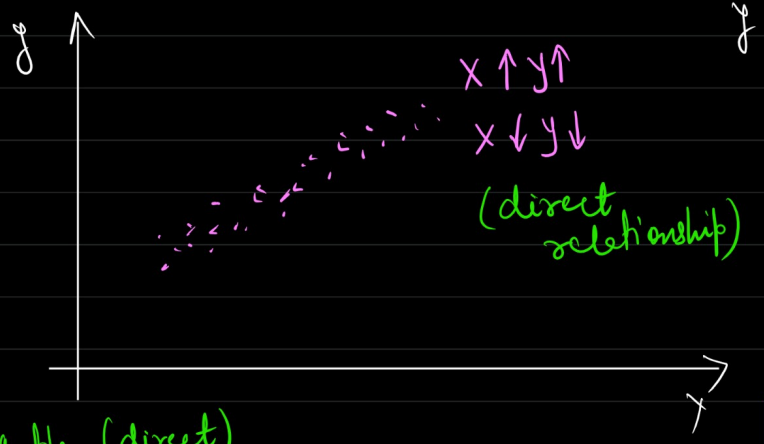
Covariance and Correlation

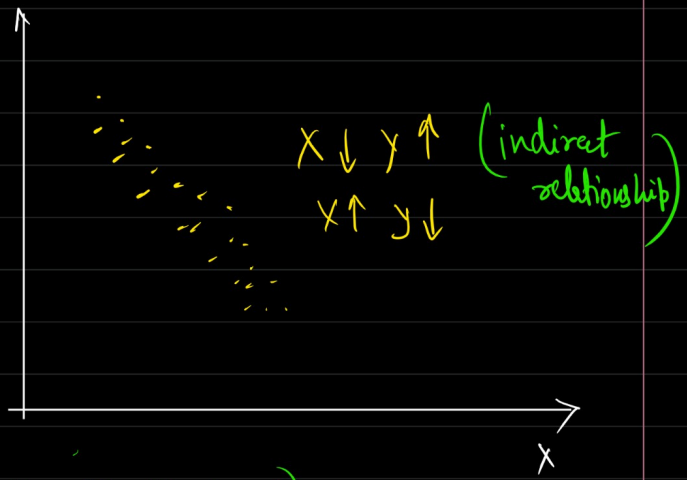
- $y\uparrow x\uparrow$
- $y\uparrow x\downarrow$
- $y\downarrow x\uparrow$
- $y\downarrow x\downarrow$

$\Rightarrow$ relationship b/w two features.

(y) transaction count	(x) total transaction amount
5	1 lakh
4	50K.
–	–
–	–
–	–

* Understanding the relationship



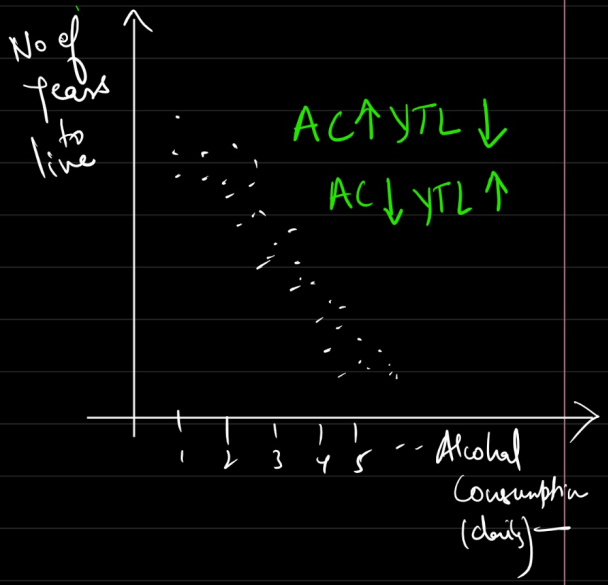


Example (direct)

* Predict price of house based on area of house.

AH↑ P↑
AH↓ P↓

Example (Indirect)



* To quantify / measure the relationship

① Covariance

$$Cov(x,y) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

where x & y are two features

Co - Variance
- Co → Existing together
- Variance → Spread of data

$$Var(x) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1} = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(x_i - \bar{x})}{n-1}$$

$Var(x)$ → Spread of data

⇓

relationship of a feature with itself.

* Covariance means, you are trying to understand the relationship of a feature with respect to other feature.

* Covariance – relationship b/w two variables.

→ $x \uparrow y \uparrow$, $x \downarrow y \downarrow$ (+ve covariance) or $x \uparrow y \downarrow$, $x \downarrow y \uparrow$ (-ve covariance)

$$cov(x,y) = \sum_{i=1}^{n} \frac{(x - \bar{x})(y - \bar{y})}{n-1}$$

x	y	$x - \bar{x}$	$y - \bar{y}$
-2	3	(2-3)	(3-5.5)
-3	5	(3-3)	(5-5.5)
-6	6	(6-3)	(6-5.5)
-1	8	(1-3)	(8-5.5)
$\bar{x}=3$	$\bar{y}=5.5$		

$$\Rightarrow \frac{(-3)(-3.5) + (3-3)(5-5.5) + (6-3)(6-5.5) + (-3)(8-5.5)}{4-1}$$

$$\Rightarrow \frac{(-1)(-2.5) + 0 + 3 \times 0.5 + (-2) \times 2.5}{3}$$

$$= \frac{2.5 + 0 + 1.5 - 5}{3} = \frac{-1}{3} = \boxed{-0.33}$$

$\rightarrow -0.33 \rightarrow$ interpretation $\rightarrow$ The two features X & Y are negatively related.

X	Y
2	3
4	5
6	7
$\bar{x}=4$	$\bar{y}=5$

$$Cov(x,y) = \frac{\sum_{i=1}^{n}(x-\bar{x})(y-\bar{y})}{n-1}$$

$$= \frac{(2-4)(3-5) + (4-4)(5-5) + (6-4)(7-5)}{3-1}$$

(the $(4-4)(5-5)$ term $\rightarrow 0$)

$$= \frac{4+0+4}{2} = 4 \ (+ve)$$

$Cov(x,y) = 4$

interpretation $\rightarrow$ x & y are having a positive relationship

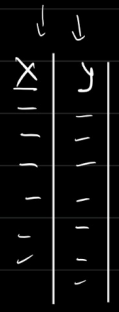
* Advantage

$\rightarrow$ Relationship b/w x, y

+ve, -ve.

* Disadvantage of Covariance

$$Cov = \frac{\sum(x-\bar{x})(y-\bar{y})}{n-1}$$



① X Y

$Cov(x,y) = 50$

A B

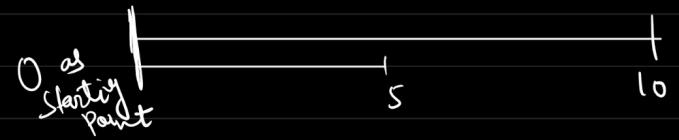
$Cov(A,B) = 100$

Can I say the relationship b/w X and Y is twice of A and B ?? $\Rightarrow$ No

Cov (x,y)

$\Downarrow$

range $\rightarrow -\infty$ to ∞ $\longrightarrow$ you cannot compare the strength of relationship.



→ No comparison of strength of relationship in covariance

→ No any standardised scale to interpret the strength.

② Covariance has dimension

$$Cov(x,y) = \sum_{i=1}^{n} \frac{(x-\bar{x})(y-\bar{y})}{n-1}$$

(x → ht., $\bar{x}$ → ht, n is no of rows.)

x_1 height (ft)	x_2 Weight Kg	x_3 transaction amount Rs
⋮	⋮	⋮

$Cov(\text{transaction Amt}, ht) = Rs \cdot ft \Rightarrow 450\ Rs \cdot ft$

$Cov(ht, wt) = ft \cdot kg \Rightarrow 600\ ft \cdot kg$

* we can not compare two different dimension.

$Cov(x,y)$ → Rs ft

$Cov(y,z)$ → ft kg

→ Not comparable → different dimension.

Soln

→ −1 to 1

→ dimensionless

② Pearson Correlation Coefficient [−1 to 1]

$$\rho_{(x,y)} = \frac{Cov(x,y)}{\sigma_x \sigma_y} \Rightarrow [-1 \text{ to } 1]$$

→ dimensionless quantity

No relationship

-1 ← 0 → 1

← the more negatively correlated the features are. (indirect)

→ The more positively correlated features are. (direct)

$\rightarrow \rho_{x,y} = 0.4$, $\rho_{A,B} = 0.8$

$\rightarrow$ feature A, B is highly correlated as compared to x, y.

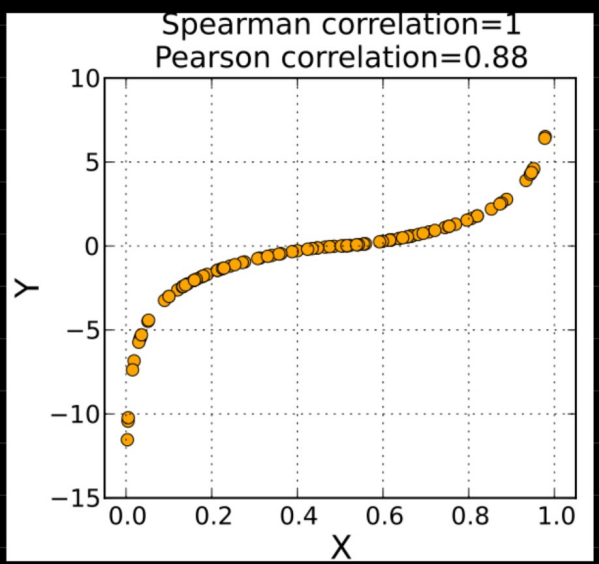
$\rightarrow \rho_{x,y} = -0.2$ $\quad \rho_{AB} = -0.5$

$x \uparrow y \downarrow$ $\quad$ $A \uparrow B \downarrow$

$x \downarrow y \uparrow$ $\quad$ $A \downarrow B \uparrow$

AB is more negative correlated as compared to x, y

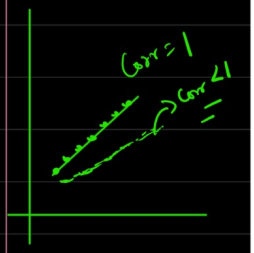
* Pearson correlation coefficient always measures the linear relationship

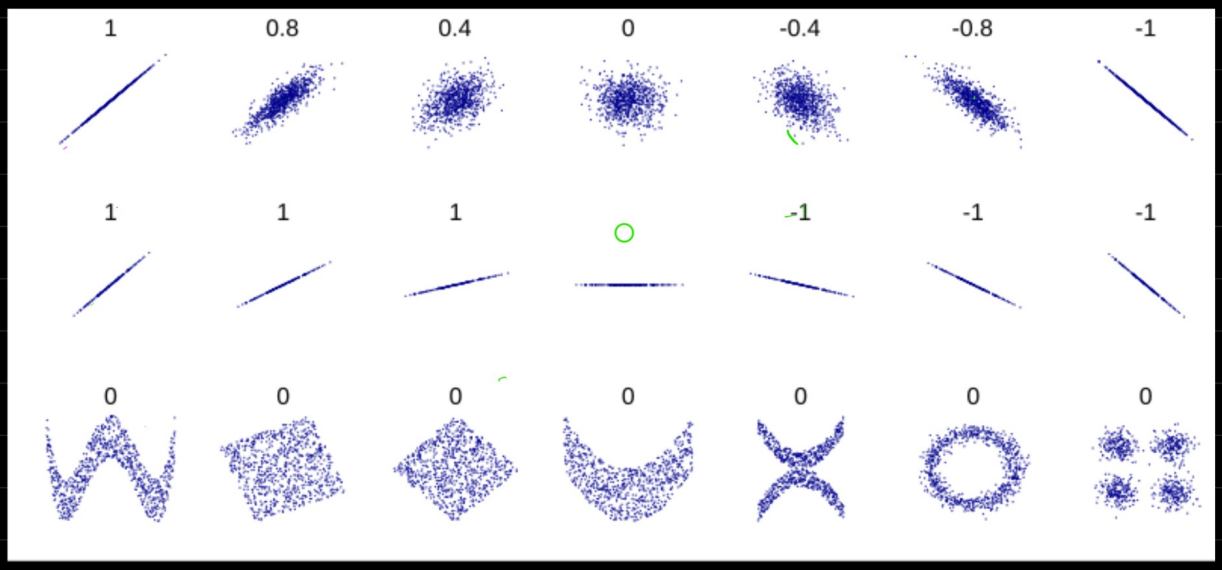


Person correlation

$x \uparrow y \uparrow$
$x \downarrow y \downarrow$
$x \uparrow y \downarrow$
$x \downarrow y \uparrow$

$\Rightarrow x \uparrow y \uparrow$





* Correlation is not slope
* Non linear data will have Pearson correlation 0.

* What to do for understanding Non-linear relationship

Spearman Rank Correlation

$$\rho_s = \frac{Cov(R(x), R(y))}{\sigma(R(x)) \cdot \sigma(R(y))}$$

R(x) - Rank of x
R(y) - Rank y.

x	y	R(x)	R(y)
5	6	3	1
7	4	2	2
8	3	1	3
1	1	5	5
2	2	4	4

$x \rightarrow 5, 7, 8, 1, 2 \Rightarrow$ Sort the value

5th 4th 3rd 2nd 1st

$1, 2, 5, 7, 8$

Highest no will be rank 1

* Use Case

x_1	x_2	x_3	. . . .	x_{1000}	Price ($\bar{y}$)

$x_1 - y$
$x_2 - y$
$x_3 - y$
$x_4 - y$

Corr ≈ 0 → Correlation is used for feature selection in ML modelling