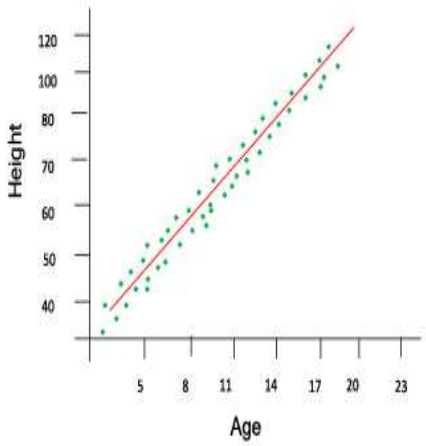
Topic :

1. Logistic regression

Description:

Regression :

Regression is a statistical relationship between two or more variables in which a change in the independent variable is associated with a difference in the dependent variable.

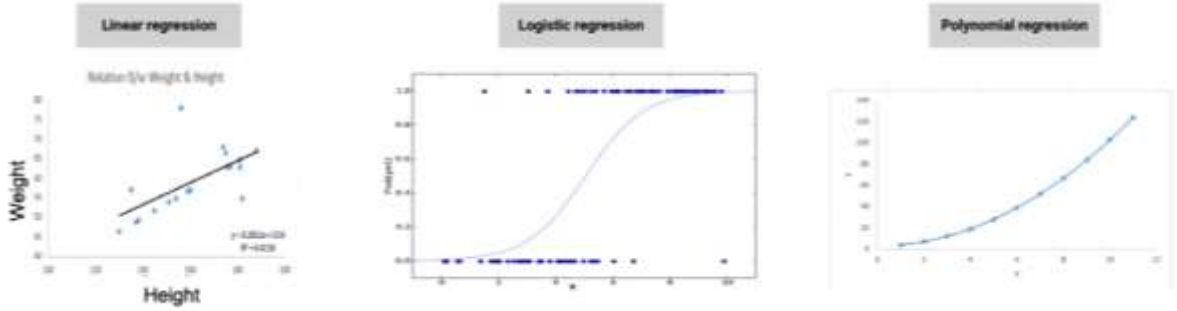


The plot in the middle shows the clear linear relationship between age and height, which is indicated by the solid red line called a trendline, or a regression line, or the **line of best fit**. Here, the height is the dependent variable, and age is the independent variable.

Types of Regression

There are various types of regression:

1. Linear regression logistic
2. Logistic regression
3. Polynomial regression

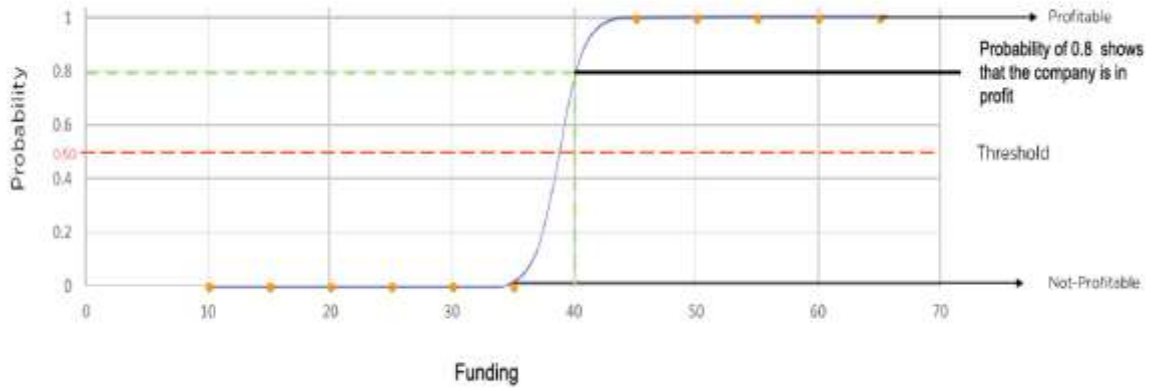


When the Y value in the graph is categorical—such as yes or no, true or false, the subject did or did not do something—then you would use logistic regression. Logistic regression is when the Y value on the graph is categorical and depends on the X variable.

