Frequently Asked Questions

1. What do you mean by curve of regression?
   Answer:
   If the variables in a bivariate distribution are related, we will find that the points in the scatter diagram will cluster around some curve called the "curve of regression".

2. What is line of regression?
   Answer:
   The line of regression is the line, which gives the best estimate to the value of one variable for any specific value of the other variable. Thus, the line of regression is the line of "best fit" and is obtained by Principle of least squares.

3. Find the regression line Y on X.
   Answer:
   Consider the equation, $\Sigma y_i = na + b\Sigma x_i$
   On dividing by n, we get, $\bar{y} = na + b\bar{x} - - - - - - - - (1)$
   Thus, the line of regression of Y on X passes through $(\bar{x}, \bar{y})$
   Now

$$\mu_{11} = Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y} \Rightarrow \frac{1}{n}\sum_{i=1}^{n} x_i y_i = \mu_{11} + \bar{x}\bar{y} - - - (2)$$

   Also,

$$\sigma_X{}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i{}^2 - \bar{x}^2 \Rightarrow \frac{1}{n}\sum_{i=1}^{n} x_i{}^2 = \sigma_X{}^2 + \bar{x}^2 - - - - - - (3)$$

   Now, consider the equation, $\Sigma x_i y_i = a\Sigma x_i + b\Sigma x_i{}^2 - - (4)$
   Dividing by n and substituting for $\Sigma x_i y_i/n$, $\Sigma x_i{}^2/n$, we get,
   $\mu_{11} + \bar{x}\bar{y} = a\bar{x} + b(\sigma_X{}^2 + \bar{x}^2) - - - - - - (5)$
   Multiplying the equation (4) by $\bar{x}$ and subtracting from the equation (5), we get,

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

   Since b is the slope of the line of regression Y on X and since the line of regression passes through the point $(\bar{x}, \bar{y})$, its equation is,

$$Y - \bar{y} = b(X - \bar{x}) = \frac{\mu_{11}}{\sigma_X{}^2}(X - \bar{x}) \Rightarrow Y - \bar{y} = r\frac{\sigma_Y}{\sigma_X}(X - \bar{x})$$

4. Write the regression equation X on Y.
   Answer:

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

5. State the properties of Regression Coefficients.
   Answer:
   - Correlation coefficient is the geometric mean between the regression coefficients.
   - If one of the regression coefficients is greater than unity, the other must be less than unity.

- The modulus value of the arithmetic mean of the regression coefficient is not less than the modulus value of the correlation coefficient r.
- Regression coefficients are independent of the change of origin but not of scale.

6. Prove that Correlation coefficient is the geometric mean between the regression coefficients.
   Answer:
   Multiplying b_{XY} and b_{YX}, we get, $b_{XY} \times b_{YX} = r\frac{\sigma_X}{\sigma_Y} r\frac{\sigma_Y}{\sigma_X} = r^2 \Rightarrow r = \sqrt{b_{XY} \times b_{YX}}$

7. Prove that if one of the regression coefficients is greater than unity, the other must be less than unity.
   Answer:
   Let one of the regression coefficient say b_{YX} is greater than unity then we have to show that $b_{XY} < 1$
   $b_{YX} > 1$ implies, $1/ b_{YX} < 1$
   Also $r^2 \leq 1$ implies $b_{YX} . b_{XY} \leq 1$
   i.e. $b_{XY} \leq 1/ b_{YX}$
   Hence, $b_{XY} < 1$

8. Show that the modulus value of the arithmetic mean of the regression coefficient is not less than the modulus value of the correlation coefficient r.
   Answer:
   We have to prove that, $\left|\frac{1}{2}(b_{YX} + b_{XY})\right| > |r|$

   $\Rightarrow \left|\frac{1}{2}\left(r\frac{\sigma_Y}{\sigma_X} + r\frac{\sigma_X}{\sigma_Y} \right)\right| \geq |r^2| \Rightarrow \frac{\sigma_Y}{\sigma_X} + \frac{\sigma_X}{\sigma_Y} \geq 2, (\because |r| > 0)$

   $\Rightarrow {\sigma_Y}^2 + {\sigma_X}^2 - 2\sigma_X\sigma_Y > 0 \Rightarrow (\sigma_Y - \sigma_X)^2 > 0$
   Which is always true, since the square of a real quantity is≥0?

