Introduction to Statistical Methods and Minitab Software

VIP Machine Learning Course

Basic Concepts in Statistics

- Median
- Mean
- Variance and covariance

Median

{13, 23, 11, 16, 15, 10, 26}

{10, 11, 13, 15, 16, 23, 26}

{3, 13, 7, 5, 21, 23, 23, 40, 23, 14, 12, 56, 23, 29}

{3, 5, 7, 12, 13, 14, 21, 23, 23, 23, 23, 29, 40, 56}

21 + 23 = 44

44 ÷ 2 = 22

Mean

29 , 23 , 56 , 12 , 14 , 23 , 40 , 23 , 39 , 23 , 20 , 13 , 5 , 7 , 3

330 ÷ 15 = 22

Difference Between Variance and Covariance

- Variance is a measure of the spread of data points around the mean of a single variable.
- Covariance, on the other hand, measures how two variables change together relative to their means.
- Covariance is typically used to analyze the relationship between two variables.
- If two variables are identical, their variance and covariance will be equal.

Calculating Covariance

Assuming the statistical set has two specific properties A and B, the dataset would be as follows:

{ (a1, b1)...(an, bn) }

$$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

Calculating Covariance

x	y	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
12	20	-9.3	-21.2	197.16
30	60	8.7	18.8	163.56
15	27	-6.3	-14.2	89.46
24	50	2.7	8.8	23.76
14	21	-7.3	-20.2	147.46
18	30	-3.3	-11.2	36.96
28	61	6.7	19.8	132.66
26	54	4.7	12.8	60.16
19	32	-2.3	-9.2	21.16
27	57	5.7	15.8	90.06
$\dot{x} = 21.3$	$\bar{y} = 41.2$			$\Sigma = 962.4$

Calculating Covariance

x	y	$(x_i - \bar{x})(y_i - \bar{y})$
12	20	197.16
30	60	163.56
15	27	89.46
24	50	23.76
14	21	147.46
18	30	36.96
28	61	132.66
26	54	60.16
19	32	21.16
27	57	90.06
$\bar{x} = 21.3$	$\bar{y} = 41.2$	$\Sigma = 962.4$

$$Cov(x, y) = s_{xy} = \frac{962.4}{n - 1}$$

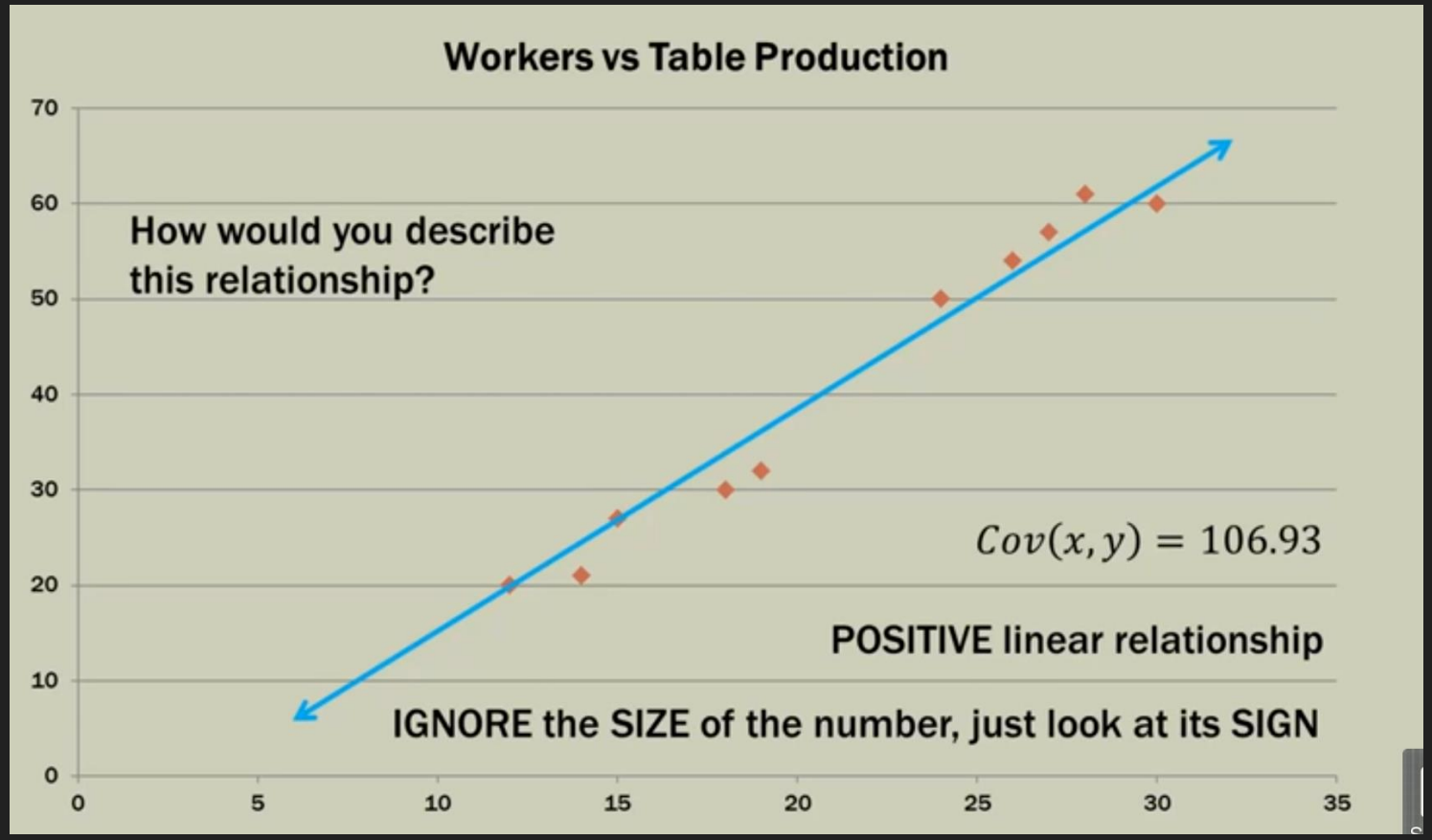
$$\frac{962.4}{9}$$

$$Cov(x, y) = 106.93$$

Positive and Negative Covariance

- The magnitude of covariance is not very important; rather, its sign is crucial.
- A positive covariance indicates positive linear changes, while a negative covariance indicates negative linear changes.

Calculating Covariance

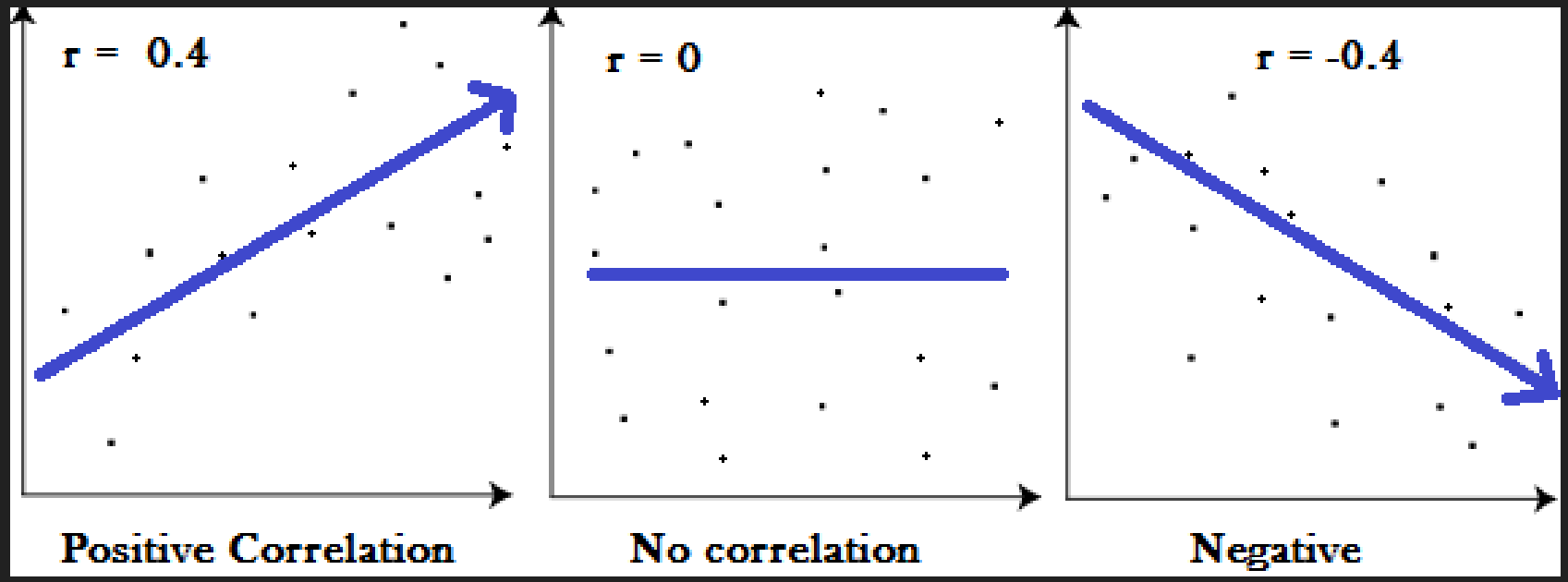


Correlation Coefficient

- The statistical method Pearson's Product Moment Coefficient is used to calculate correlation between numerical data.
- Covariance alone does not have a specific interpretation as it does not represent the strength of the relationship between specific properties. Therefore, the correlation coefficient should be used for this purpose.

Correlation Coefficient

- The correlation coefficient can range from -1 to +1.
- A value of 0 means no correlation.
- The closer the number is to +1, the stronger the positive correlation, and the closer it is to -1, the stronger the negative correlation.



Calculating Correlation Coefficient

Subject	Age x	Glucose Level y	xy	x^2	y^2
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**

Calculating Correlation Coefficient

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

$\Sigma x = 247$

$\Sigma y = 486$

$\Sigma xy = 20485$

$\Sigma x2 = 11409$

$\Sigma y2 = 40022$

$n = 6$

The Correlation coefficient =

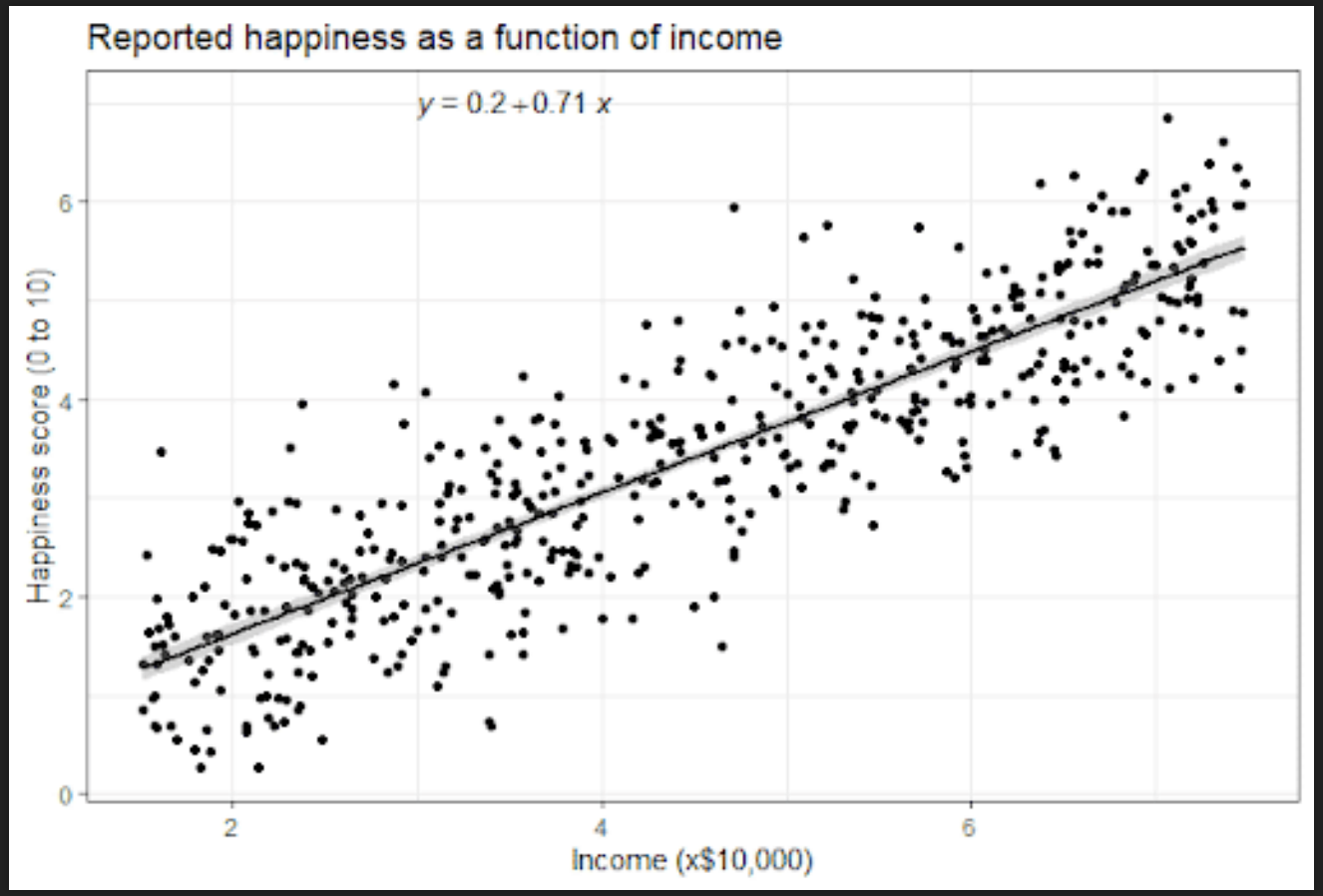
$6(20485) – (247 \times 486) / [\sqrt{[[6(11409) – (2472)] \times [6(40022) – 4862]]}]$

$= 0.5298$

Regression

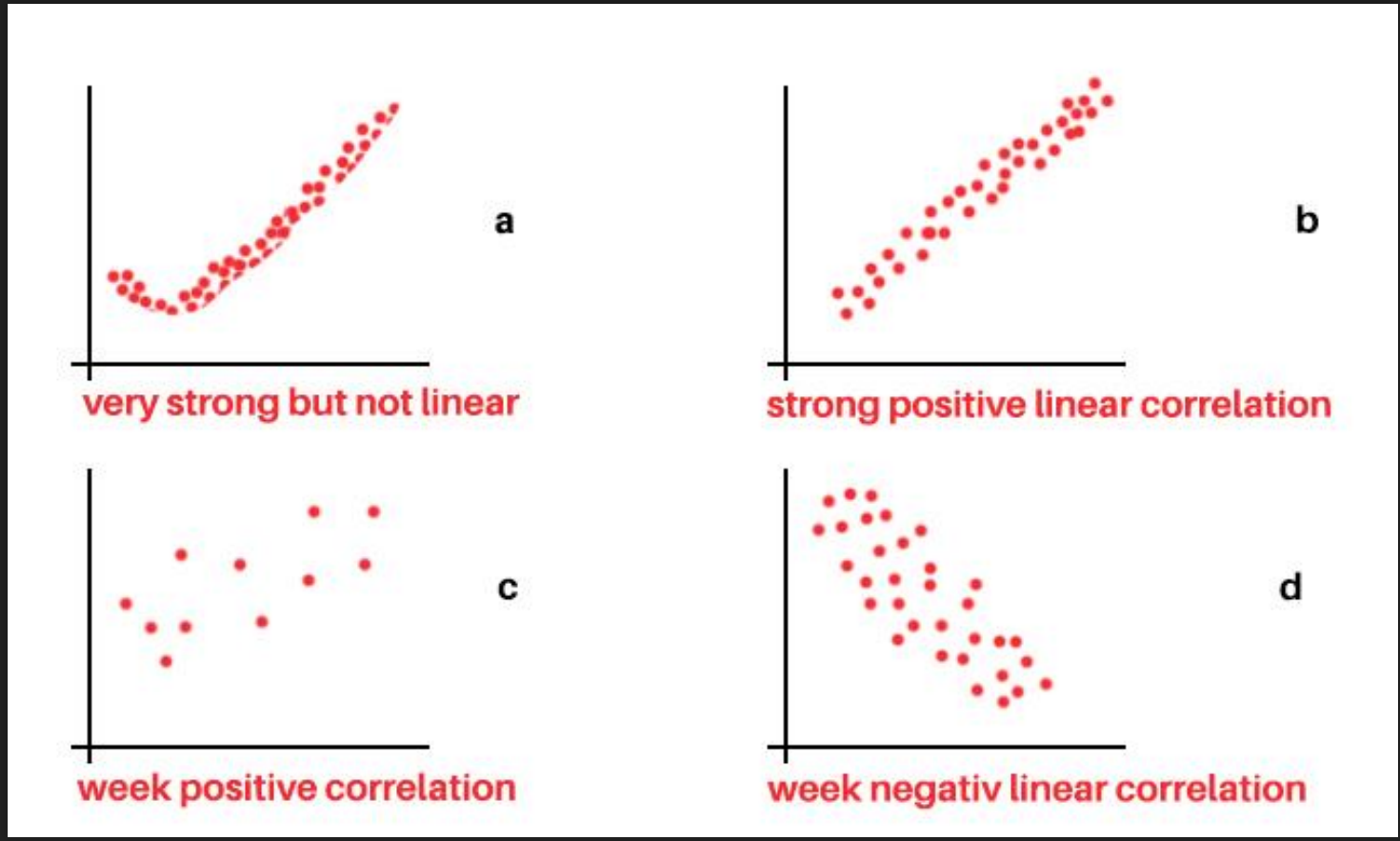
- It is used to predict continuous values such as prices, income levels, height, weight, and so on.
- In various scientific fields, such as mechanics, most problems can be solved using regression methods.

Regression

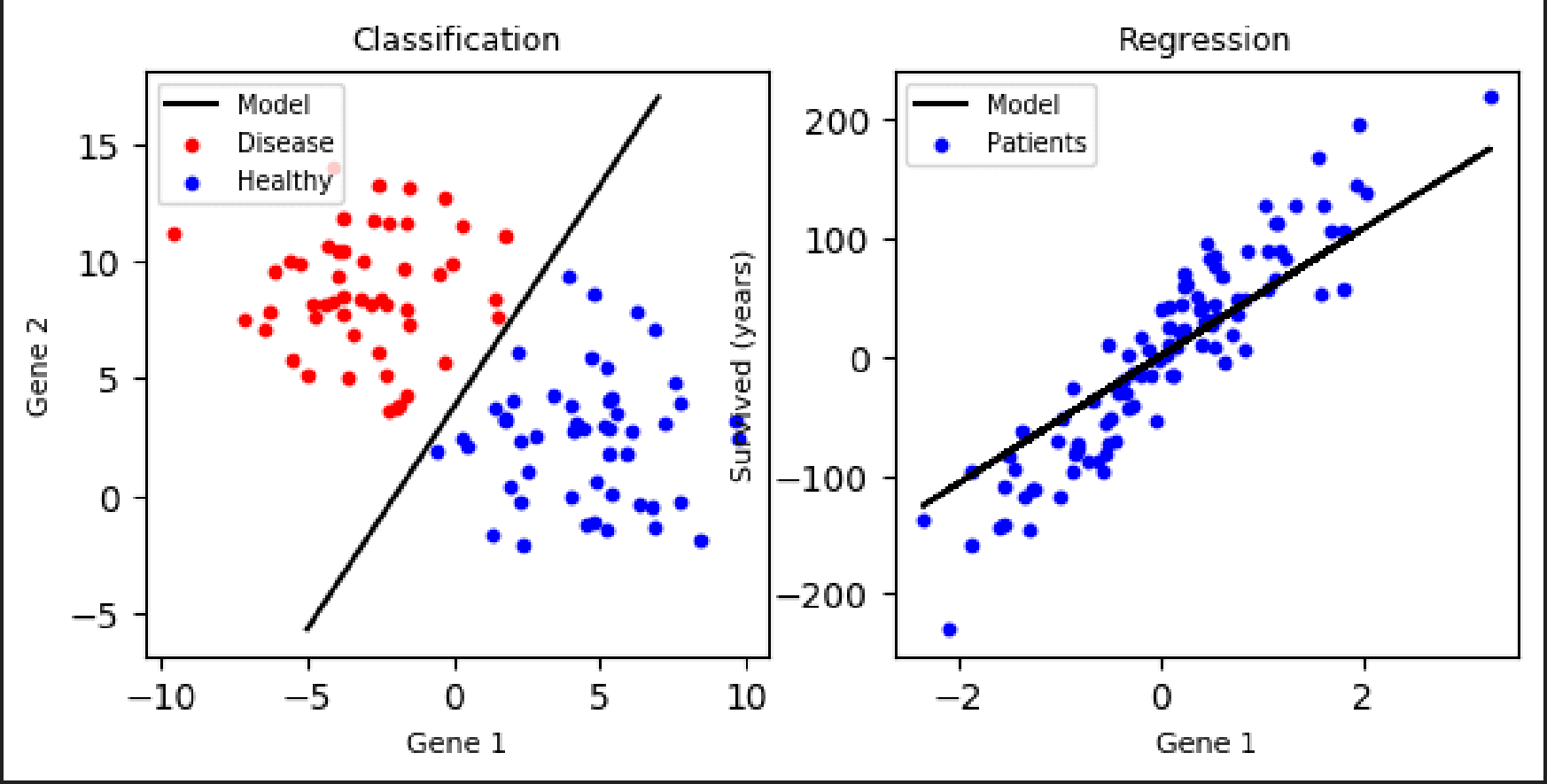


Y = a + bX

Regression


a
very strong but not linear
b
strong positive linear correlation
c
week positive correlation
d
week negativ linear correlation

Regression VS Classification



Design Of Experiments

- Let's consider an experiment involving 4 factors (agents, parameters, or variables) that interact with each other. If each factor can take three levels, denoted as a, b, and c, the total number of experiments that can be conducted is:
- 3 raised to the power of 4, which equals 81 experiments or L^F, where L represents the level and F represents the factor. The scenario that considers all possible experiment combinations is called Full Factorial, while one that considers only a subset is called Fractional Factorial.
- Now, suppose achieving the same results as before with just 16 experiments, which is 20% of the total experiments. This concept has led to the development of various experimental design methods.
- By using experimental design, it is possible to create an input parameter matrix (similar to a randomized trial), conduct experiments, obtain results, and analyze them.

Advantages of experimental design in the development stage

- Reduction of process output variability
- Decreased development time
- Lower overall costs

Advantages of experimental design in the engineering design stage

- Evaluation and comparison of basic design schemes
- Evaluation of different materials
- Selection of design factors so that the product works under different conditions
- Determine the key design factors that affect product performance

Response Surface Methodology (RSM)

- Response Surface Methodology (RSM) is used when the number of experiments becomes impractically large due to a high number of independent variables (high number of factors).
- In such cases, two main issues arise. First, the cost of conducting numerous experiments becomes prohibitive. Second, time constraints prevent the completion of all experiments (making it impossible to conduct a Full Factorial design).

Response Surface Methodology (RSM)

- In Response Surface Methodology (RSM), levels are encoded using the numbers 0, 1, and -1 to denote the middle level (0), the high level (1), and the low level (-1) respectively.
- To work with RSM, experiments should be planned after an in-depth study to identify appropriate levels and factors, and to determine suitable responses for designing experiments correctly.

Response Surface Methodology (RSM)

- In summary, Response Surface Methodology (RSM) focuses on designing an experimental study and models the combined effects of multiple variables. It then establishes a regression model to relate responses to factors.

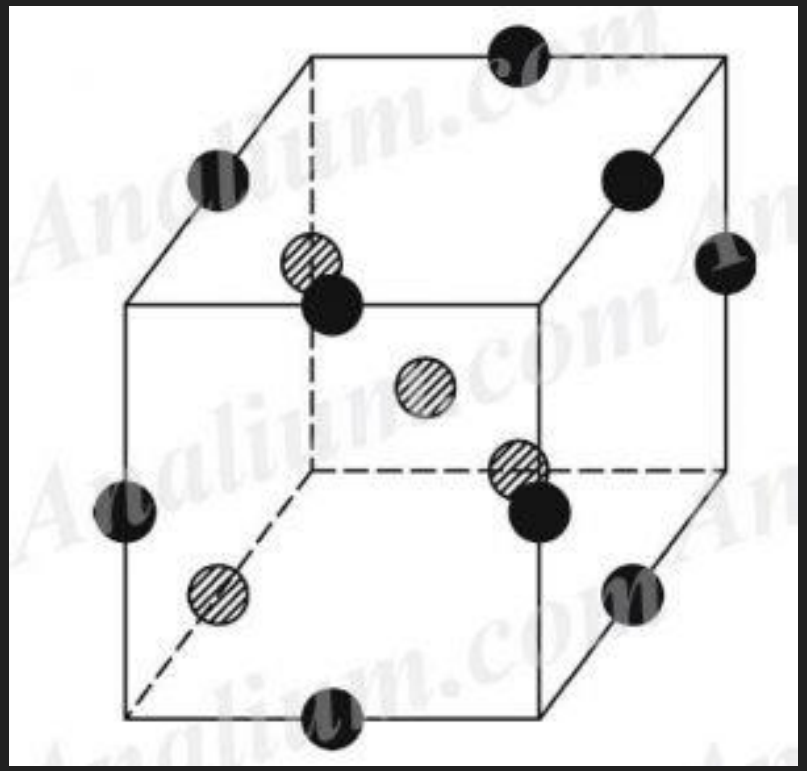
- Some of its key objectives include:
- Improving processes or identifying optimal inputs
- Addressing process issues and weaknesses
- Enhancing process stability

Some of the methods of Response Surface Methodology

- Central composite design
- Box-Behnken design

Box-Behnken Design

- The figure shows an example of Box-Behnken design for three factors.
- The number of experiments for this method is calculated using the formula:
- $N=2^K\times(K-1)+C0$
- where:
- NNN is the total number of experiments,
- KKK is the number of factors,
- C0C0C0 is the number of center points, typically equal to 3.



The End

Thank you for your attention. I wish you pleasant times ahead.