Polynomial regression is when the relationship between the dependent variable Y and the independent variable X is in the nth degree of X. In a plot, you can see that the relationship is not linear; there's a curve to that best-fit trendline.

Why Logistic Regression?

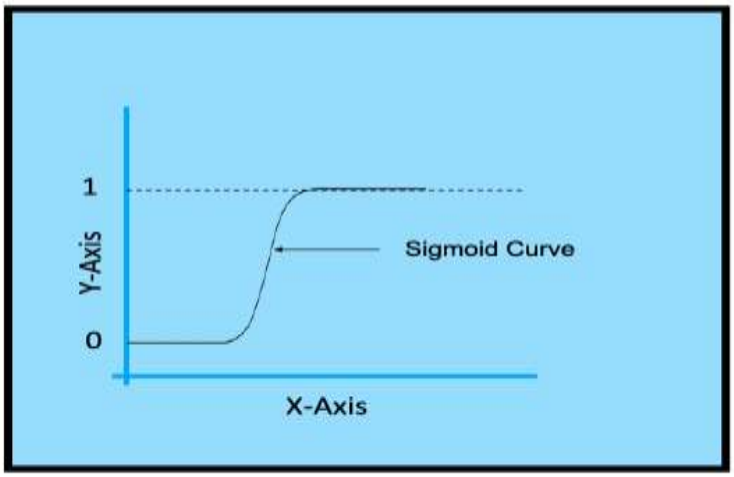
Make use of logistic regression, when there are two outcomes—in our case, profitable and not profitable.



In this example, given the amount of funding, we can calculate the probability that a company will be profitable or not profitable. If you use the threshold line of 0.5, then you have your classifier. If the probability is 0.5 or higher, the company is profitable; if the probability is lower than 0.5, it's not profitable.

What is Logistic Regression?

Using linear regression, you can't divide the output into two distinct categories—yes or no. To divide our results into two categories, you would have to clip the line between 0 and 1. If you recall, probabilities can be between only 0 and 1, and if we're going to use probability on the y-axis, then you can't have anything that is below 0 or above 1.



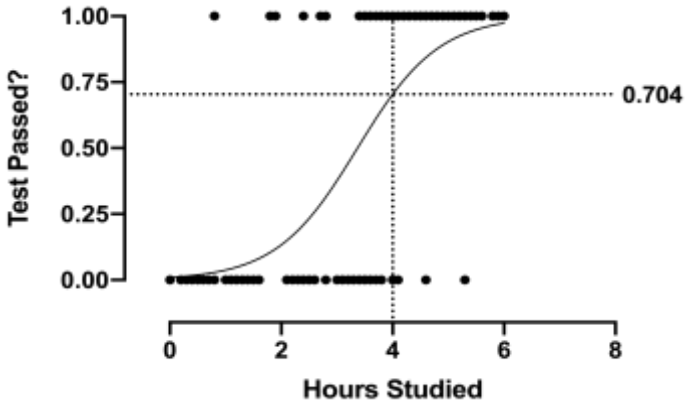
For logistic regression, you will make use of a sigmoid function, and the sigmoid curve is the line of best fit. Notice that it's not linear, but it does satisfy our requirement of using a single line that does not need to be clipped.

Because you cannot use a linear equation for binary predictions, you need to use the sigmoid function, which is represented by the equation:

p = 1/(1+e-y)

e is the base of the natural logs.

The logistic fit is the S-curve that models the probability of success as a function of hours of study. In this example, instructors will be glad to observe that few students who studied 4 hours failed the exam. Indeed, for a student who studied 4 hours, the model predicts the probability of passing to be around 70%.



The S-curve is a byproduct of the way the logistic function estimates the probability. Note that probabilities are bound between 0 and 1, which makes sense: you can't have a "negative probability" of an event happening, and a probability greater than 100% also doesn't make any sense. As such, the upper and lower bounds of the S-curve are also limited by these values. But what this means is that - unlike with linear regression - the values we get from the model don't give us direct estimates for the values we expect to observe. At X = 4, the value of the model is 0.704. However, for any observation that we make at X = 4, the outcome will ONLY be 0 or 1; the observed

value would never be 0.704. The model simply tells us that we can expect ~70% of our outcomes to be 1 at X = 4. This is a critical point to understand for logistic regression.
Examples: