demonstrate curve linear negative slope and curve linear positive slope.
5. Use a graph to show the maximum and low points of curves.
6. Use the formulas for a and b to explain ordinary least squares.
7. Provide a step-by-step explanation of the OLS algorithm.
8. What is the regression's standard error? To represent the same, make a graph.
9. Provide an example of multiple linear regression.
10. Describe the regression analysis assumptions and the BLUE principle.
11. Describe two major issues with regression analysis.
12. How can the linear regression model's accuracy be improved?
13. Using an example, describe the polynomial regression model in detail.
14. Provide a detailed explanation of logistic regression.
15. What are the logistic regression assumptions?
16. Go through the details of maximum likelihood estimation.

Q. 1. Using a graph to illustrate slope and intercept, define basic linear regression.



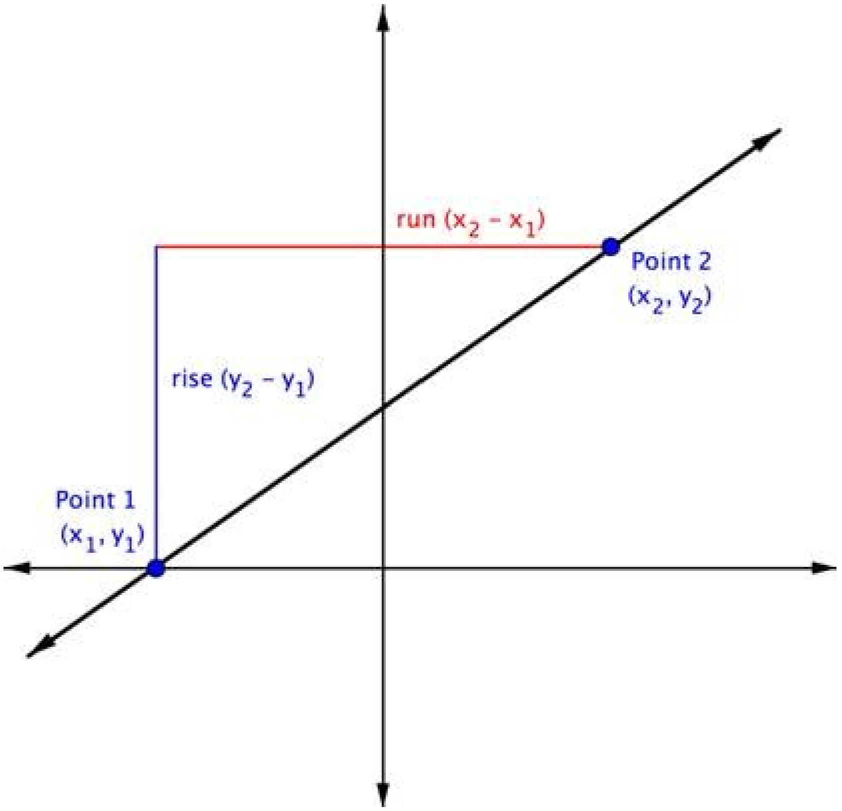
Sol:


LINEAR REGRESSION
The thing we want to explain
DEPENDENT VARIABLE
y
i.e 77% of the variance in y is explained by x. Below c.30% means they're hardly connected. Above 95% and they're practically the same.
$R^2 = 0{\cdot}77$
If you only had data on x, this line provides your best estimate of y. If the fit is strong and no major ourliers, x could be used as a surrogate or forecast of y.
LINE OF BEST FIT
DATA POINT
95% CONFIDENCE BAND
If a data point falls outside these lines, you're 95% sure there is something special about it causing it to do better or worse than others - an 'outlier' worth understanding
OUTLIER
INDEPENDENT VARIABLE
x
The factor we think might influence the dependent variable

