21BDS0340

Abhinav Dinesh Srivatsa

__/__/____

Applications of correlation and regression in real life

The applications of correlation and regression are diverse because of the fact that our age has found value in collecting data and analysing relations between them. Correlation deals with finding relations between independent or dependent collections of data. It is important to know correlation to understand how events and data are connected. Regression is the calculations of a line of best fit for the data. This useful to extrapolate or find a dependent value based on given input independent variables.

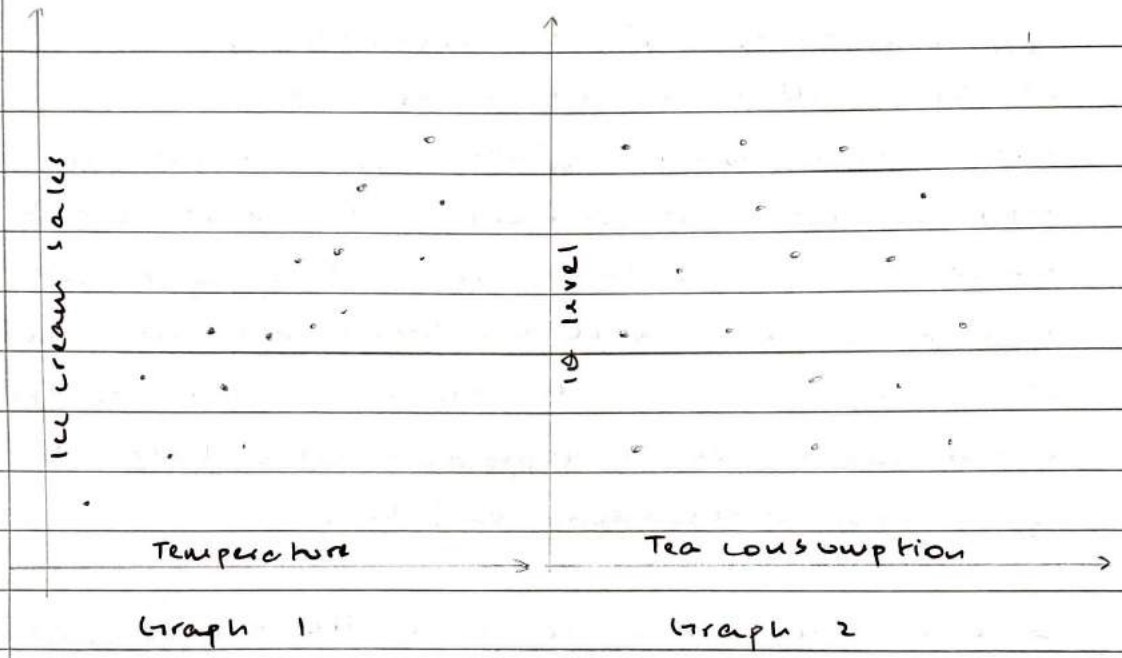
Correlation has many use cases, like data analysis, cause-effect relations and predictive machine learning models. Correlation is seeing a lot more widespread use mainly because the amount of data we are able to collect now is higher than we ever have.

Data analysis is the biggest place where correlation is used today. Data analysis is essentially the analysis and deconstruction of relations in certain data sets. Because of the massive amount of data we have at our disposal now, many companies and institutions are heavily investing in studying their users and relations to better understand cause-effect relations.

Cause-effect relations are one of the most important relations we have at our disposal. From predicting when volcanoes to knowing how exams lead to stress,

__/__/____

we have a lot of data to predict and find relations between a lot variables. A few examples with data can be as follows:



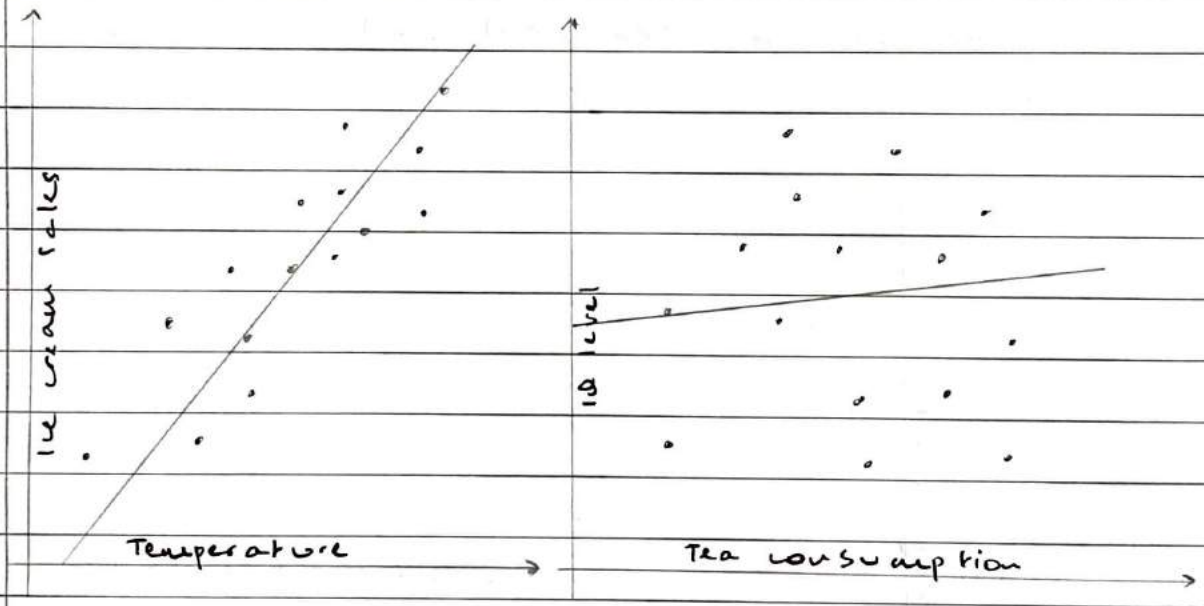
Graph 1 Graph 2

From these two graphs we can clearly see that the data in graph 1 is more predictable and that when our independent variable - temperature, we can see that as temperature increases, the sales of ice cream increase. We call this as a positive high correlation. whereas in the graph 2, the amount of tea consumed tells us almost nothing of a persons IQ level. This type of correlation is almost nothing, or almost zero.

The other major use of correlation is for predictive algorithms. Predictive machine learning models take in a lot of data and can find patterns between independent and dependent variables. This is useful in creating regression lines to predict values that are

seemingly new to the model, but has now learned how the data is related.

This leads into regression, which is a way to fit data into a line ~~of~~ or an n-1 dimensional shape. For the previous examples of the temperature vs. ice cream sales and tea consumption vs. IQ level we can draw the following regression lines:



Graph 3 Graph 4

We can clearly see that in graph 3, the line of best fit actually fits well. If we try to find the ice cream sales with respect to a new temperature, we should get an accurate value. whereas in the graph 4, we did not get a good line of best fit. This difference in quality in lines of best fit can be explained by the correlation in data. A high value, or extremely low value ~ -1, will tend to produce better lines of fit. A low correlation or around zero will not produce a good line of best fit.

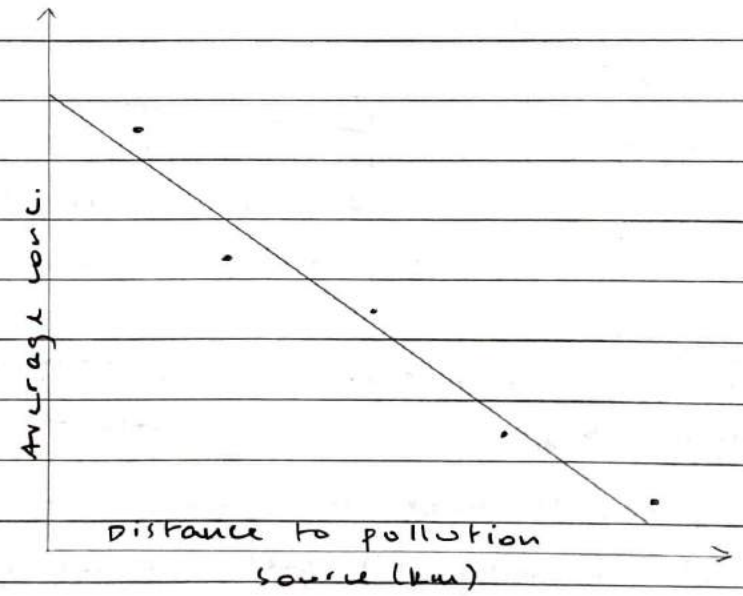
__/__/____

Linear regression can help us solve problems like the following:

Water pollution:

Distance to pollution source (kms)	2	4	6	8	10
Average concentration	11.5	10.2	10.3	9.68	9.32

From looking at the data, we can clearly see that as the distance increases, the average concentration decreases. From this, we can say that the data is highly negatively correlated. We can plot a graph of the data and add a line of regression.



From the graph, we can see that the data is indeed highly negatively correlated. By actually finding the equation of the line:

① $\Sigma y = m \Sigma x + cn$

② $\Sigma xy = m \Sigma x^2 + c \Sigma x$

where x = distance, y = concentration

By solving the values:

$\Sigma x = 30$

$\Sigma y = 51$

$\Sigma xy = 296.24$

$\Sigma x^2 = 220$

① $51 = 30m + 5c$

② $296.24 = 220m + 30c$

By solving these equations:

$m = -0.244$

$c = 11.664$

By creating an equation: $y = -0.244x + 11.664$, we can find values of average concentration by giving an input distance x.

For example at $x = 12$, $y = 8.736$. From this, our regression line tells us that at a distance 12 km from the pollution source, the average concentration will be 8.736.

The big thing that is trending today is the field of machine learning. In a simplified form we can predict anything by knowing all the variables that form its relations. Linear regression deals with only 2 dimensional, while machine learning applies this to multiple dimensions. By computing a best fit of the data, an algorithm can accurately predict an outcome that is dependent on a list of variables.

One simple application is to find the quality of wine based on factors. These factors include: alcohol content, malic acid, ash, alkalinity of ash, magnesium, total phenols, flavanoids, phenols, proanthocyanins, color, hue, OD280/OD315, proline. These factors can decide whether wine is wine or not. Because of the number of independent variables, it can be hard to visualise the data, but a machine learning model can plot and find relations between the data no matter the dimensions. To conclude the model and check for output, simply give it your data and it'll tell you what wine if it is.

From these applications, it is clear that correlation and regression ~~theory~~ have massive uses and implications on shaping the world around us. From machine learning to deciphering relations, analysing the large amounts of data we have nowadays is an essential field to improve and grow society and applications.