What is the difference between correlation and regression?

Authored by

stats writer

April 25, 2024

Correlation and regression are two statistical techniques commonly used to measure the relationship between two variables. However, they serve different purposes and have distinct characteristics.

Correlation is a measure of the strength and direction of the linear relationship between two variables. It can range from -1 to 1, where a value of 1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 indicates no correlation. Correlation does not imply causation, meaning that even if two variables are strongly correlated, it does not necessarily mean that one causes the other.

Regression, on the other hand, is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It aims to find the best-fitting line or curve that describes the relationship between the variables. Regression can help predict the value of the dependent variable based on the values of the independent variables. Unlike correlation, regression can indicate causation, as it looks at how changes in the independent variable(s) affect the dependent variable.

In summary, correlation measures the strength and direction of a relationship between two variables, while regression helps to model and predict the values of a dependent variable based on one or more independent variables.

Correlation vs. Regression: What's the Difference?

Correlation and regression are two terms in statistics that are related, but not quite the same.

In this tutorial, we'll provide a brief explanation of both terms and explain how they're similar and different.

What is Correlation?

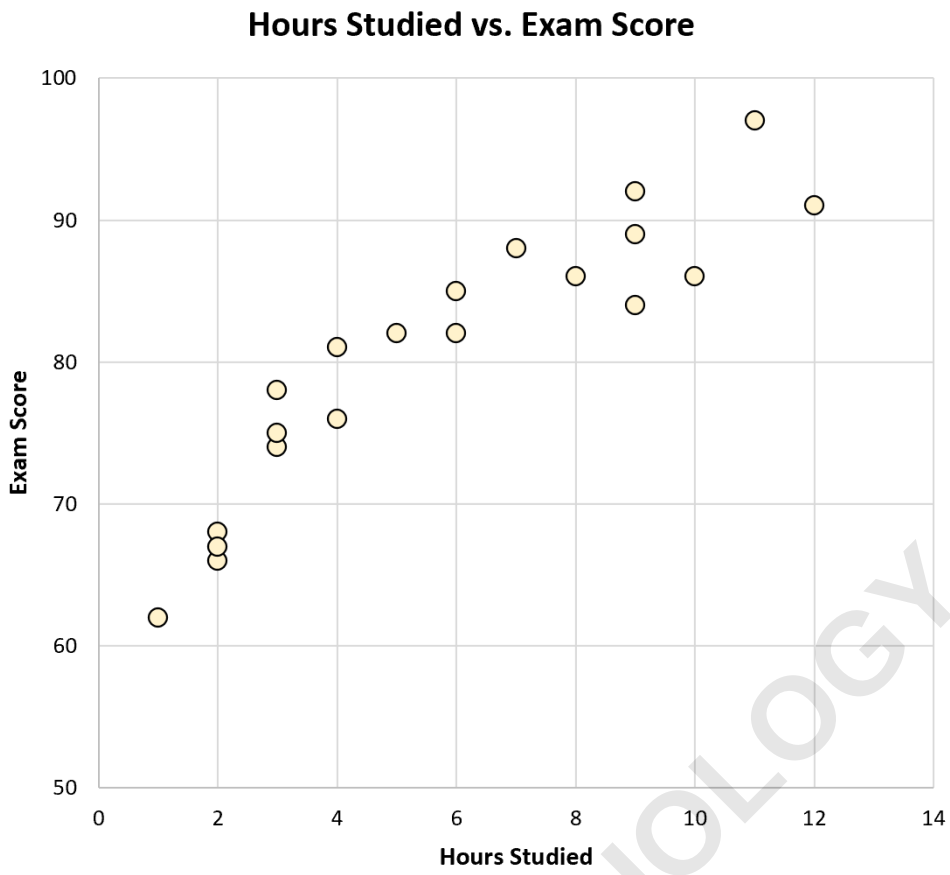
Correlation measures the linear association between two variables, *x* and *y*. It has a value between -1 and 1 where:

-1 indicates a perfectly negative linear correlation between two variables0 indicates no linear correlation between two variables1 indicates a perfectly positive linear correlation between two variables

For example, suppose we have the following dataset that contains two variables: (1) Hours studied and (2) Exam Score received for 20 different students:

Hours Studied	Exam Score
1	62
2	66
2	68
2	67
3	74
3	78
3	75
4	76
4	81
5	82
6	82
6	85
7	88
8	86
9	84
9	89
9	92
10	86
11	97
12	91

If we created a scatterplot of hours studied vs. exam score, here's what it would look like:



Just from looking at the plot, we can tell that students who study more tend to earn higher exam scores. In other words, we can visually see that there is a positive correlation between the two variables.

Using a calculator, we can find that the correlation between these two variables is r = 0.915. Since this value is close to 1, it confirms that there is a strong positive correlation between the two variables.

What is Regression?

