1 Understanding Regression

From a young age, we learn to associate being truthful and kind to others as good traits and lying as a bad trait. We draw from experience that expressing good traits are usually rewarding in nature whereas bad traits do not. We understand that there exists a relationship between our actions and their consequences. Finding out this relationship between two events; an action and a consequence, is at the core of regression.

When given two sets of data, a regression analysis helps us understand the relationship between those two. For example, say you're a fresh college sophomore looking for a place to live. When browsing through places to live you notice a clear trend. The closer your address to your campus, the more expensive the rent.

Distance from campus (m)	Cost of a room (Rs.)
400	8000
700	7100
1000	6200
1300	5300
1600	4400
1900	3500

Figure 1: Table showing distance from campus and cost of room

If we plot this into a graph, it looks like this:

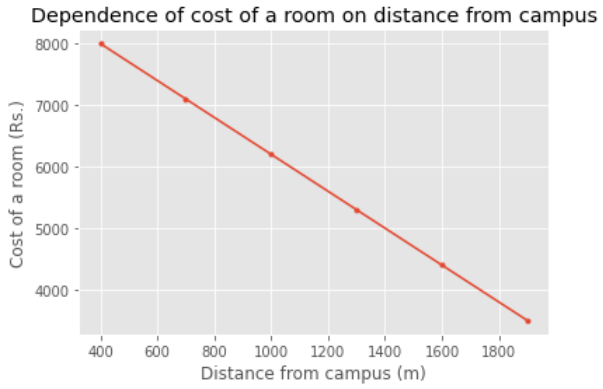

Figure 2: Graph showing distance from campus vs cost of room

Here, we see a clear line that fits all the points perfectly. The slope of the line is -3 and the x-intercept is 9200.

However data in real life looks something like this:

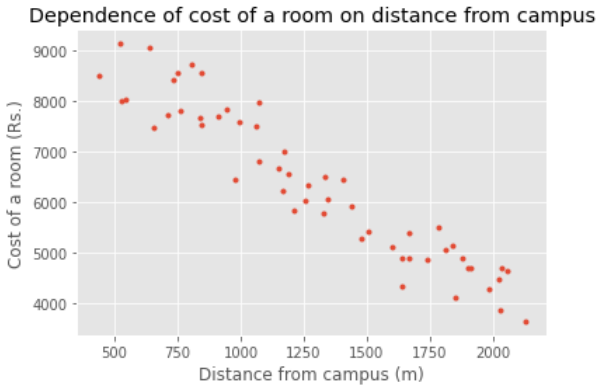

Figure 3: Distance from campus vs cost of room in real life

It is quite clear that finding a straight line that fits all the point is impossible to do. The best we can do is find a line that most satisfies the condition such that the distance from a point to the line is as minimum as possible. Formalizing this dependence in a form of a clear equation is regression.

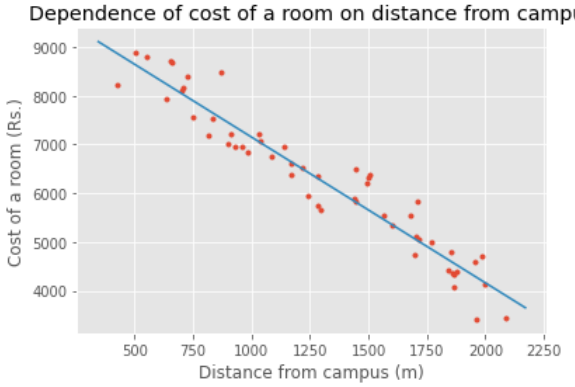

Figure 4: A linear regression line through given points

A more formal definition states: **a regression analysis is a set of statistical processes for estimating the relationship between a dependent variable and one or more independent variables.**

There exists many kind of regressions. In this reading, we will be looking at the two most common: **Linear** and **Logistic** regression

2 Linear Regression

Linear regression is the simplest form of regression. The equation for linear regression resembles one of the first equations we learn: the equation of a line in slope-intercept form:

$$y = m.x + c$$

(This is obvious because, in regression, we are trying to fit a line that satisfies a set of points as closely as possible)
In linear regression analysis, we write the equation a little differently:

$$y = \theta_0 + \theta_1.x$$

Here, the values for θ_0 and θ_1 are called biases. Depending on how many independent variables, we add additional theta components to our existing equation. For example, for three independent variables that influence a dependent variable, the equation looks like:

$$y = \theta_0 + \theta_1.x_1 + \theta_2.x_2 + \theta_3.x_3$$

In matrix form, the equation looks like:

$$y = \theta T.X$$

, where θ is the row matrix containing all the values of biases [θ_0, θ_1, θ_2, θ_3] and X is the row matrix containing all the values of input data [x_0, x_1, x_2, x_3] and $x_0 = 1$.

Linear regression is useful for predicting the continuous dependent variable with the help of independent variable(s). The value from a linear regression can exist from any range.

3 Logistic Regression

Logistic regression is another popular form of regression. It is a non-linear function useful in classification problems and outputs a probability of occurrence rather than a concrete value. The equation for logistic regression is as follows:

$$log\left[\frac{y}{1-y}\right] = \theta_0 + \theta_1.x_1 + \theta_2.x_2 + \theta_3.x_3 + ...$$

Similar to linear regression, the values for θ_0 and θ_1 are called biases. Depending on how many independent variables, we add additional theta components to our existing equation.

Logistic regression is useful for predicting where the probabilities between two classes is required. To solve a logistic regression problem, we require additional concept such as gradient descent. The value from a logistic regression can exist only exist for a range between (0, 1).