Regression

Tushar B. Kute,
http://tusharkute.com

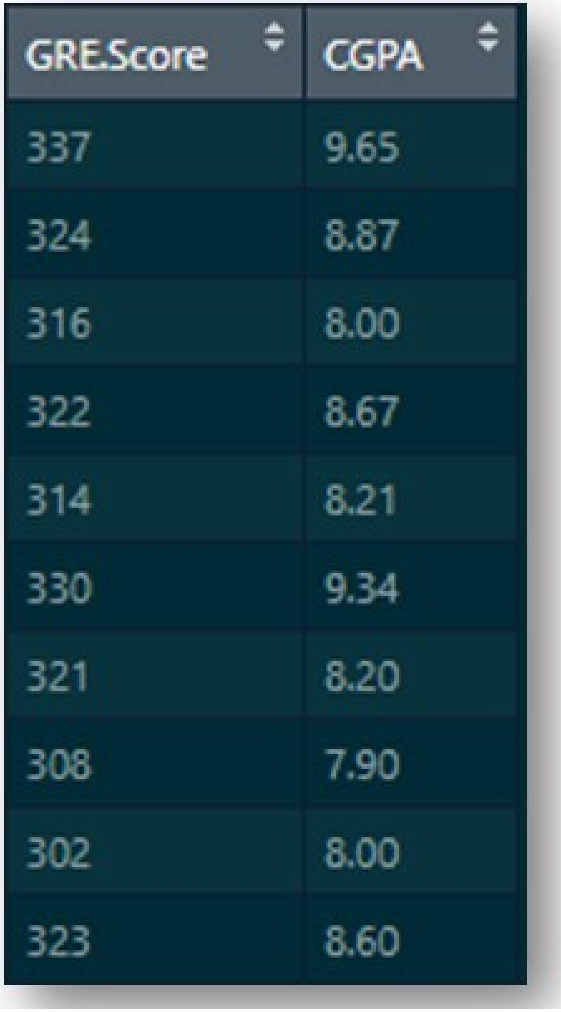
Regression?

- Regression analysis is a statistical method that helps us to analyse and understand the relationship between two or more variables of interest.
- The process that is adapted to perform regression analysis helps to understand which factors are important, which factors can be ignored and how they are influencing each other.

Regression?

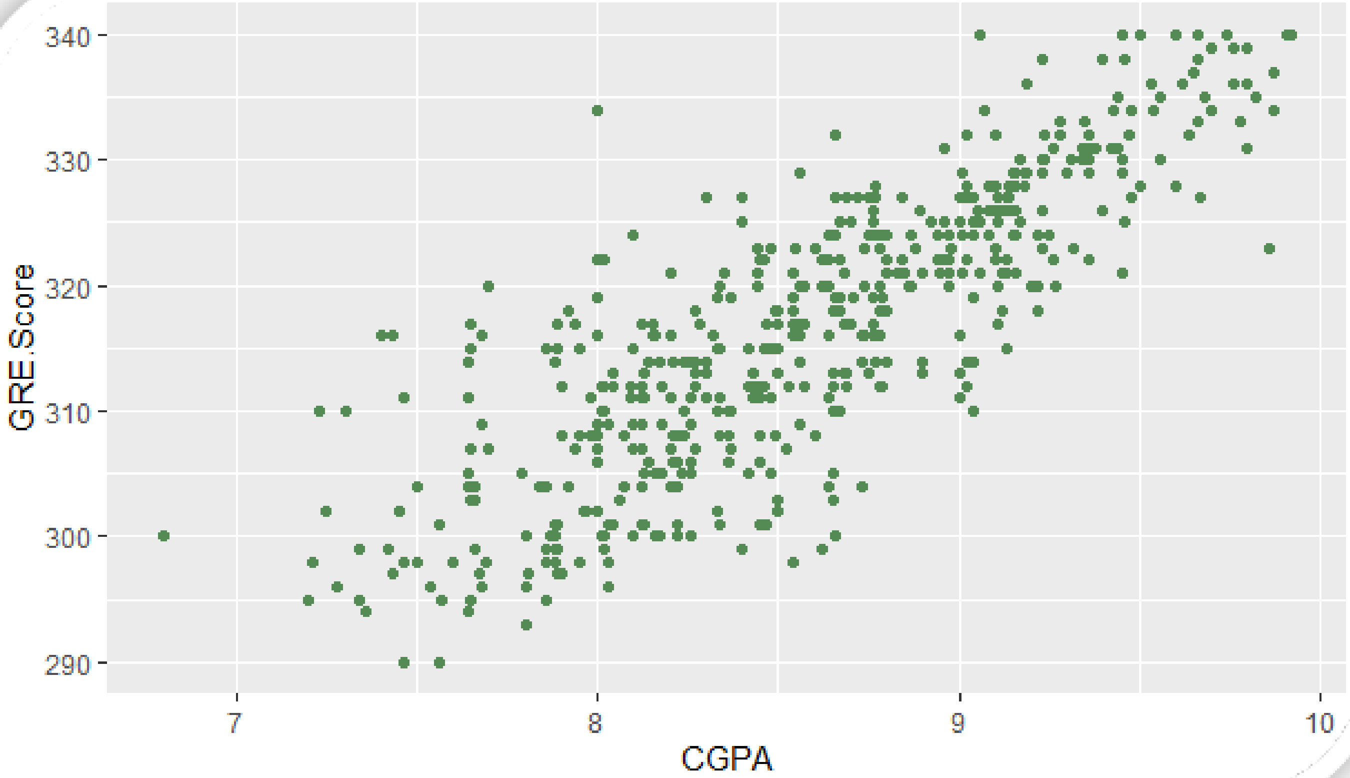
- For the regression analysis is be a successful method, we understand the following terms:
  - Dependent Variable: This is the variable that we are trying to understand or forecast.
  - Independent Variable: These are factors that influence the analysis or target variable and provide us with information regarding the relationship of the variables with the target variable.

Example:



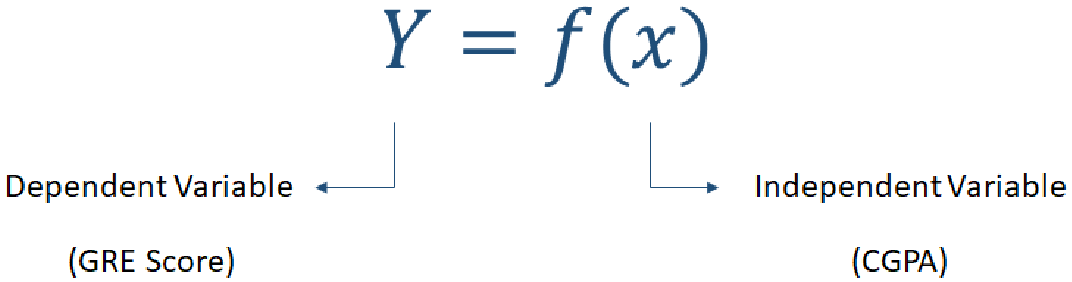
GRE.Score	CGPA
337	9.65
324	8.87
316	8.00
322	8.67
314	8.21
330	9.34
321	8.20
308	7.90
302	8.00
323	8.60

Example:



Regression

- In regression, we normally have one dependent variable and one or more independent variables.
- Here we try to “regress” the value of dependent variable “Y” with the help of the independent variables.
- In other words, we are trying to understand, how does the value of ‘Y’ change w.r.t change in ‘X’.



$$Y = f(x)$$

Dependent Variable

(GRE Score)

Independent Variable

(CGPA)

Regression: Simple Applications

- Marks scored by students based on number of hours studied (ideally)- Here marks scored in exams are independent and the number of hours studied is independent.
- Predicting crop yields based on the amount of rainfall- Yield is a dependent variable while the measure of precipitation is an independent variable.
- Predicting the Salary of a person based on years of experience- Therefore, Experience becomes the independent while Salary turns into the dependent variable.

Terminologies

- Outliers
  - Suppose there is an observation in the dataset that has a very high or very low value as compared to the other observations in the data, i.e. it does not belong to the population, such an observation is called an outlier.
  - In simple words, it is an extreme value. An outlier is a problem because many times it hampers the results we get.

Terminologies

- Multicollinearity
  - When the independent variables are highly correlated to each other, then the variables are said to be multicollinear.
  - Many types of regression techniques assume multicollinearity should not be present in the dataset.
  - It is because it causes problems in ranking variables based on its importance, or it makes the job difficult in selecting the most important independent variable.

Terminologies

- Heteroscedasticity
  - When the variation between the target variable and the independent variable is not constant, it is called heteroscedasticity.
  - Example-As one’s income increases, the variability of food consumption will increase.
  - A poorer person will spend a rather constant amount by always eating inexpensive food; a wealthier person may occasionally buy inexpensive food and at other times, eat expensive meals.
  - Those with higher incomes display a greater variability of food consumption.

Assumptions

- The variables should be measured at a continuous level. Examples of continuous variables are time, sales, weight and test scores.
- Use a scatterplot to find out quickly if there is a linear relationship between those two variables.
- The observations should be independent of each other (that is, there should be no dependency).
- Your data should have no significant outliers.
- Check for homoscedasticity — a statistical concept in which the variances along the best-fit linear-regression line remain similar all through that line.
- The residuals (errors) of the best-fit regression line follow normal distribution.

Types of Regression

- Linear Regression
- Multiple Regression
- Logistic Regression
- Polynomial Regression
- Regularized Models
  - Ridge Regression
  - Lasso Regression
  - ElasticNet Regression
- Outlier Based Model
  - RANSAC

Linear Regression

- The simplest of all regression types is Linear Regression where it tries to establish relationships between Independent and Dependent variables.
- The Dependent variable considered here is always a continuous variable.
- Linear Regression is a predictive model used for finding the linear relationship between a dependent variable and one or more independent variables.

$$Y = a + bx$$

Dependent Variable
is continuous

Linear Regression

- In Linear Regression these two variables are related through an equation, where exponent (power) of both these variables is 1.
- Mathematically a linear relationship represents a straight line when plotted as a graph.
- A non-linear relationship where the exponent of any variable is not equal to 1 creates a curve.
- The general mathematical equation for a linear regression is –

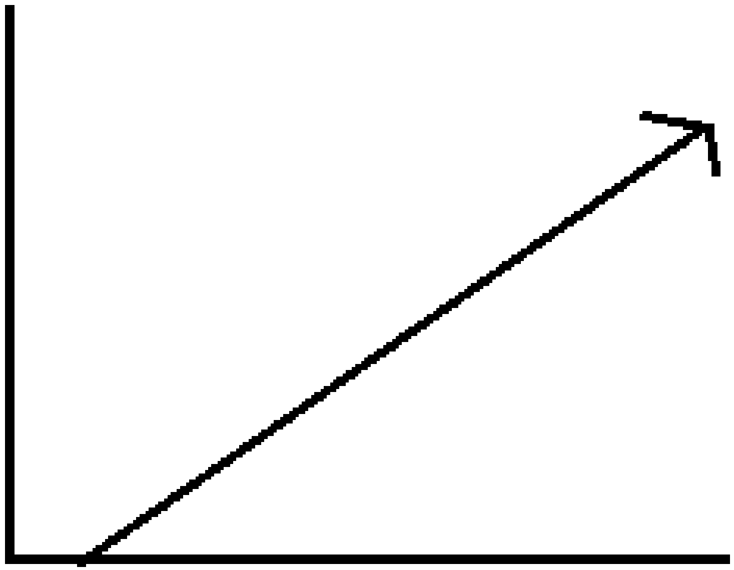
```
y = mx + c
```

y is the response variable.

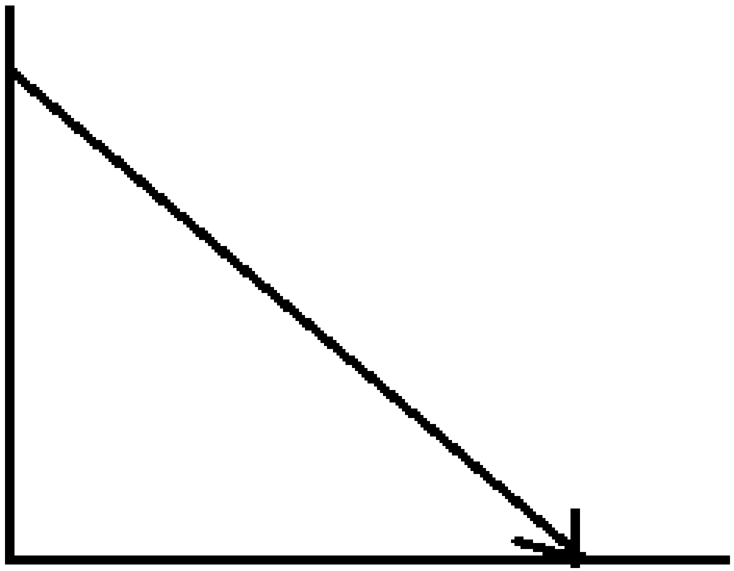
x is the predictor variable.

m and c are constants which are called the coefficients.

Linear Regression



Positive Linear Relationship

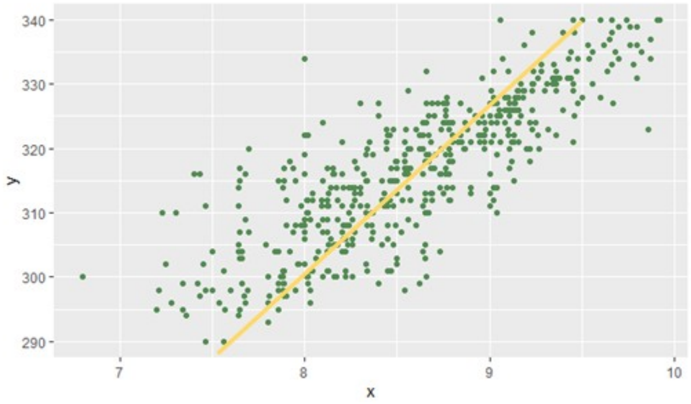


Negative Linear Relationship

Linear Regression

- Here, ‘Y’ is our dependent variable, which is a continuous numerical and we are trying to understand how does ‘Y’ change with ‘X’.
- So, if we are supposed to answer, the above question of “What will be the GRE score of the student, if his CCGPA is 8.32?” our go to option should be linear regression.

$$GRE = 261 + 6.8CGPA \rightarrow GRE = 261 + 6.8(8.32) \rightarrow GRE = 317.57$$



Simple Linear Regression

- As the model is used to predict the dependent variable, the relationship between the variables can be written in the below format.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- Where,
  - Y_i – Dependent variable
  - β_0 — Intercept
  - β_1 – Slope Coefficient
  - X_i – Independent Variable
  - ε_i – Random Error Term

Simple Linear Regression

- The main factor that is considered as part of Regression analysis is understanding the variance between the variables. For understanding the variance, we need to understand the measures of variation.
  - SST = total sum of squares (Total Variation)
    - Measures the variation of the Y i values around their mean Y
  - SSR = regression sum of squares (Explained Variation)
    - Variation attributable to the relationship between X and Y
  - SSE = error sum of squares (Unexplained Variation)
    - Variation in Y attributable to factors other than X

Steps to establish Linear Regression

- A simple example of regression is predicting weight of a person when his height is known. To do this we need to have the relationship between height and weight of a person.
- The steps to create the relationship is –
  - Carry out the experiment of gathering a sample of observed values of height and corresponding weight.
  - Create the object of linear regression.
  - Train the algorithm with dataset of input and output.
  - Get a summary of the relationship model to know the average error in prediction. Also called residuals.
  - To predict the weight of new persons, use the predict() function.

Example – Input data

- Below is the sample data representing the observations –

  # Values of height

  151, 174, 138, 186, 128, 136, 179, 163, 152, 131

  # Values of weight.

  63, 81, 56, 91, 47, 57, 76, 72, 62, 48

Least Square Regression

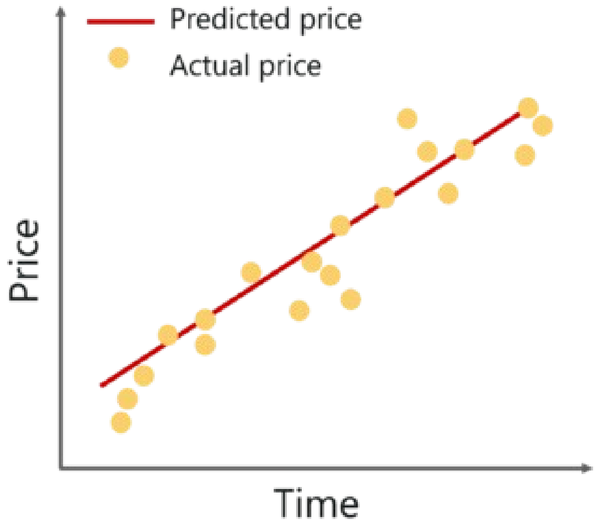
- The least-squares regression method is a technique commonly used in Regression Analysis.
- It is a mathematical method used to find the best fit line that represents the relationship between an independent and dependent variable.
- To understand the least-squares regression method lets get familiar with the concepts involved in formulating the line of best fit.

What is line of best fit ?

- Line of best fit is drawn to represent the relationship between 2 or more variables. To be more specific, the best fit line is drawn across a scatter plot of data points in order to represent a relationship between those data points.
- Regression analysis makes use of mathematical methods such as least squares to obtain a definite relationship between the predictor variable (s) and the target variable.
- The least-squares method is one of the most effective ways used to draw the line of best fit. It is based on the idea that the square of the errors obtained must be minimized to the most possible extent and hence the name least squares method.

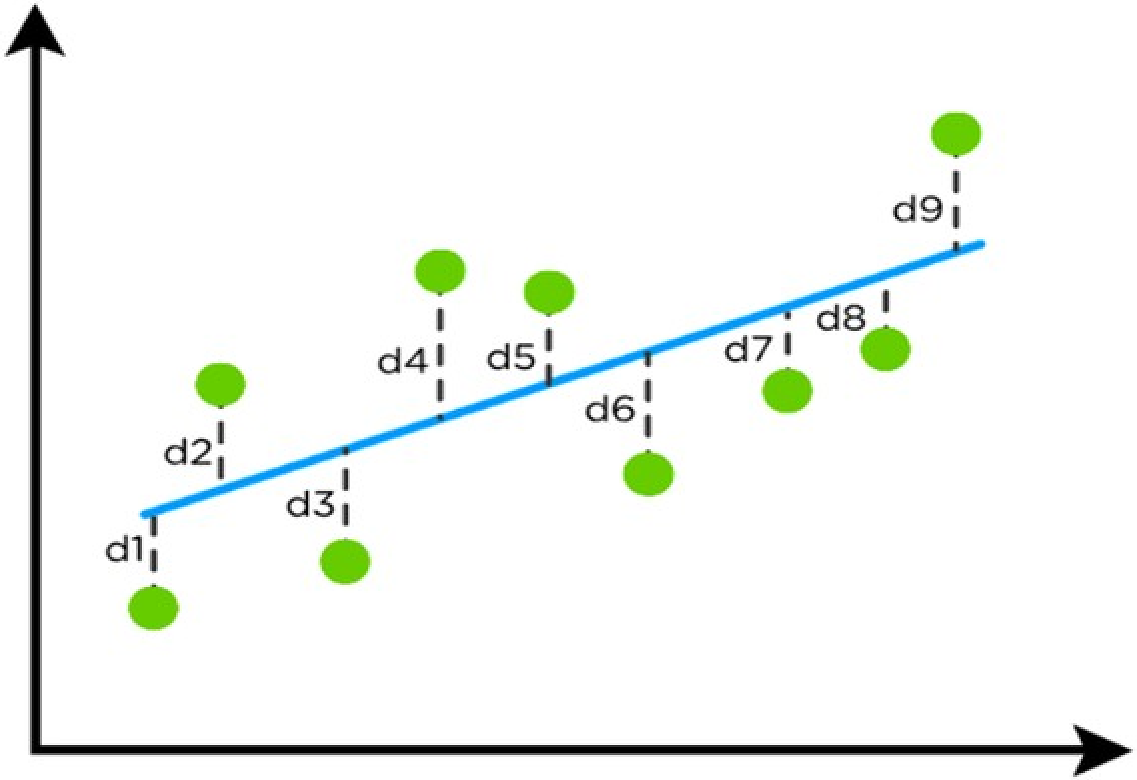
Visualizing

- If we were to plot the best fit line that shows the depicts the sales of a company over a period of time, it would look something like this:



- Notice that the line is as close as possible to all the scattered data points. This is what an ideal best fit line looks like.

Visualizing



$D = d1^2 + d2^2 + d3^2 + d4^2 + d5^2 + d6^2 + d7^2 + d8^2 + d9^2$

The regression line (blue) has the least value of D

Calculate line of best fit

- To start constructing the line that best depicts the relationship between variables in the data, we first need to get our basics right. Take a look at the equation below:

$$y = mx + c$$

- Surely, you've come across this equation before. It is a simple equation that represents a straight line along 2 Dimensional data, i.e. x-axis and y-axis. To better understand this, let's break down the equation:

  y: dependent variable

  m: the slope of the line

  x: independent variable

  c: y-intercept

Calculate line of best fit

- Step 1: Calculate the slope ‘m’ by using the following formula:

$$m = \frac{n\sum xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$$

- Step 2: Compute the y-intercept (the value of y at the point where the line crosses the y-axis):

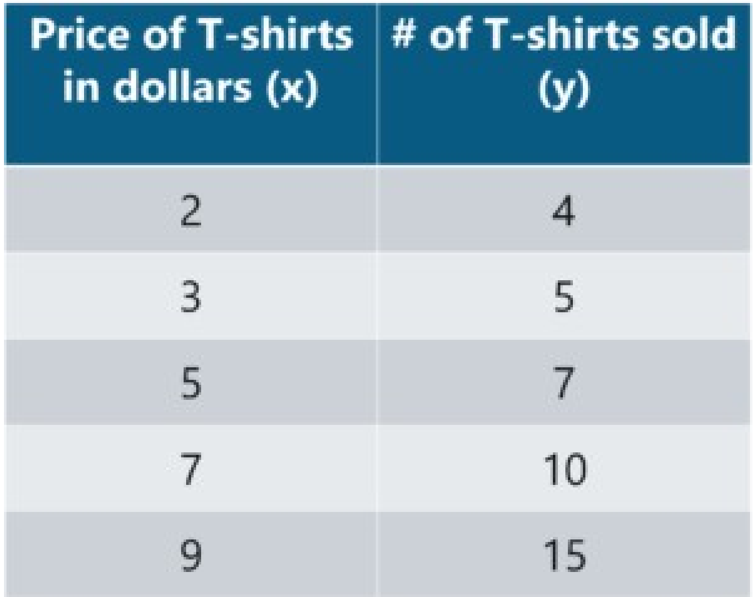
$$\mathrm{c} = y - mx$$

- Step 3: Substitute the values in the final equation:

$$y = mx + c$$

Example

- Consider an example. Tom who is the owner of a retail shop, found the price of different T-shirts vs the number of T-shirts sold at his shop over a period of one week.
- He tabulated this like shown below:



Price of T-shirts in dollars (x)	# of T-shirts sold (y)
2	4
3	5
5	7
7	10
9	15

Step-1

- Let us use the concept of least squares regression to find the line of best fit for the above data.
- Step 1: Calculate the slope ‘m’ by using the following formula:

$$m = \frac{n\sum xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$$

- After you substitute the respective values, m = 1.518 approximately.

Step-2

- Step 2: Compute the y-intercept value

$$c = y - mx$$

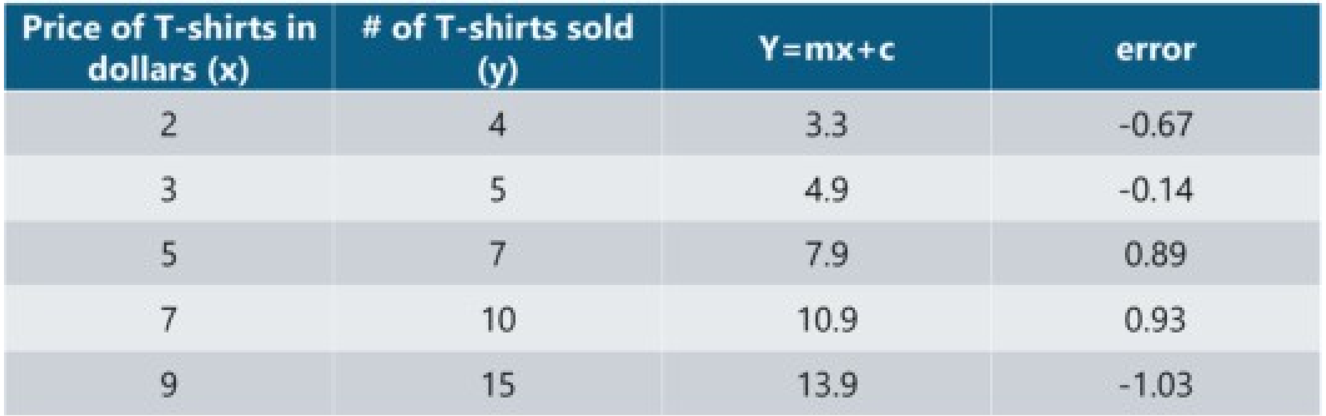
- After you substitute the respective values, c = 0.305 approximately.

Step-3

- Step 3: Substitute the values in the final equation

$$y = mx + c$$

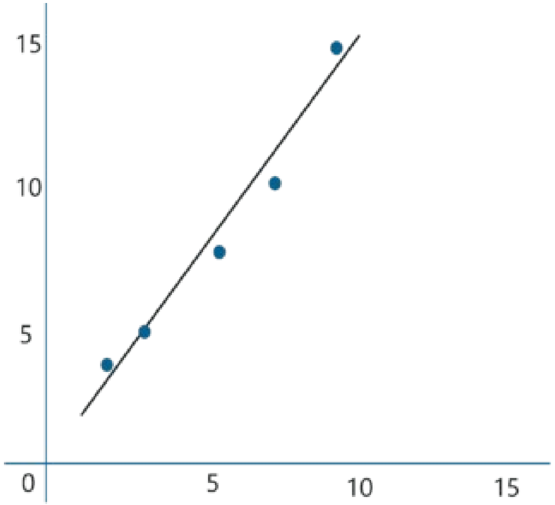
- Once you substitute the values, it should look something like this:



Price of T-shirts in dollars (x)	# of T-shirts sold (y)	Y=mx+c	error
2	4	3.3	-0.67
3	5	4.9	-0.14
5	7	7.9	0.89
7	10	10.9	0.93
9	15	13.9	-1.03

Predicting

- Let’s construct a graph that represents the y=mx + c line of best fit:



- Now Tom can use the above equation to estimate how many T-shirts of price $8 can he sell at the retail shop.

y = 1.518 x 8 + 0.305 = 12.45 T-shirts

Real Life Example #1

- Businesses often use linear regression to understand the relationship between advertising spending and revenue.
- For example, they might fit a simple linear regression model using advertising spending as the predictor variable and revenue as the response variable. The regression model would take the following form:

$$\text{revenue} = \beta 0 + \beta 1(\text{ad spending})$$

- The coefficient β0 would represent total expected revenue when ad spending is zero.

Real Life Example #2

- Medical researchers often use linear regression to understand the relationship between drug dosage and blood pressure of patients.
- For example, researchers might administer various dosages of a certain drug to patients and observe how their blood pressure responds.
- They might fit a simple linear regression model using dosage as the predictor variable and blood pressure as the response variable. The regression model would take the following form:

$$\text{blood pressure} = \beta 0 + \beta 1(\text{dosage})$$

Real Life Example #3

- Agricultural scientists often use linear regression to measure the effect of fertilizer and water on crop yields.
- For example, scientists might use different amounts of fertilizer and water on different fields and see how it affects crop yield.
- They might fit a multiple linear regression model using fertilizer and water as the predictor variables and crop yield as the response variable. The regression model would take the following form:

  crop yield = β0 + β1(amount of fertilizer) + β2(amount of water)

Real Life Example #4

- Data scientists for professional sports teams often use linear regression to measure the effect that different training regimens have on player performance.
- For example, data scientists in the NBA might analyze how different amounts of weekly yoga sessions and weightlifting sessions affect the number of points a player scores.
- They might fit a multiple linear regression model using yoga sessions and weightlifting sessions as the predictor variables and total points scored as the response variable. The regression model would take the following form:

  points scored = β0 + β1(yoga sessions) + β2(weightlifting sessions)

Useful web resources

- www.mitu.co.in
- www.scikit-learn.org
- www.towardsdatascience.com
- www.medium.com
- www.analyticsvidhya.com
- www.kaggle.com
- www.stephacking.com
- www.github.com

Thank you

@mitu_skillologies

/mITuSkillologies

@mitu_group

/company/mitu-skillologies

MITUSkillologies

Web Resources
https://mitu.co.in
http://tusharkute.com

contact@mitu.co.in

tushar@tusharkute.com