Unit 2: Correlation and Regressions

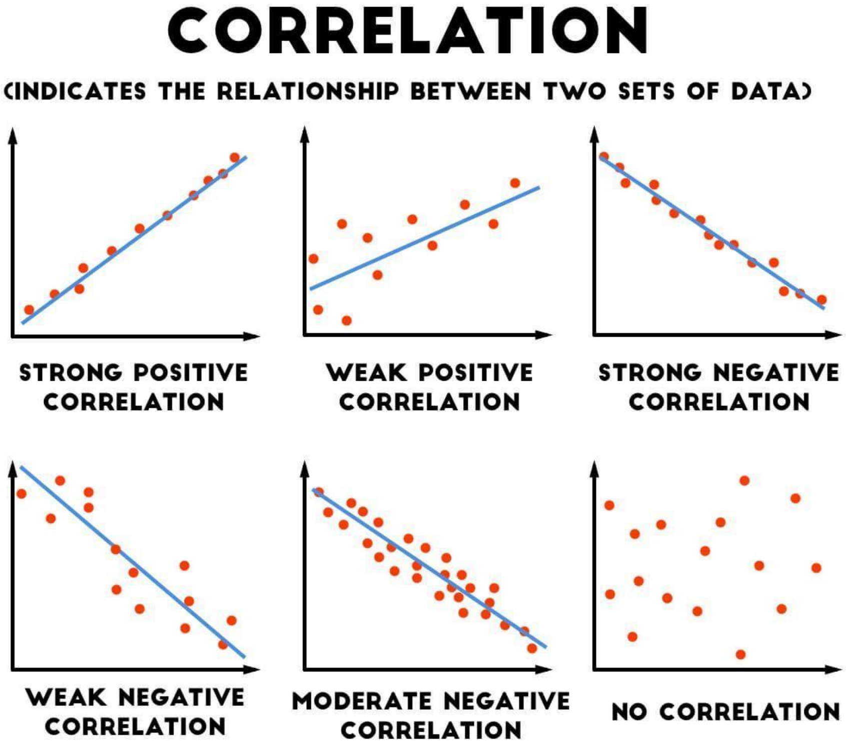
Correlation: If the change in one variable affects a change in the other variable, the variables are said to be correlated.

- **Positive Correlation**
- **Negative Correlation**

(Definitions discussed in class)

Scatter Diagram

It is a way of diagrammatic representation of bivariate date. It is also known as dotted representation.



KARL PEARSON'S COEFFICIENT OF CORRELATION

It is numerical measure of linear relationship between two variables. It is denoted as $r(X,\ Y)$.

$$r(X,\ Y) = \frac{Cov(X,\ Y)}{\sigma_X \sigma_Y}$$

Cov(X, Y) is also denoted as σ_{XY} $or\ \mu_{11}$.

- **Karl Pearson's correlation coefficient** is also called **product-moment correlation coefficient.**
- Two independent variables are uncorrelated, but the converse is not true.
- Correlation coefficient is independent of change of origin.
  (Proof discussed in class)
- Correlation coefficient lies between ±1.

$$-1 \leq r \leq 1$$

$$r = \frac{Cov(X,Y)}{\sigma_X \sigma_X} = \frac{\frac{1}{n}\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n}\sum(x_i-\bar{x})^2}\sqrt{\frac{1}{n}\sum(y_i-\bar{y})^2}}$$

$$r^2 = \frac{\left(\sum(x_i-\bar{x})(y_i-\bar{y})\right)^2}{\sum(x_i-\bar{x})^2\sum(y_i-\bar{y})^2}$$

$$r^2 = \frac{\left(\sum a_i b_i\right)^2}{\sum a_i^2 \sum b_i^2} \leq 1$$

$$(a_1b_1 + a_2b_2)^2 = a_1^2b_1^2 + a_2^2b_2^2 + 2a_1a_2b_1b_2$$

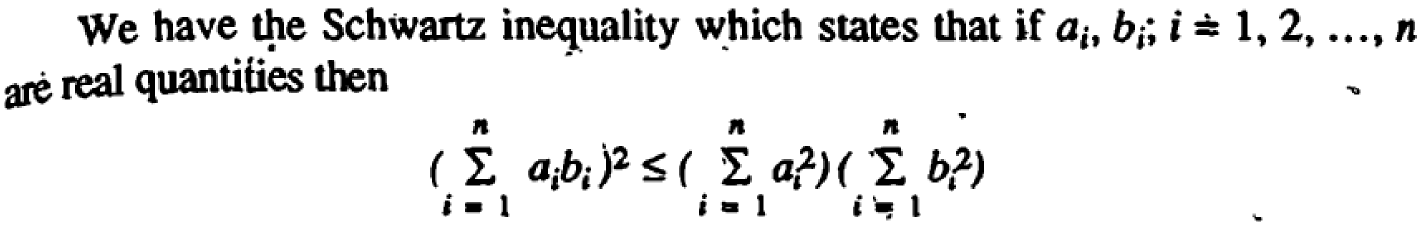
$$(a_1^2 + a_2^2)(b_1^2 + b_2^2) = a_1^2b_1^2 + a_1^2b_2^2 + a_2^2b_1^2 + a_2^2b_2^2$$

$$\text{den} - \text{num} = a_1^2b_2^2 + a_2^2b_1^2 - 2a_1a_2b_1b_2$$

$$= (a_1b_2 - a_2b_1)^2 \geq 0$$

$$r^2 \leq 1$$

$$-1 \leq r \leq 1$$



We have the Schwartz inequality which states that if $a_i, b_i; i = 1, 2, \ldots, n$ are real quantities then

$$\left(\sum_{i=1}^{n} a_i b_i\right)^2 \leq \left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} b_i^2\right)$$

` Q1. Calculate the correlation coefficient for the following heights of fathers (X) and their sons (Y):

X:	65	66	67	67	68	69	70	72
Y:	67	68	65	68	72	72	69	71

Q2: A computer while calculating correlation coefficient between two variable X and Y from 25 pairs of observations obtained the following results:

$$n = 25 \,, \sum X = 125, \sum X^2 = 650, \sum Y = 100, \ \sum Y^2 = 460, \ \sum XY = 508$$

If was, however, later discovered at the time of checking that he had copied down two pairs as

X | Y while correct values were X | Y .

6 | 14 8 | 12

9 | 6 6 | 8

Obtain the correct value of correlation coefficient.

Q3. The joint probability distribution of X and Y is given below:

X → Y ↓	-1	1
0	1/8	3/8
1	2/8	2/8

Find correlation coefficient between X and Y.

Q4. The following table gives, according to age, the frequency of marks obtained by 100 students in an intelligent test:

Age in Years (X) → Marks(Y) ↓	18	19	20	21	TOTAL
10-20	4	2	2	-----------	8
20-30	5	4	6	4	19
30-40	6	8	10	11	35
40-50	4	4	6	8	22
50-60	------	2	4	4	10
60-70	-------	2	3	1	6
TOTAL	19	22	31	28	100

Find correlation coefficient.

Q5. X, Y and Z are random variables each with expectation 10 and variances 1, 4 and 9 resp. The correlation coefficients are $\mathrm{r(X,\ Y) = 0,\ r(Y,\ Z) = r(Z, X) = 1/4}$.

Find

(i) $\mathrm{E(X + Y - 2Z)}$

(ii) $\mathrm{Cov(X + 3,\ Y + 3)}$

(iii) $\mathrm{V(Y - 2Z)}$

(iv) $\mathrm{Cov(3X,\ 5Z)}$

(v) $r(3X, 5Z)$

Spearmann's Rank Correlation Coefficient: It is denoted as $\rho(X,\ Y)$.

$$\rho(\mathrm{X},\ \mathrm{Y}) = 1 - \frac{6\sum_{i=1}^{n} d_i^{\,2}}{n(n^2-1)}$$

Assuming that no two individuals are bracketed equal in eith classification, each of the variables X and Y takes the values 1, 2, ..., n.

Hence $$\bar{x} = \bar{y} = \frac{1}{n}(1 + 2 + 3 + \ldots + n) = \frac{n+1}{2}$$

$$\sigma_X^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = \frac{1}{n}(1^2 + 2^2 + \ldots + n^2) - \left(\frac{n+1}{2}\right)^2$$

$$= \frac{n(n+1)(2n+1)}{6n} - \left(\frac{n+1}{2}\right)^2 = \frac{n^2-1}{12}$$

$$\therefore \qquad \sigma_X^2 = \frac{n^2-1}{12} = \sigma_Y^2$$

In general $x_i \neq y_i$. Let $d_i = x_i - y_i$

$$\therefore \qquad d_i = (x_i - \bar{x}) - (y_i - \bar{y}) \qquad (\because \bar{x} = \bar{y})$$

Squaring and summing over i from 1 to n, we get

$$\sum d_i^2 = \sum\{(x_i - \bar{x}) - (y_i - \bar{y})\}^2$$

$$= \sum(x_i - \bar{x})^2 + \sum(y_i - \bar{y})^2 - 2\sum(x_i - \bar{x})(y_i - \bar{y})$$

Dividing both sides by n, we get

$$\frac{1}{n}\sum d_i^2 = \sigma_X^2 + \sigma_Y^2 - 2\ \mathrm{Cov}\ (X,\ Y) = \sigma_X^2 + \sigma_Y^2 - 2\rho\ \sigma_X\sigma_Y$$

where ρ is the rank correlation coefficient between A and B.

$$\therefore \qquad \frac{1}{n}\sum d_i^2 = 2\sigma_X^2 - 2\rho\sigma_X^2 \ \Rightarrow\ 1 - \rho = \frac{\sum d_i^2}{2n\sigma_X^2}$$

$$\Rightarrow \qquad \rho = 1 - \frac{\sum_{i=1}^{n} d_i^2}{2n\sigma_X^2} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \qquad \ldots(10$$

Q6. The marks secured by recruits in the selection test (X) and in the proficiency test (Y) are given below:

X:	10	15	12	17	13	16	24	14	22	20

Y:	30	42	45	46	33	34	40	35	39	38

Calculate rank correlation coefficient.

REPEATED RANKS

(Process of assigning ranks is explained in class.)

$$\rho(\mathrm{X},\ \mathrm{Y}) = 1 - \frac{6\left(\sum_{i=1}^{n} d_i^2\ +\ \ correction\ factor\ for\ each\ tie\right)}{\mathrm{n}(\mathrm{n}^2 - 1)}$$

where $correction\ factor = \frac{m(m^2-1)}{12}$, m represents repetition of each tied rank.

Q7. A sample of 12 fathers and their eldest sons gave the following data about their height in inches:

Father:	65	63	67	64	68	62	70	66	68	67	69	71
Son:	68	66	68	65	69	66	68	65	71	67	68	70

Calculate coefficient of rank correlation.

Q8. Obtain the rank correlation coefficient

X:	68	64	75	50	64	80	75	40	55	64
Y:	62	58	68	45	81	60	68	48	50	70

Q9. The coefficient of rank correlation between marks in Statistics and marks in Mathematics obtained by a certain group of students is 0.8. if the sum of the squares of the difference of ranks is given to be 33. Find no. of students in the group.

REGRESSION

It is mathematical measure of the average relationship between two or more variables in terms of the original units of the data.

- The variables whose value is influenced or is to be predicted is called dependent variable (Regressed or Explained variable) and the variable who influences the values or is used for prediction is called independent variable (Repressor or Explanatory variable).
- The curve around which points of scatter diagram cluster is called **curve of regression.**

- If the curve is a straight line, it is called line of regression (**Linear Regression**), otherwise **Curvilinear.**
- The line of regression is the line which gives the best estimate to the value of one variable for any specific value of other variable **(Line of Best Fit).**

Principle of Least Squares

It consists in minimising the sum of the square of the deviations of the actual values of dependent variable from their estimated values as given by the line of best fit.

Regression line for Y **on** X

$$(Y-\bar{Y})=\frac{r\sigma_Y}{\sigma_X}(X-\bar{X}), \qquad b_{YX}=\frac{r\sigma_Y}{\sigma_X}$$

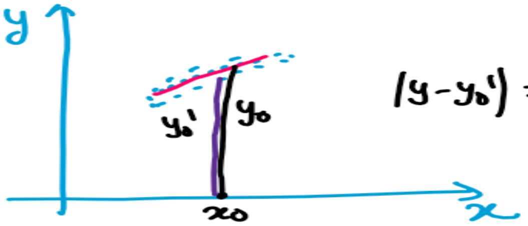
b_{YX} is known as regression coefficient of Y on X.

Regression line for X **on** Y

$$(X-\bar{X})=\frac{r\sigma_X}{\sigma_Y}(Y-\bar{Y}), \qquad b_{XY}=\frac{r\sigma_X}{\sigma_Y}$$

b_{XY} is known as regression coefficient of X on Y.