9. Regression coefficients are independent of the change of origin but not of scale.
   Answer:
   Let, $U = \frac{X-a}{h}, V = \frac{Y-b}{k} \Rightarrow X = a + hU, Y = b + kV$
   Where a, b h(>0) and k(>0) are constants.
   $Cov(X,Y) = hkCov(U,V), {\sigma_X}^2 = h^2{\sigma_U}^2, {\sigma_Y}^2 = k^2{\sigma_V}^2$
   Then,
   $\therefore b_{YX} = \frac{Cov(X,Y)}{{\sigma_X}^2} = \frac{hkCov(U,V)}{h^2{\sigma_U}^2} = \frac{k}{h}\frac{Cov(U,V)}{{\sigma_U}^2} = \frac{k}{h}b_{VU}$

10. Obtain an angle between two lines of regression.
    Answer:
    Equations of the lines of regression of Y on X and X on Y are, $Y - \bar{y} = r\frac{\sigma_Y}{\sigma_X}(X - \bar{x})$
    and $X - \bar{x} = r\frac{\sigma_X}{\sigma_Y}(Y - \bar{y})$
    Slopes of these lines are $r\frac{\sigma_Y}{\sigma_X}$ and $\frac{\sigma_Y}{r\sigma_X}$ respectively.

If θ is the angle between the two lines of regression, then,

$$\tan\theta = \left| \frac{r\frac{\sigma_Y}{\sigma_X} - \frac{\sigma_Y}{r\sigma_X}}{1 + r\frac{\sigma_Y}{\sigma_X}.\frac{\sigma_Y}{r\sigma_X}} \right| = \left| \frac{r^2 - 1}{r} \right| \left(\frac{\sigma_X \sigma_Y}{{\sigma_X}^2 + {\sigma_Y}^2} \right)$$

$$= \frac{1 - r^2}{|r|} \left(\frac{\sigma_X \sigma_Y}{{\sigma_X}^2 + {\sigma_Y}^2} \right), (\because r^2 \le 1)$$

$$\therefore \theta = \tan^{-1} \left\{ \frac{1 - r^2}{|r|} \left(\frac{\sigma_X \sigma_Y}{{\sigma_X}^2 + {\sigma_Y}^2} \right) \right\}$$

11. How the two regression lines arranged when two variables are uncorrelated?
Answer:
When two variables are uncorrelated, i.e. If r=0, tanθ = ∞ implies, θ=∏/2. Thus if the two variables are uncorrelated, the lines of regression becomes perpendicular to each other.

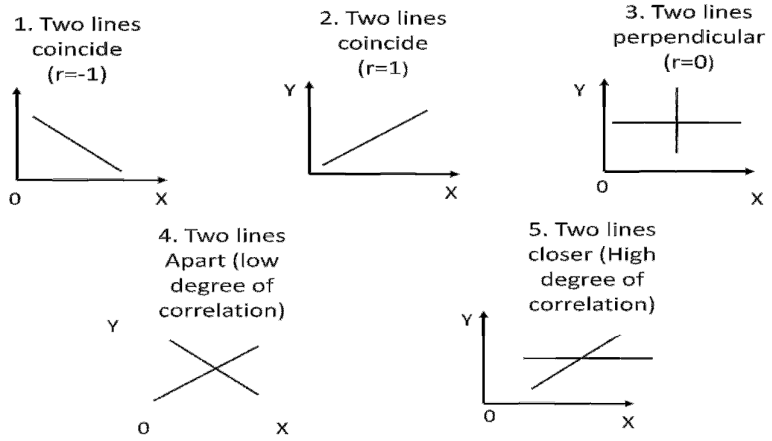
12. How the two regression lines arranged when two variables are perfectly correlated?
Answer:
If r=±1, tanθ=0 implies θ=0 or ∏. In this case, either the two lines of regression coincide or they are parallel to each other. However, since both lines of regression pass through the point $(\bar{x}, \bar{y})$, they cannot be parallel. Hence, in the case of perfect correlation, positive or negative, the two lines of regression coincide.

13. How the angle between two regression lines is used to study the correlation between the variables?
Answer:
The fact that if r=0 (variables uncorrelated), the two lines of regression are perpendicular to each other. If r=±1, θ=0, i.e. the two lines coincide, leads us to the conclusion that for higher degree of correlation between the variables, the angle between the lines is smaller, i.e. the two lines of regression are nearer to each other.

On the other hand, if the lines of regression make a larger angle, they indicate a poor degree of correlation between the variables and ultimately for θ=∏/2, r=0, i.e. the lines becomes perpendicular if no correlation exists between the variables.

14. Plot the two regression lines and give the inference about the correlation between the variables.

Answer:



15. From the following data, find the most likely price in Mumbai corresponding to price of Rs. 70 at Kolkata.

	Kolkata	Mumbai
Average Price	65	67
Standard Deviation	2.5	3.5

Correlation coefficient between the prices of commodities in the two cities is 0.8

Answer:

Let the prices (in Rupees) in Kolkata and Mumbai be denoted by X and Y respectively. Then we have given,

$\overline{X} = 65, \overline{Y} = 67, \sigma_X = 2.5, \sigma_Y = 3.5, r = 0.8$

Line of regression Y on X is $Y - \overline{y} = r\frac{\sigma_Y}{\sigma_X}(X - \overline{x})$

$$Y - 67 = 0.8\frac{3.5}{2.5}(X - 65)$$

When X=70, Y=72.6

Hence, the most likely price in Mumbai corresponding to the price of Rs. 70 at Kolkata is Rs 72.60.