Regression is a method we can use to understand how changing the values of the *x* variable affect the values of the *y* variable.

A regression model uses one variable, *x*, as the predictor variable, and the other variable, *y*, as the . It then finds an equation with the following form that best describes the relationship between the two variables:

? = b0 + b1x

where:

?: The predicted value of the response variableb0: The y-intercept (the value of y when x is equal to zero)b1: The regression coefficient (the average increase in y for a one unit increase in x)x: The value of the predictor variable

Hours Studied	Exam Score
1	62
2	66
2	68
2	67
3	74
3	78
3	75
4	76
4	81
5	82
6	82
6	85
7	88
8	86
9	84
9	89
9	92
10	86
11	97
12	91

Using a , we find that the following equation best describes the relationship between these two variables:

Predicted exam score = 65.47 + 2.58*(hours studied)

The way to interpret this equation is as follows:

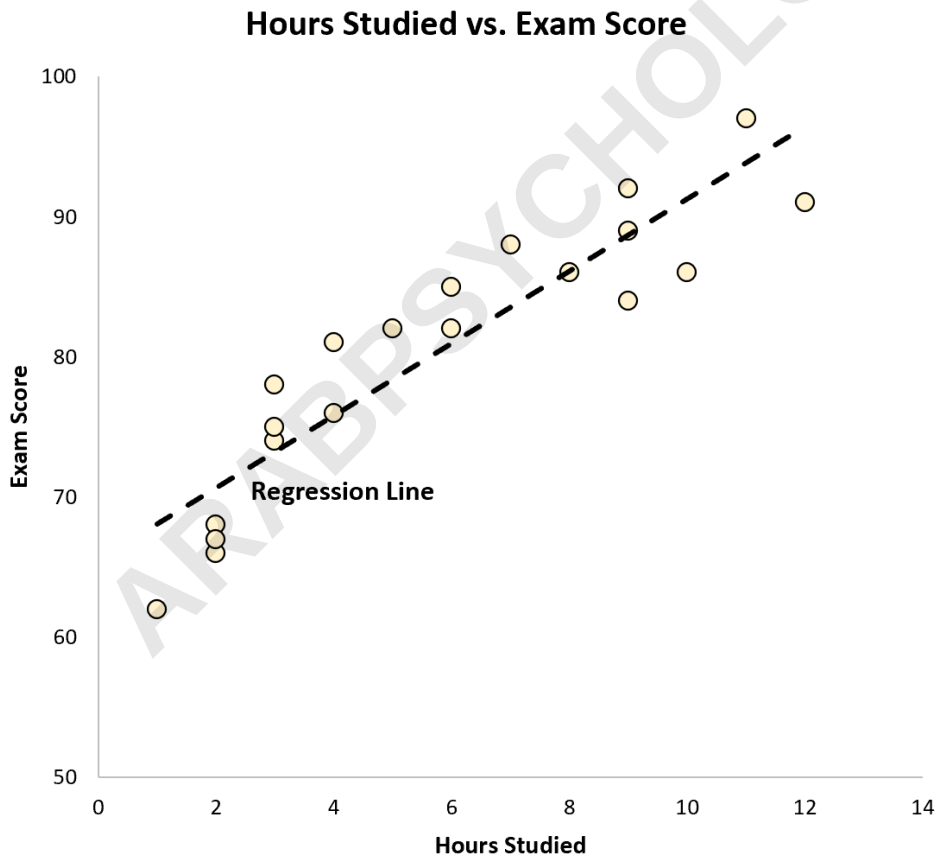
The predicted exam score for a student who studies zero hours is 65.47.The average increase in exam score associated with one additional hour studied is 2.58.

We can also use this equation to predict the score that a student will receive based on the number of hours studied.

For example, a student who studies 6 hours is expected to receive a score of 80.95:

Predicted exam score = 65.47 + 2.58*(6) = 80.95.

We can also plot this equation as a line on a scatterplot:



We can see that the regression line "fits" the data quite

well.

Recall earlier that the correlation between these two variables was r = 0.915. It turns out that we can square this value and get a number called "r-squared" that describes the total proportion of in the response variable that can be explained by the predictor variable.

In this example, r2 = 0.9152 = 0.837. This means that 83.7% of the variation in exam scores can be explained by the number of hours studied.

Correlation vs. Regression: Similarities & Differences

Here is a summary of the similarities and differences between correlation and regression:

Similarities:

Both quantify the direction of a relationship between two variables.Both quantify the strength of a relationship between two variables.

Differences:

Regression is able to show a cause-and-effect relationship between two variables. Correlation does

not do this.Regression is able to use an equation to predict the value of one variable, based on the value of another variable. Correlation does not does this.Regression uses an equation to quantify the relationship between two variables. Correlation uses a single number.

The following tutorials offer more in-depth explanations of topics covered in this post.