STATISTICS FOR DATA SCIENCE

Power Test &
Simple Linear Regression

Dr. Karthiyayini
Department of Science and Humanities

STATISTICS FOR DATA SCIENCE

Unit 5 : Power Test & Simple Linear Regression

Session : 5 (Continued Session)

Sub Topic : Correlation & Regression Analysis

Dr. Karthiyayini
Department of Science & Humanities

Regression Analysis

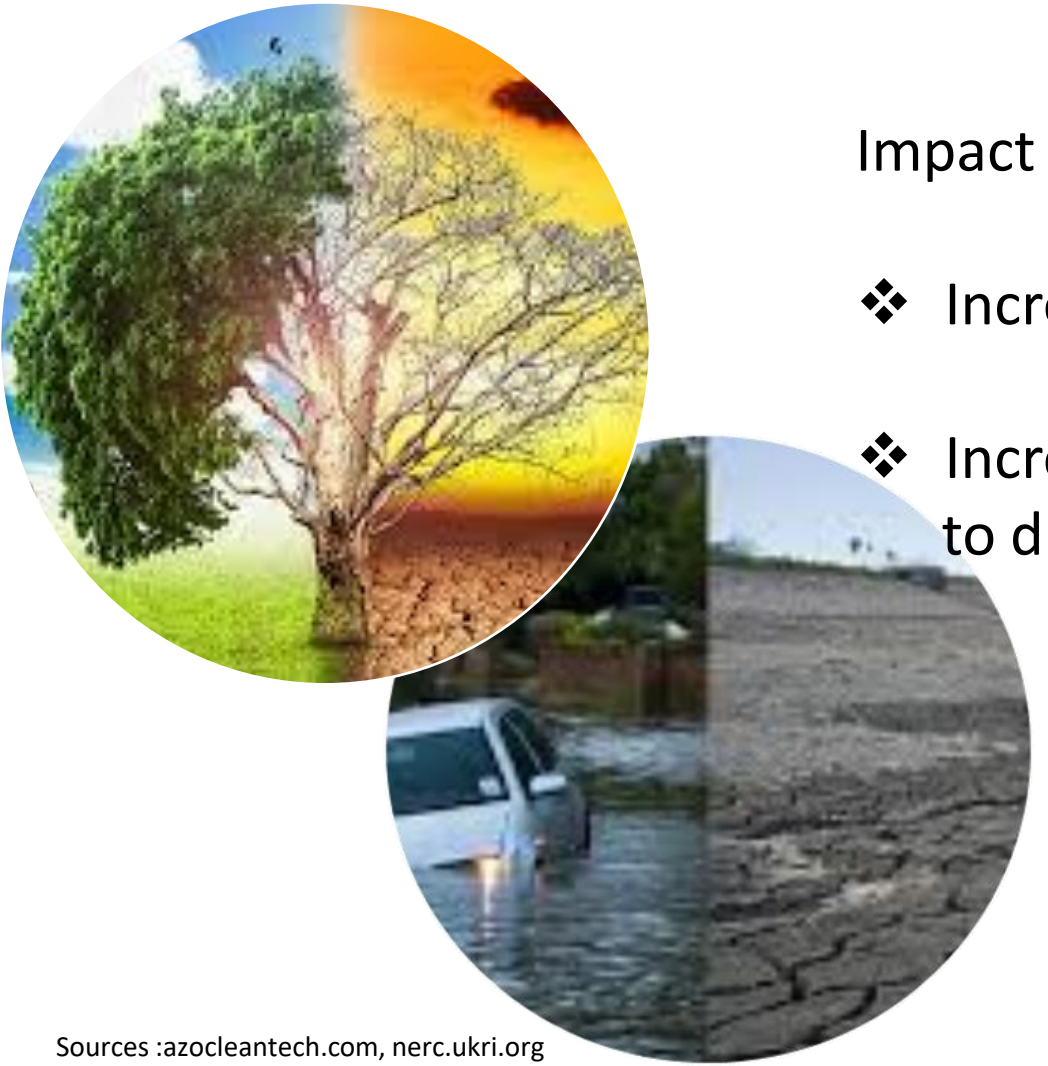
❖Regression Analysis is basically the study of a set of data to make the best guess or some kind of prediction.

