Unit-3

Regression

What is Regression ?

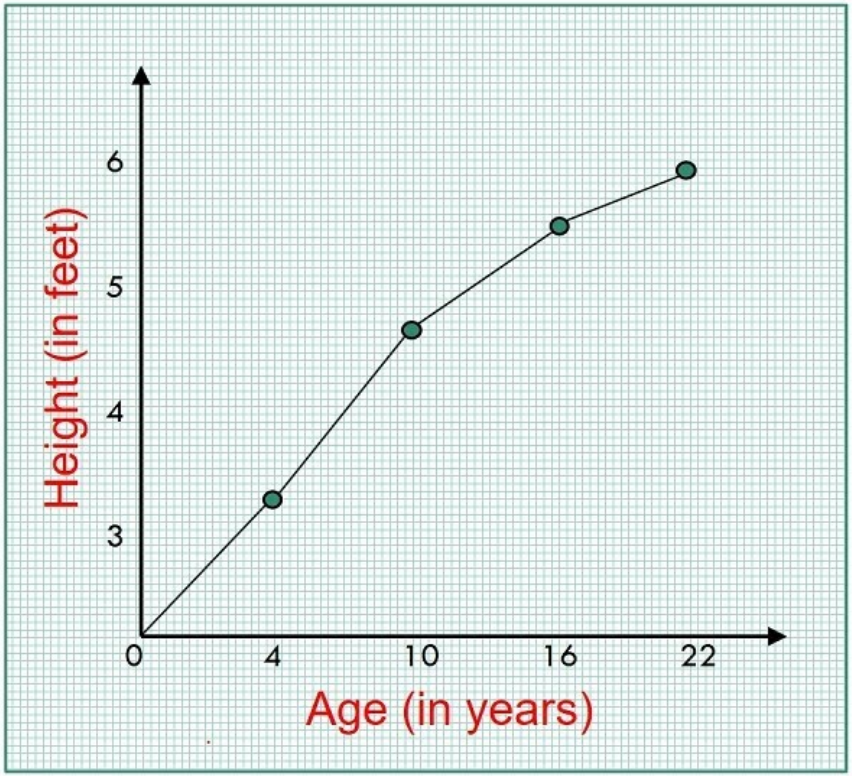
- Regression searches for relationships among variables.
- Regression analysis is a set of statistical methods used for the estimation of relationships between a dependent variable and one or more independent variables.
- It can be utilized to assess the strength of the relationship between variables and for modeling the future relationship between them.

https://realpython.com/linear-regression-in-python/

independent & dependent variable

- The **independent variable** is the cause. Its value is **independent** of other variables in your study.
- The **dependent variable** is the effect. Its value **depends** on changes in the independent variable.

Dependent Vs Independent Variables



Key Differences

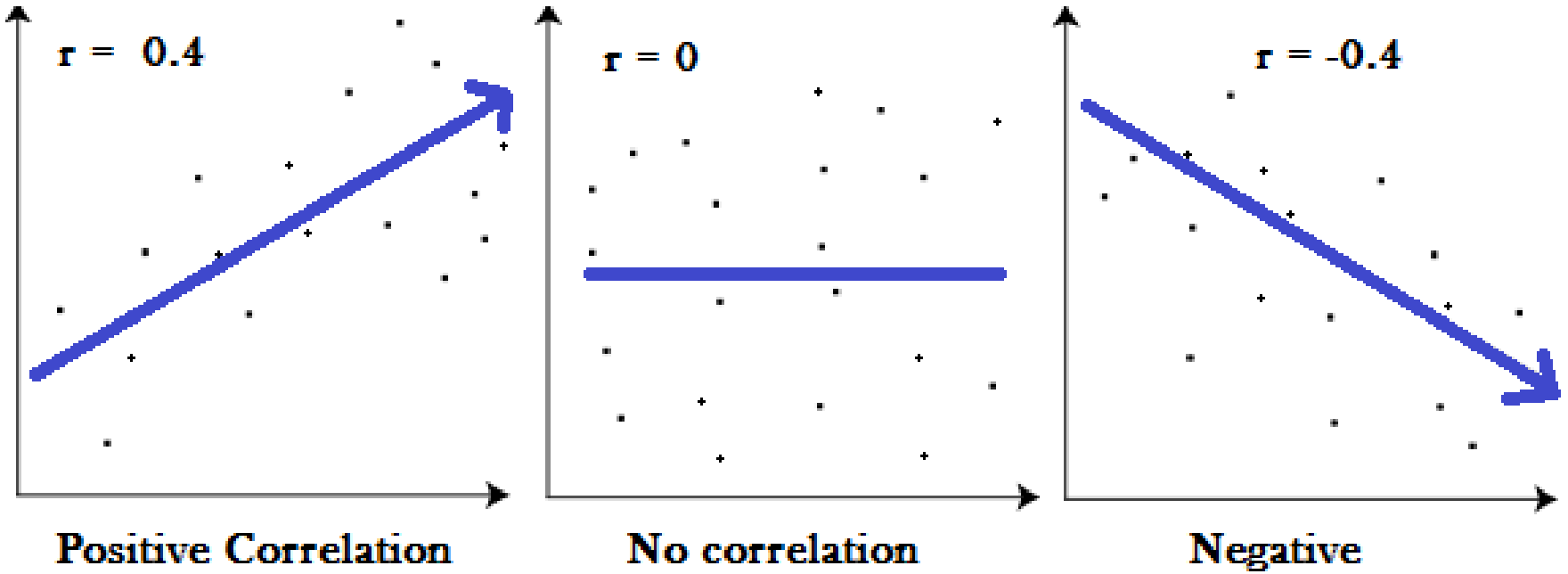
Correlation Coefficient

- Correlation coefficients are used to measure the **strength of the relationship** between two variables.
- Pearson correlation is the one most commonly used in statistics. This measures the strength and direction of a linear relationship between two variables.

Pearson Correlation Coefficient

- There are several types of correlation coefficients, but the one that is most common is the **Pearson correlation (*r*).** This measures the strength and direction of the linear relationship between two variables.

- Correlation coefficient formulas are used to find how strong a relationship is between data. The formulas return a value between -1 and 1
  - 1 indicates a strong positive relationship.
  - -1 indicates a strong negative relationship.
  - A result of zero indicates no relationship at all.


r = 0.4
r = 0
r = -0.4
Positive Correlation
No correlation
Negative

Formula of Correlation

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

Find Correlation Coefficient

Serial No	Age (years) X	Weight (Kg) Y
1	7	12
2	6	8
3	8	12
4	5	10
5	6	11
6	9	13

Solution

Serial n.	Age (years) (x)	Weight (Kg) (y)	xy	X^2	Y^2
1	7	12	84	49	144
2	6	8	48	36	64
3	8	12	96	64	144
4	5	10	50	25	100
5	6	11	66	36	121
6	9	13	117	81	169
Total	∑x= 41	∑y= 66	∑xy= 461	∑x2= 291	∑y2= 742

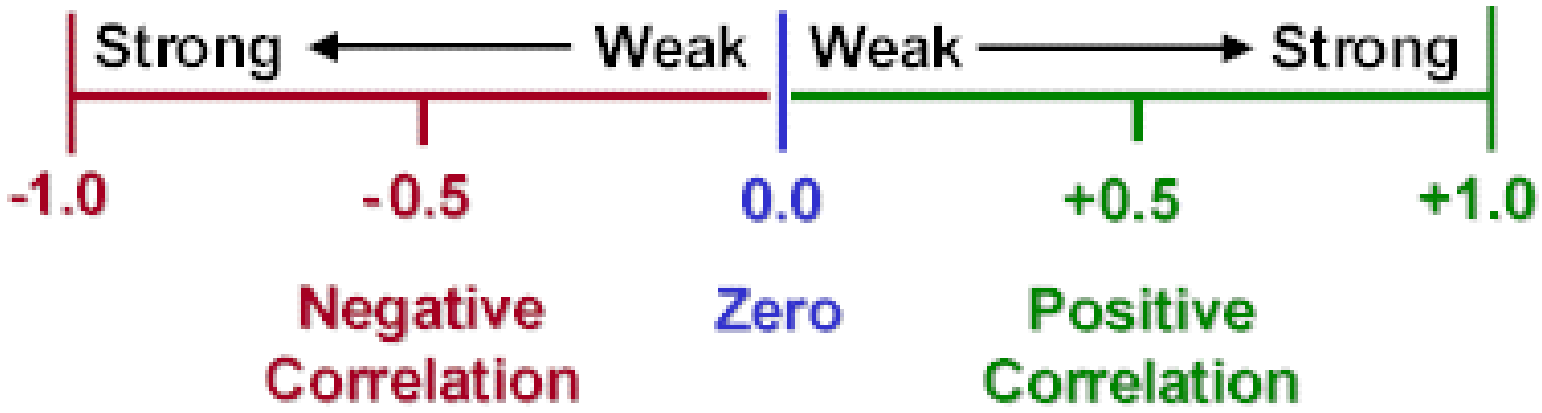
Answer r = 0.759

Result Discussion

- If r = Zero this means no association or correlation between the two variables.
- If $0 < r < 0.25$ = weak correlation.
- If $0.25 \leq r < 0.75$ = intermediate correlation.
- If $0.75 \leq r < 1$ = strong correlation.
- If r = I = perfect correlation.

Correlation Coefficient

Shows Strength & Direction of Correlation



http://www.edugyan.in/2017/02/correlation-coefficient.html

Need of Regression

- Regression is a **statistical approach** used in finance, investment, and other fields to identify the **strength and type of a connection** between one dependent variable (typically represented by Y) and a sequence of other variables (known as independent variables).
- Regression is essentially the **"best guess"** at utilizing a collection of data to generate some form of forecast. It is the process of fitting a set of points to a graph.

Applications of Regression

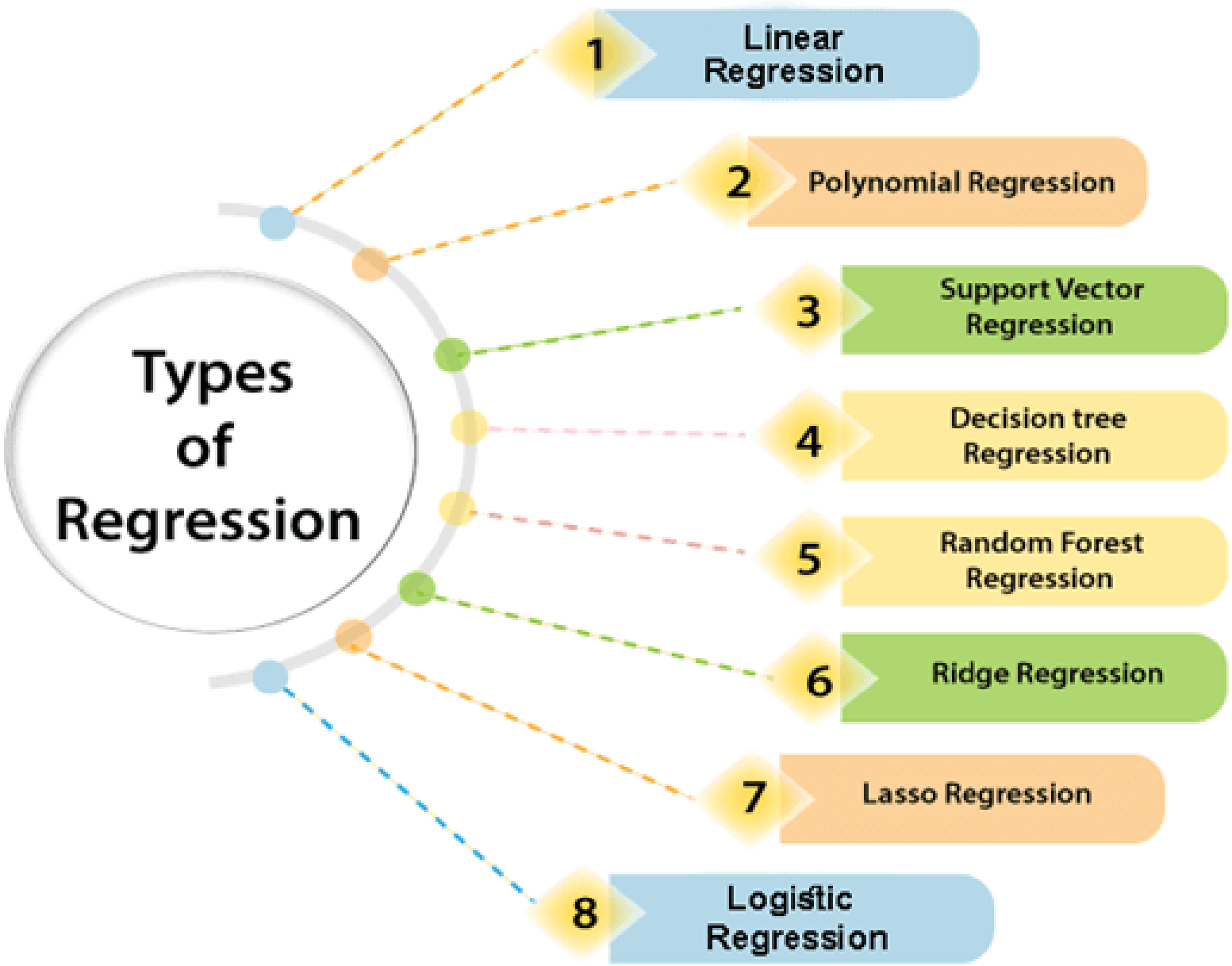
- Linear regression can be used in market research studies and customer survey results analysis
- Linear regressions can be used in business to evaluate trends and make estimates or forecasts.
- Linear Regression can be also used to assess risk in financial services or insurance domain. For example, a car insurance company might conduct a linear regression to come up with a suggested premium table using predicted claims to Insured Declared Value ratio.

Applications of Regression

- Height and weight — as height increases, you'd expect the weight to increase, but not perfectly.
- Alcohol consumed and blood alcohol content — as alcohol consumption increases, you'd expect one's blood alcohol content to increase, but not perfectly.
- Vital lung capacity and pack-years of smoking — as the amount of smoking increases (as quantified by the number of pack-years of smoking), you'd expect lung function (as quantified by vital lung capacity) to decrease, but not perfectly.
- Driving speed and gas mileage — as driving speed increases, you'd expect gas mileage to decrease, but not perfectly.

Types of Regression

- Simple Linear Regression
- Multiple Linear Regression
- Polynomial Regression


Types of Regression
1 Linear Regression
2 Polynomial Regression
3 Support Vector Regression
4 Decision tree Regression
5 Random Forest Regression
6 Ridge Regression
7 Lasso Regression
8 Logistic Regression

Simple Linear Regression

- Simple linear regression is a statistical approach that allows us to study and summarize the relationship between two continuous quantitative variables.
- Out of the two variables, one variable is called the dependent variable, and the other variable is called the independent variable.
- Our goal is to predict the dependent variable's value based on the value of the independent variable.
- A simple linear regression aims to find the best relationship between X (independent variable) and Y (dependent variable).

Mathematically the relationship can be represented with the help of following equation

$$Y=A+Bx$$

Y is the dependent variable.
X is the independent variable.
A and **B** are constants which are called the coefficients.

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

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

Exercise - 1

SUBJECT	AGE X	GLUCOSE LEVEL Y
1	43	99
2	21	65
3	25	79
4	42	75
5	57	87
6	59	81

Solution of Exercise - 1

SUBJECT	AGE X	GLUCOSE LEVEL Y	XY	X2	Y2
1	43	99	4257	1849	9801
2	21	65	1365	441	4225
3	25	79	1975	625	6241
4	42	75	3150	1764	5625
5	57	87	4959	3249	7569
6	59	81	4779	3481	6561
Σ	247	486	20485	11409	40022

From the above table, **Σx** = 247, **Σy** = 486, **Σxy** = 20485, $\mathbf{\Sigma x^2}$ = 11409, $\mathbf{\Sigma y^2}$ = 40022.
n is the sample size (6, in our case).

y' = a + bx
y' = 65.14 + 0.385225x

Exercise - 2

A sample of 6 persons was selected the value of their age (x variable) and their weight is demonstrated in the following table. Find the regression equation and what is the predicted weight when age is 8.5 years.

Serial no.	Age (x)	Weight (y)
1	7	12
2	6	8
3	8	12
4	5	10
5	6	11
6	9	13

Solution of Exercise - 2

Serial no.	Age (x)	Weight (y)	xy	X^2	Y^2
1	7	12	84	49	144
2	6	8	48	36	64
3	8	12	96	64	144
4	5	10	50	25	100
5	6	11	66	36	121
6	9	13	117	81	169
Total	41	66	461	291	742

$$Y = 4.675 + 0.92x$$