Q. 2. In a graph, explain the terms rise, run, and slope.

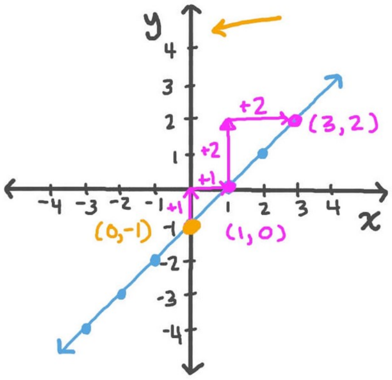
Sol: Using two of the points on the line, you can find the slope of the line by finding the rise and the run. The vertical change between two points is called the rise, and the horizontal change is called the run. The slope equals the rise divided by the run: Slope =rise/run.

```
Slope = rise/ run
```



Q. 3. Use a graph to demonstrate slope, linear positive slope, and linear negative slope, as well as the different conditions that contribute to the slope.

Find the slope m and the y-intercept b of this straight line.



$$\text{slope } (m) = \frac{\text{rise}}{\text{run}} = \frac{\text{changes in } y}{\text{changes in } x}$$

y-intercept (b) = the point where the line crosses the y-axis

$$m = \frac{+2}{+2} = 1$$

$$\boxed{\begin{matrix} m=1 \\ b=-1 \end{matrix}}$$

$$y = mx + b$$

$$y = x - 1$$

Sol:

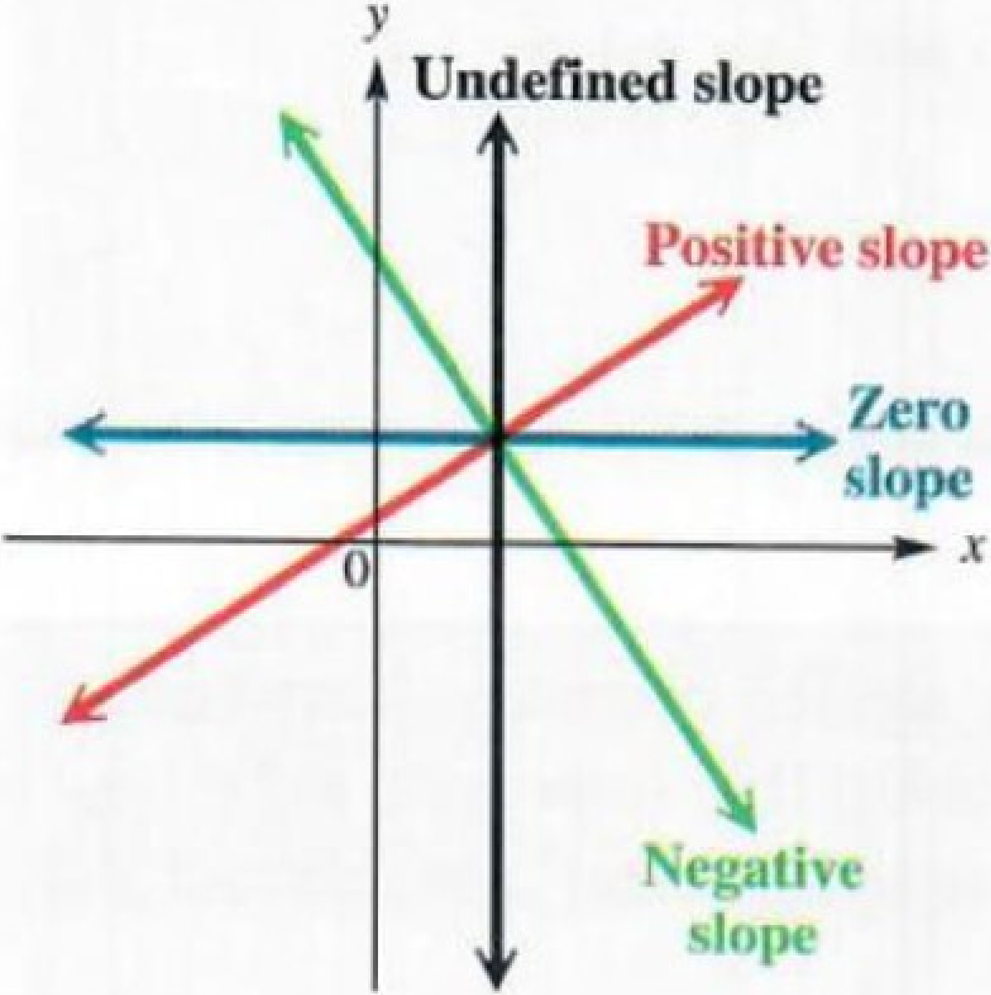
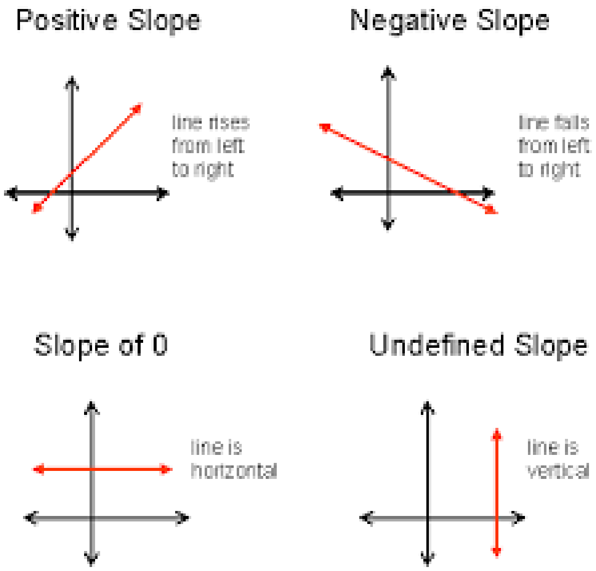
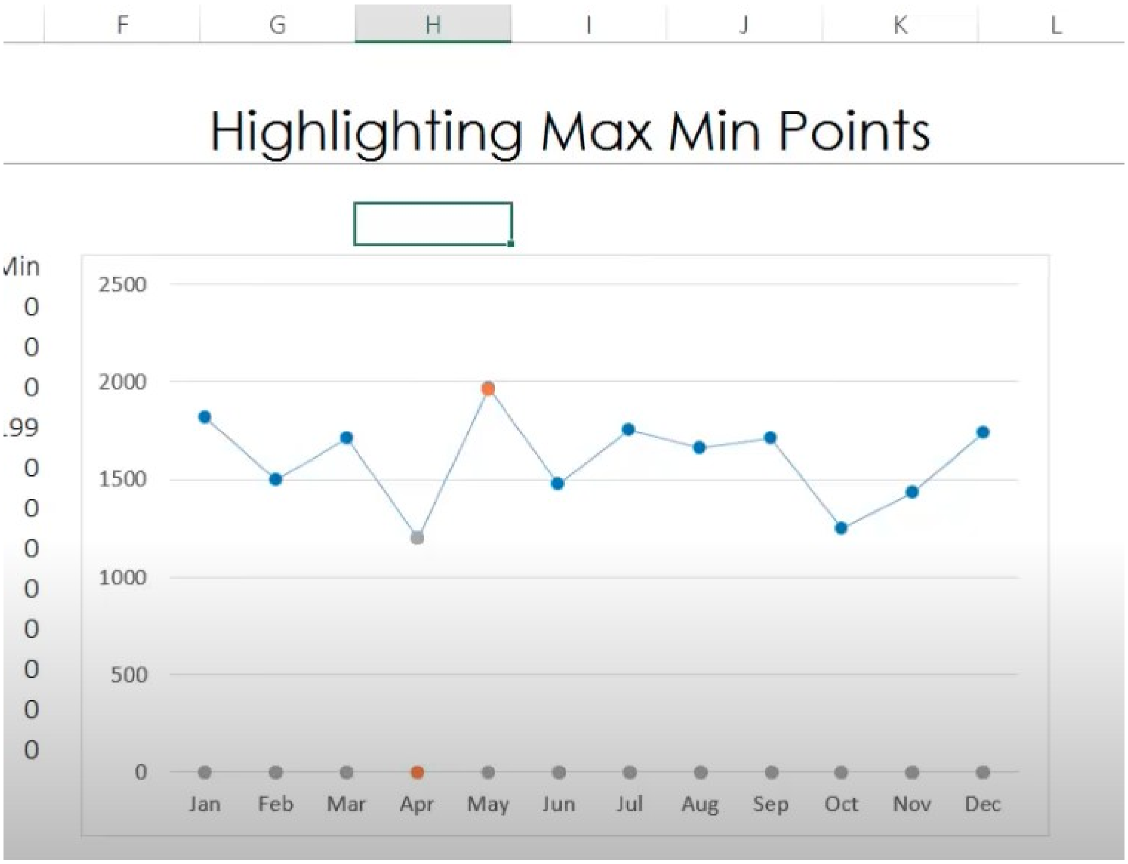

