



Linear Regression

Variables

Linear regression is a model used in data science. It is used to find relationships between variables. For example, linear regression might be used to answer a question like: 'What is the relationship between hours of sleep and productivity?'

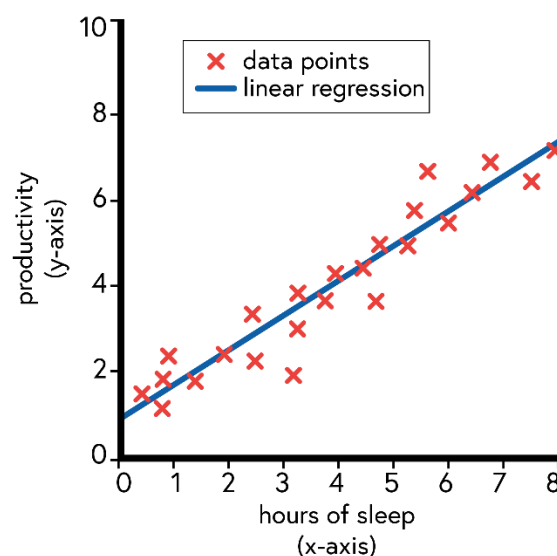
In this question, the two variables are: "hours of sleep" and "productivity". Variables can be independent or dependent.

- Independent variable: hours of sleep
- Dependent variable: productivity

The independent variable is the cause. The dependent variable is the effect. This means that the dependent variable changes as a result of the independent variable. Therefore, in the example, productivity changes as a result of how many hours of sleep a person gets.

Graphs

In a graph, the independent variable is the x-axis (the horizontal line) and the dependent variable is the y-axis (the vertical line). You can see these labelled in the graph below. The graph shows the relationship between the two variables.



The red dots are data points. These show the results of all the individual data that was collected. The blue line is the linear regression. It is sometimes called a "trendline" or a "line



of best fit". This line shows the approximate relationship between the two variables. Some of the data is above this line and some is below, but overall, the data is following this line.

Correlations

So, what does this linear regression tell us about our question: 'What is the relationship between hours of sleep and productivity?' It tells us that as the number of hours of sleep increases, the amount of productivity increases. This is called a "positive correlation". In a linear relationship, there is also "negative correlation" and "no correlation" as you can see in the graphs below.