P**roof of Regression line *Y* on *X***



$|y - y_0'| = \text{small.}$

Let line of regression for Y on X

$Y = a + bX$

a → point on line

b → slope of line

Error = minimum

$$\text{Error} = \sum_{i=1}^{n} (y_i - (a+bx_i))^2 = \text{minimum}$$

$$\frac{\partial E}{\partial a} = \sum_{i=1}^{n} 2(y_i - a - bx_i)\cdot(-1)$$

$$\sum_{i=1}^{n} (y_i - a - bx_i) = 0$$

$$\sum_{i=1}^{n} y_i - na - b\sum_{i=1}^{n} x_i = 0$$

Divide by n

$$\frac{\sum y_i}{n} - a - b\frac{\sum x_i}{n} = 0$$

$$\bar{Y} - a - b\bar{X} = 0 \quad \text{—}①$$

$$\bar{Y} = a + b\bar{X}$$

Mean will lie on regression line

$$\frac{\partial E}{\partial b} = \sum_{i=1}^{n} 2(y_i - a - bx_i)\cdot(-x_i)$$

$$\sum (x_i y_i - a x_i - b x_i^2) = 0$$

Divide by n

$$\frac{\sum x_i y_i}{n} - a\frac{\sum x_i}{n} - b\frac{\sum x_i^2}{n} = 0$$

$E(XY) = Cov(X,Y) + E(X)\,E(Y)$

$$(\mu_{11} + \bar{X}\bar{Y}) - a\bar{X} - b(\sigma_X^2 + \bar{X}^2) = 0 \quad \text{—}②$$

$$② - ①\bar{X}$$

$$\mu_{11} + \cancel{\bar{X}\bar{Y}} - \cancel{a\bar{X}} - b\sigma_X^2 - \cancel{b\bar{X}^2}$$
$$-(\cancel{\bar{X}\bar{Y}} - \cancel{a\bar{X}} - \cancel{b\bar{X}^2}) = 0$$

$$\mu_{11} - b\sigma_X^2 = 0$$

$$b = \frac{\mu_{11}}{\sigma_X^2} = \frac{r\sigma_X\sigma_Y}{\sigma_X^2}$$

$$b = \text{slope} = \frac{r\sigma_Y}{\sigma_X} = \text{Slope of } Y \text{ on } X.$$

$(\bar{X}, \bar{Y})$ → Point on line , Slope $= \frac{r\sigma_Y}{\sigma_X} = b_{YX} =$ Regression coefficient of Y on X

$$(Y - \bar{Y}) = \frac{r\sigma_Y}{\sigma_X}(X - \bar{X})$$

$$(Y - \bar{Y}) = b_{YX}\,(X - \bar{X})$$

Imp points to remember

- Both the lines of regression pass through the point $(\bar{X},\ \bar{Y})$.
- Two lines of regression are different. The regression line Y on X is obtained on minimising the sum of the squares of the errors parallel to Y −axis, while the regression equation of X on Y is obtained by minimsing the sum of squares of the errors parallel to X −axis.
- For perfect correlation, both the lines of regression coincide.

Properties of Regression Coefficients (Proofs discussed in class)

- Correlation coefficient is the geometric mean between the regression coefficients.
- r, b_{XY}, b_{YX} are of same sign.
- If one of the regression coefficients is greater than unity, the other must be less than unity.
- Regression coefficients are independent of change of origin but not of scale.

r(X,Y) is independent of change of origin always

" " " " scale when scales are of same sign

$U = aX+b, \quad V = cY+d$

$a, c > 0$

$r(U,V) = r(X,Y)$

$\sigma_U^2 = a^2\sigma_X^2, \quad \sigma_V^2 = c^2\sigma_Y^2$

$\sigma_U = a\sigma_X, \quad \sigma_V = c\sigma_Y$

$$b_{VU} = \frac{r(U,V)\sigma_V}{\sigma_U} = \frac{r(X,Y)\cdot c\sigma_Y}{a\sigma_X} = \frac{c}{a} b_{YX}$$

- The modulus value of the arithmetic mean of the regression coefficients is not less than the modulus of the correlation coefficient r.

$A.M \geq \underset{r}{\underline{G.M}}$

$$\left|\frac{b_{YX}+b_{XY}}{2}\right| \geq |r|$$

$$\left|\frac{\frac{r\sigma_Y}{\sigma_X}+\frac{r\sigma_X}{\sigma_Y}}{2}\right| \geq |r| \Rightarrow \frac{\sigma_Y^2+\sigma_X^2}{2\sigma_X\sigma_Y} \geq 1$$

$$\sigma_X^2+\sigma_Y^2-2\sigma_X\sigma_Y \geq 0$$

$$(\sigma_X-\sigma_Y)^2 \geq 0$$

- Angle between two lines of regression

$$\tan\theta = \left|\frac{m_1 - m_2}{1 + m_1 m_2}\right|$$

$$(y-\bar{y}) = \frac{r\sigma_y}{\sigma_x}(x-\bar{x}) \qquad (x-\bar{x}) = \frac{r\sigma_x}{\sigma_y}(y-\bar{y})$$

$$(y-\bar{y}) = \frac{\sigma_y}{r\sigma_x}(x-\bar{x})$$

$$m_1 = \frac{r\sigma_y}{\sigma_x}$$

$$m_2 = \frac{\sigma_y}{r\sigma_x}$$

$$\tan\theta = \left|\frac{\frac{r\sigma_y}{\sigma_x} - \frac{\sigma_y}{r\sigma_x}}{1 + \frac{r\sigma_y}{\sigma_x}\cdot\frac{\sigma_y}{r\sigma_x}}\right| = \left|\frac{r^2-1}{r}\right|\left(\frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}\right)$$

$$\tan\theta = \frac{1-r^2}{|r|}\;\frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}$$

Use this formula, if data is given

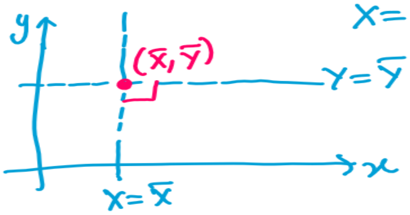
$\tan\theta = \frac{m_1 - m_2}{1 + m_1 m_2}$ Use this formula, if regression lines are given

Special cases of θ

Case I:- $r = 0$ Uncorrelated $\theta = \frac{\pi}{2}$

$(y-\bar{y}) = \frac{r\sigma_y}{\sigma_x}(x-\bar{x})$, $(x-\bar{x}) = \frac{r\sigma_x}{\sigma_y}(y-\bar{y})$

$Y = \bar{Y}$ $X = \bar{X}$



Case II Perfect correlation $r = \pm 1$

$\tan\theta = 0, \quad \theta = 0$

Lines of regression will coincide.

Q10. Obtain the equations of two lines of regression for the following heights of fathers (X) and their sons (Y). Also obtain the estimate of height of Father when height of son is 70.

X:	65	66	67	67	68	69	70	72
Y:	67	68	65	68	72	72	69	71.

Q11. Using the following data, find regression line Y on X.

	X	Y
No. of items	15	15
Arithmetic mean	25	18
Sum of squares of deviations from mean	136	138

Also, summation of product of deviations of X and Y from their respective arithmetic mean is 122.

Q12. Out of the two regression lines given by $X + 2Y - 5 = 0,\ 2X + 3Y - 8 = 0.$

(i) Find which one is regression line of Xon Y?

(ii) Mean Values of X and Y

(iii) Correlation coefficient

(iv) Angle between regression lines.

- **Standard Error of estimate**

$$s_X = \sigma_X\sqrt{1 - r^2}\,, \qquad s_Y = \sigma_Y\sqrt{1 - r^2}$$

- **Ratio of Coefficient of variability for** X **on** Y is

$$\frac{\sigma_X^2}{\sigma_Y^2} = \frac{b_{XY}}{b_{YX}}$$

- **Correlation coefficient between estimated and observed value**$=r(X,\hat{X}) = r(Y,\hat{Y}) = r(X,Y)$
- **Coefficient of variation**$=\frac{\sigma}{\mu}$

Q12. For 10 observations on price X and supply Y the following data were obtained.

$$\sum X = 130,\ \sum Y = 220,\ \sum X^2 = 2288,\ \sum Y^2 = 5506,\ \sum XY = 3467.$$

(a) Obtain the line of regression of Y on X and estimate the supply when the price is 16 units.

$E(X)= 13$, $E(Y)=22$, $E(X^2)= 228.8$ $E(Y^2)= 550.6$

$E(XY)= 346.7$ $\sigma_X = 7.733$ $\sigma_Y = 8.160$

$Cov(X,Y)= 60.7$ $r(X,Y)=0.962$

Y on X

$$(Y-\bar{Y}) = \frac{r\sigma_Y}{\sigma_X}(X-\bar{X})$$

$$(Y-22) = \frac{(0.962)\ (8.160)}{(7.733)}(X-13)$$

$$Y = 1.015X + 8.79$$

Value of Y, when $X=16$, is $\boxed{25.045}$

(b) Find out

(i) Standard error of the estimate.

(ii) Angle between two regression lines.

(iii) Ratio of coefficient of variability of X to that of Y.

(1) $s_x = \sigma_x \sqrt{1-r^2}$, $s_y = \sigma_y \sqrt{1-r^2}$

$s_x = 2.110$ $s_y = 2.348$

(2) $$\tan\theta = \frac{1-r^2}{|r|}\left(\frac{\sigma_x \sigma_y}{\sigma_x^2+\sigma_y^2}\right) = \frac{1-(0.962)^2}{0.962} \times \left(\frac{(7.733)(8.160)}{(7.733)^2+(8.160)^2}\right)$$

$\tan\theta = 0.0387$

(3) $$\text{Ratio of coeff of var} = \frac{\sigma_x^2}{\sigma_y^2} = \left(\frac{7.733}{8.160}\right)^2$$

$$= 0.898$$