y
Undefined slope
Positive slope
Zero
slope
x
0
Negative
slope

FIGURE 3.12

Q. 4. Use a graph to demonstrate curve linear negative slope and curve linear positive slope.



Q. 5. Use a graph to show the maximum and low points of curves.



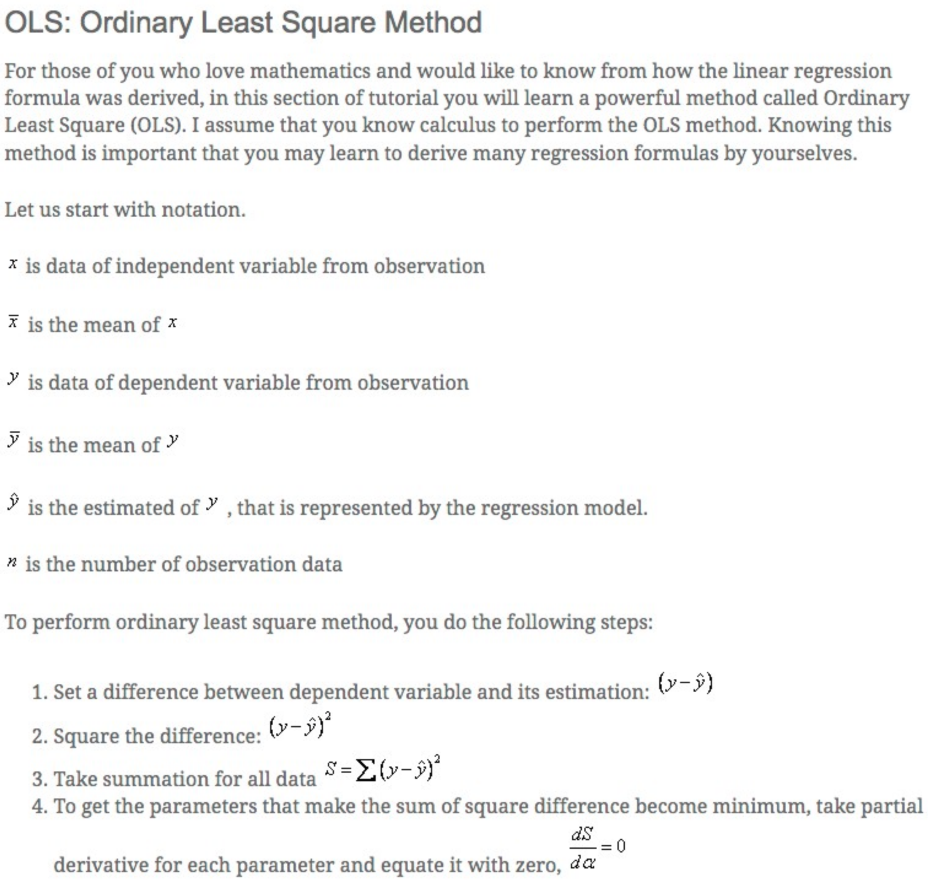
Sol:

Q. 6. Use the formulas for a and b to explain ordinary least squares.

Sol: Ordinary least squares (OLS) regression is a statistical method of analysis that estimates the relationship between one or more independent variables and a dependent variable; the method estimates the relationship by minimizing the sum of the squares in the difference between the observed and predicted values of the dependent variable configured as a straight line. In this entry, OLS regression will be discussed in the context of a bivariate model, that is, a model in which there is only one independent variable (X) predicting a dependent variable (Y). However, the logic of OLS regression is easily extended to the multivariate model in which there are two or more independent variables.