- For Example : By studying a data which provides information of how much you eat and how much you weigh, you can conclude that there exists a relationship between the two.

- Regression analysis can help you to quantify that and can help you to predict how much you will weigh in 10 years time if you continue to put on weight at the same rate.

STATISTICS FOR DATA SCIENCE

Prediction of Floods / Droughts



Impact of Global warming :

- Increase in rainfall resulting in Floods
- Increase in amount of dry land leading to droughts

Sources :azocleantech.com, nerc.ukri.org

Impact of Global Warming

Global Warming in Wet areas

- Evaporation of water from land and sea
- More rainfall
- Increase in Floods

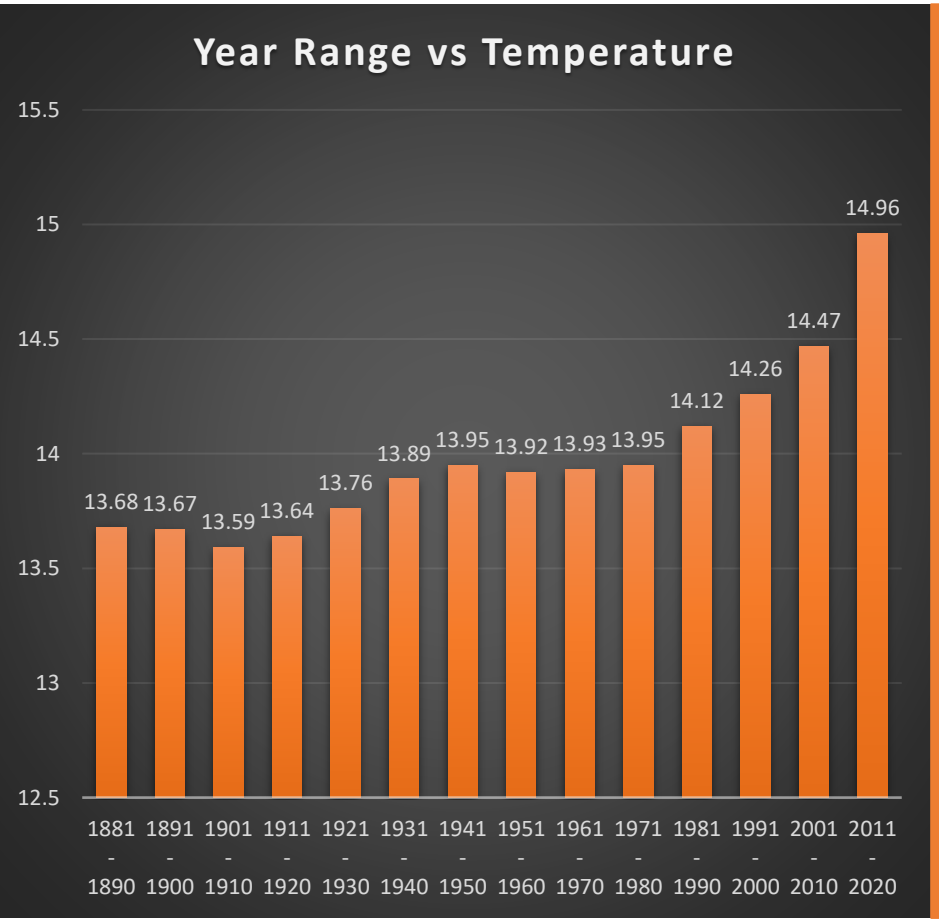
Global Warming in Dry areas

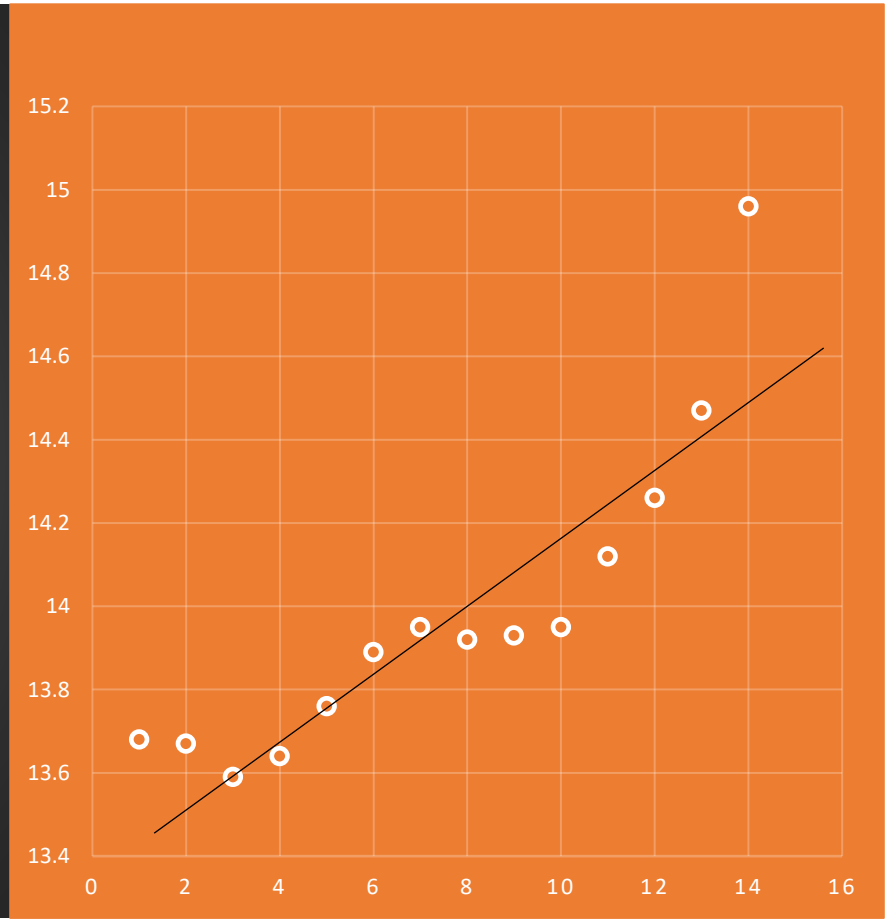
- Increase in evapotranspiration of water from land , water surfaces and plants
- Dry areas become drier
- Increase in droughts

Hence, the impact of global warming is that wet regions are likely to get wetter and the already dry regions are likely to get drier leading to an increase in floods and droughts respectively.

STATISTICS FOR DATA SCIENCE

Impact of Global warming : Graphs!!!



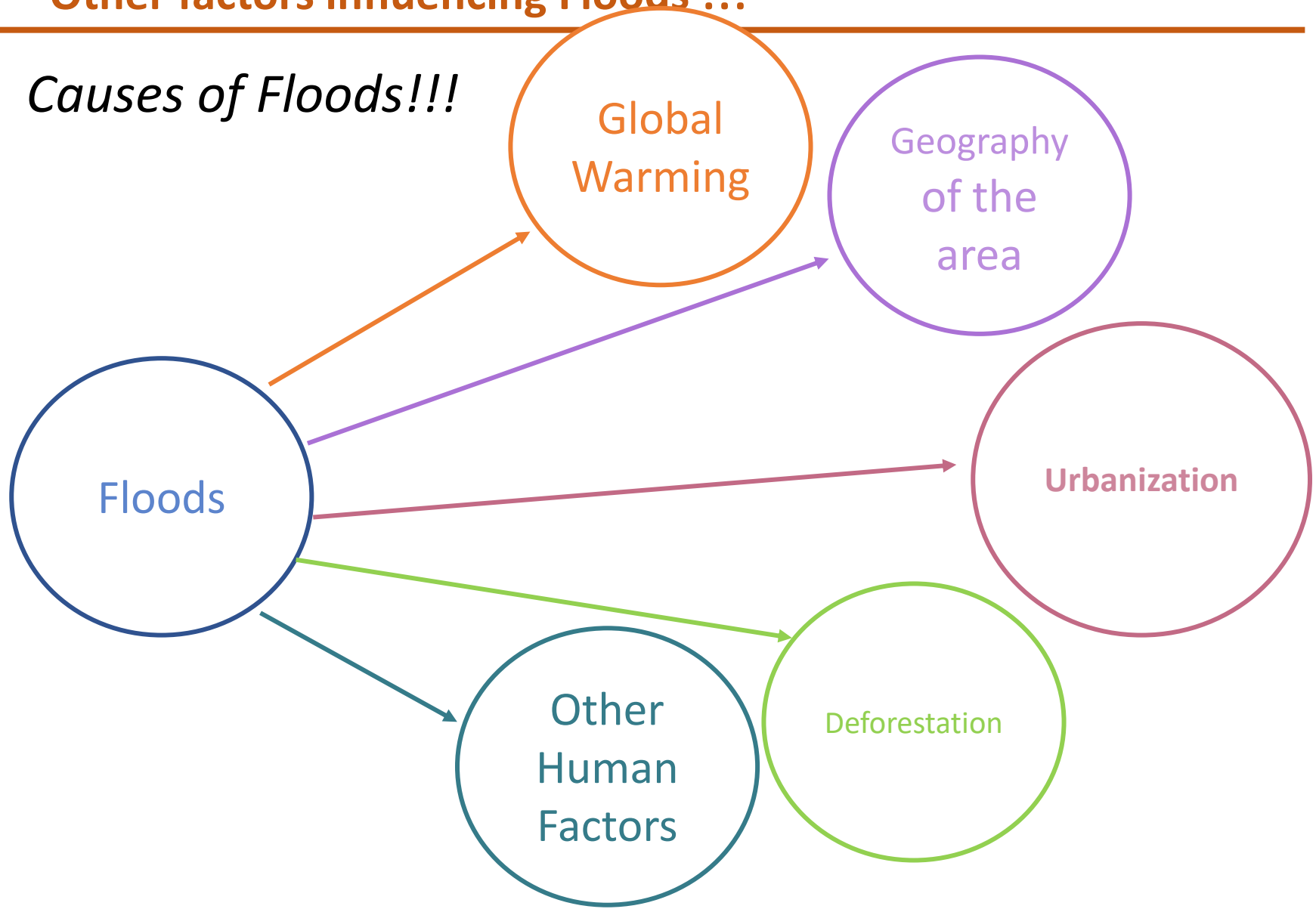


- Suppose we have the data which contains information about the amount of rainfall, year-wise.
- The using a bar chart for the data we can analyze and observe the trend of year-wise rainfall.
- For further analysis we can plot the data on a scatter plot and draw a line which passes through all the data points.
- This line is called as the Least squares line.
- The least squares line we can predict the rainfall for a given year.

STATISTICS FOR DATA SCIENCE

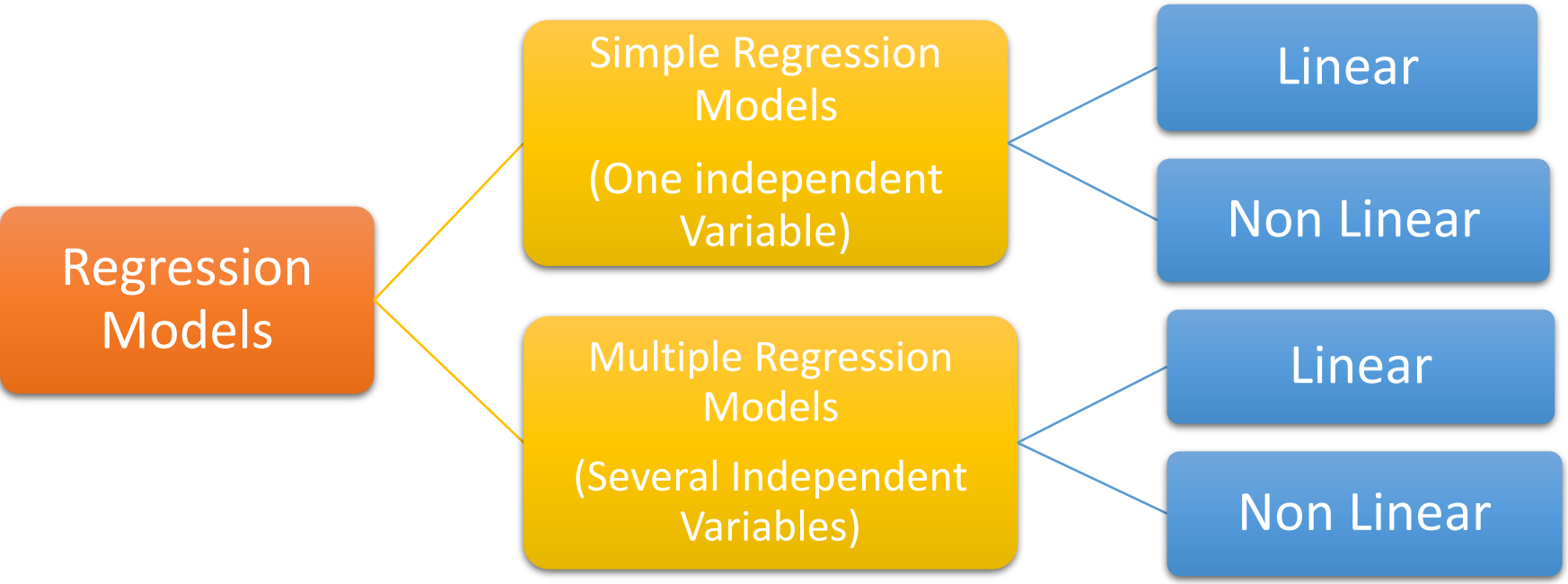
Other factors influencing Floods !!!

Causes of Floods!!!



- The other factors that could cause floods are Geography of the area, Urbanization Deforestation and other human factors like restriction of waterways, dumping of mineral waste in water beds, mining etc.
- If the analysis is related to the impact of only one factor, then it is called as Simple Regression Analysis.
- If the analysis is related to several factors, then it is called as Multiple Regression Analysis.

Regression Analysis – A Broad Classification



STATISTICS FOR DATA SCIENCE

Regression Analysis

- In statistical modeling, **regression analysis** is a set of statistical processes for estimating the relationships between a dependent variable and one or more independent variables
- It is a way of mathematically sorting out which of those variables indeed have an impact
- Which factors matter most ?
- Which can we ignore ?
- How do the factors interact with each other?
- And most importantly, how certain are we about all these factors?

STATISTICS FOR DATA SCIENCE

Some Inputs !!!

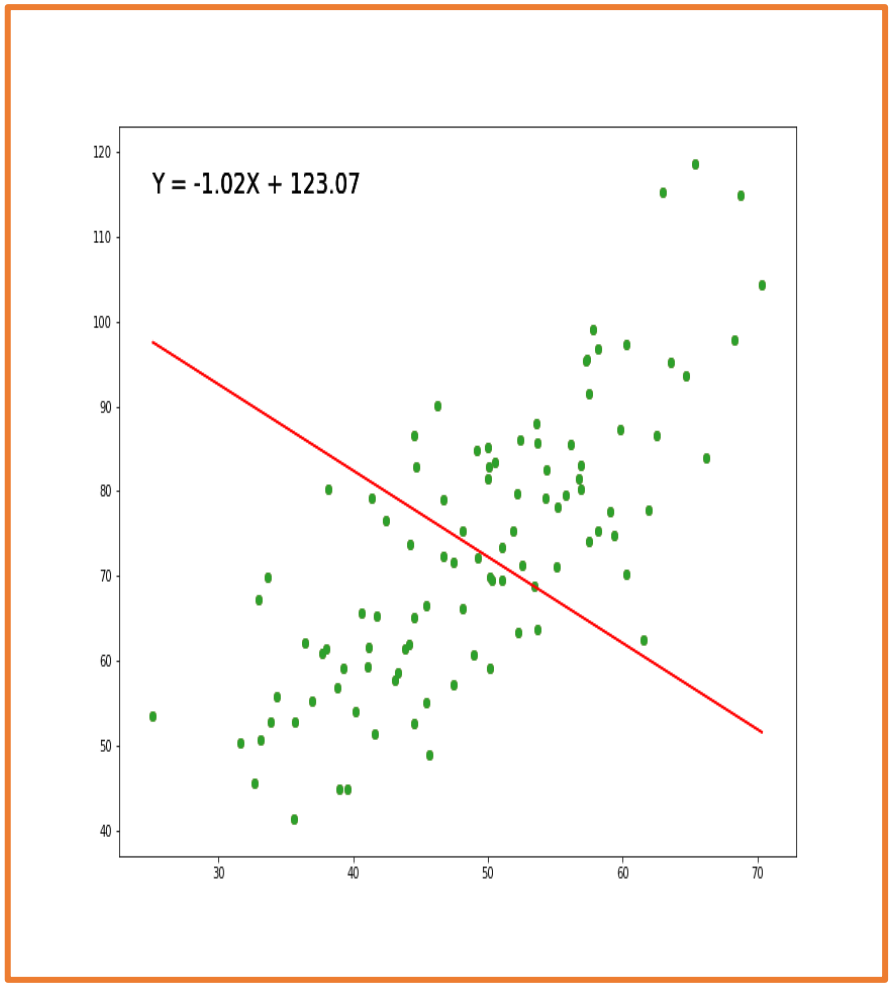
- Regression analysis is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning.
- In some situations regression analysis can be used to infer causal relationships between the independent and dependent variables.
- The term "regression" was coined by Francis Galton in the nineteenth century to describe a biological phenomenon.
- The earliest form of regression analysis is linear regression, in which a researcher finds the line that most closely fits the data according to a specific mathematical criterion.
- This line is referred to as the line of least squares, which was published by Legendre in 1805,[4] and by Gauss in 1809.[5]

STATISTICS FOR DATA SCIENCE

The Least – Squares Line

- When two variables have a linear relationship, the scatter plot tends to be clustered around a straight line.

- This line is referred to as the Least Squares Line.



Source : dphi.tech

The least squares line

- The least square line given by,

$$y_i = \widehat{\beta_0} + \widehat{\beta_1} x_i$$

where

- $\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$
- $\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$

Example :

- The details pertaining to the no. of hours spent by students in preparing for an entrance exam and the marks scored (on a scale of (0 – 100) is provided in the following table.

Using these values,

i. Estimate the marks scored by a student who has spent 2.35 hours.

ii. Predict the marks that a student can score if he/she invests 20 hours.

SL No.	No. of hours spent	Marks Scored
1	6	82
2	10	88
3	2	56
4	4	64
5	6	77
6	7	92
7	0	23
8	1	41
9	8	80
10	5	59
11	3	47

Computing the least squares line

❖ *We need to first obtain the least square line which is given by,*

$$y = \widehat{\beta_0} + \widehat{\beta_1} x$$

- $\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$
- $\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$

Example :

SL No.	No. of hours spent (x)	Marks Scored(y)	$x - \bar{x}$	$(x - \bar{x})^2$	$y - \bar{y}$	$(x - \bar{x})(y - \bar{y})$
1	6	82	1.27	1.6129	17.55	22.33
2	10	88	5.27	27.7729	23.55	124.15
3	2	56	-2.73	7.4529	-8.45	23.06
4	4	64	-0.73	0.5329	-0.45	0.33
5	6	77	1.27	1.6129	12.55	15.97
6	7	92	2.27	5.1529	27.55	62.60
7	0	23	-4.73	22.3729	-41.45	195.97
8	1	41	-3.73	13.9129	-23.45	87.42
9	8	80	3.37	11.3569	15.55	50.88
10	5	59	0.27	0.0729	-5.45	-1.49
11	3	47	-1.73	2.9929	-17.45	30.15
	4.73	64.45	0.07	94.8459	0.05	611.37

Example :

From the table we have,
$\bar{x} = 4.73 \; ; \bar{y} = 64.45$

- $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = 611.37$
- $\sum_{i=1}^{n}(x_i - \bar{x})^2 = 94.8459$
- $\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{611.37}{94.8459} = 6.49$
- $\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x} = 64.45 - (6.49)(4.73) = 30.18$
- The equation of the least squares line is given by,
  $y_i = \widehat{\beta_0} + \widehat{\beta_1}x_i \Rightarrow y = 30.18 + 6.49x$

Example :

- The equation of the least squares line is given by,
  $y = 30.18 + 6.49x$

i. To estimate the marks scored by a student who has spent 2.35 hours.

$$y = 30.18 + 6.49x = 30.18 + (6.49)(2.35) = 45.43$$

ii. To predict the marks that a student can score if he/she invests 20 hours.

$$y = 30.18 + 6.49x = 30.18 + (6.49)(20) \cong 160$$

PES UNIVERSITY ONLINE

THANK YOU

Dr. Karthiyayini
Department of Science & Humanities

Karthiyayini.roy@pes.edu

+91 80 6618 6651