OLS = ^β = (XTX)−1XTy

Q 7. Provide a step-by-step explanation of the OLS algorithm.



OLS: Ordinary Least Square Method

For those of you who love mathematics and would like to know from how the linear regression formula was derived, in this section of tutorial you will learn a powerful method called Ordinary Least Square (OLS). I assume that you know calculus to perform the OLS method. Knowing this method is important that you may learn to derive many regression formulas by yourselves.

Let us start with notation.

x is data of independent variable from observation

$\bar{x}$ is the mean of x

y is data of dependent variable from observation

$\bar{y}$ is the mean of y

$\hat{y}$ is the estimated of y , that is represented by the regression model.

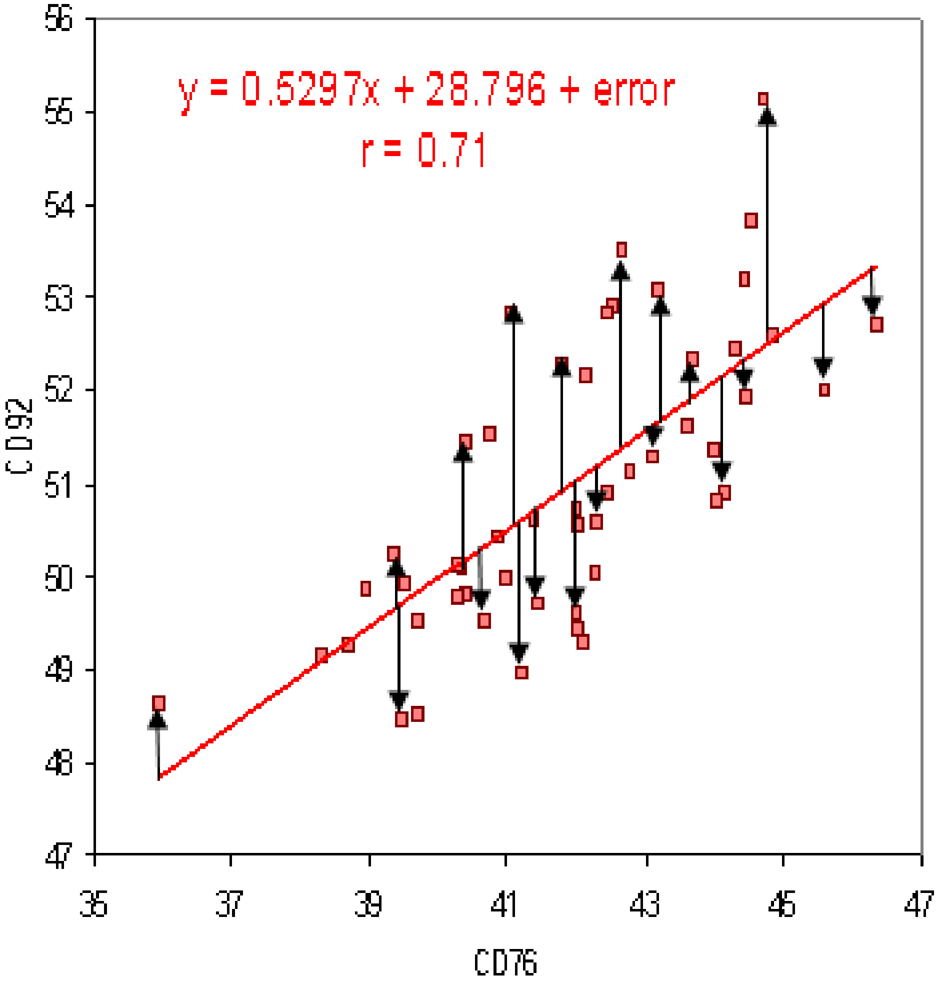
n is the number of observation data

To perform ordinary least square method, you do the following steps:

1. Set a difference between dependent variable and its estimation: $(y-\hat{y})$
2. Square the difference: $(y-\hat{y})^2$
3. Take summation for all data $S=\sum(y-\hat{y})^2$
4. To get the parameters that make the sum of square difference become minimum, take partial derivative for each parameter and equate it with zero, $\frac{dS}{d\alpha}=0$

Q. 8. What is the regression's standard error? To represent the same, make a graph.

Sol: The standard error of the regression (S), also known as the standard error of the estimate, represents the average distance that the observed values fall from the regression line. ... Smaller values are better because it indicates that the observations are closer to the



fitted line.

9. Provide an example of multiple linear regression.

Sol: Predicting the price of the house using various factors like: House area, age of the house, loction

Q. 10. Describe the regression analysis assumptions and the BLUE principle.

Sol:

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.

BLUE principle

The Gauss-Markov Theorem states that if a linear regression model fulfils the assumptions of the classical linear regression model the ordinary least squares estimator is the best linear unbiased estimator (BLUE).

Q. 11. Describe two major issues with regression analysis.

1. Non-Linearity of the response-predictor relationships.
2. Correlation of error terms.
3. A non-constant variance of the error term [Heteroscedasticity]

Q. 12. How can the linear regression model's accuracy be improved?

Sol:

Methods to Boost the Accuracy of a Model.

1. Cross Validation
2. Add more data. Having more data is always a good idea.
3. Treat missing and Outlier values.
4. Feature Engineering.
5. Feature Selection.
6. Multiple algorithms.
7. Algorithm Tuning.
8. Ensemble methods.

Q.13. Using an example, describe the polynomial regression model in detail.

Sol: In statistics, polynomial regression is a form of regression analysis in which the relationship between the independent variable x and the dependent variable y is modelled as an nth degree polynomial in x. ... For this reason, polynomial regression is considered to be a special case of multiple linear regression.

If you see the equation of polynomial regression carefully, then we can see that we are trying to estimate the relationship between coefficients and y. And the values of x and y are already given to us, only we need to determine coefficients and the degree of coefficient here is 1 only, and degree one represents simple linear regression Hence, Polynomial regression is also known as polynomial Linear regression. And this is only the simple concept behind this. I hope you got the point right?

Q. 14. Provide a detailed explanation of logistic regression.

Sol: Logistic Regression was used in the biological sciences in early twentieth century. It was then used in many social science applications. Logistic Regression is used when the dependent variable(target) is categorical. For example, To predict whether an email is spam (1) or (0) Whether the tumor is malignant (1) or not (0) Consider a scenario where we need to classify whether an email is spam or not. If we use linear regression for this problem, there is a need for setting up a threshold based on which classification can be done. Say if the actual class is malignant, predicted continuous value 0.4 and the threshold value is 0.5, the data point will be classified as not malignant which can lead to serious consequence in real time. From this example, it can be inferred that linear regression is not suitable for classification problem. Linear regression is unbounded, and this brings logistic regression into picture. Their value strictly ranges from 0 to 1.

Q. 15. What are the logistic regression assumptions?

Sol: The logistic regression method assumes that: The outcome is a binary or dichotomous variable like yes vs no, positive vs negative, 1 vs 0. There is a linear relationship between the logit of the outcome and each predictor variables.

Q. 16. Go through the details of maximum likelihood estimation.

Sol: In statistics, maximum likelihood estimation (MLE) is a method of estimating the parameters of an assumed probability distribution, given some observed data. This is achieved by maximizing a likelihood function so that, under the assumed statistical model, the observed data is most probable. The point in the parameter space that maximizes the likelihood function is called the maximum likelihood estimate. The logic of maximum likelihood is both intuitive and flexible, and as such the method has become a dominant means of statistical inference.