Regression and correlation

Correlation	Regression "Y on X"	Regression "X on Y"
$r_{XY} = r_{YX} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$	$b_{YX} = \frac{Cov(X,Y)}{Var(X)}$	$b_{XY} = \frac{Cov(X,Y)}{Var(Y)}$
$r_{YX} = r_{XX} = \frac{S_{XY}}{\sqrt{S_X^2 S_Y^2}}$	$b_{YX} = \frac{S_{XY}}{S_X^2}$	$b_{XX} = \frac{S_{XY}}{S_Y^2}$
$r_{YX} = r_{XY} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2 \sum(Y-\bar{Y})^2}}$	$b_{YX} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(X-\bar{X})^2}$	$b_{XY} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(Y-\bar{Y})^2}$
$r_{YX} = r_{XY} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{nS_X^2 nS_Y^2}}$	$b_{YX} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{nS_X^2}$	$br_{XY} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{nS_Y^2}$
$r_{YX} = r_{XY} = \frac{\sum XY - n\bar{X}\bar{Y}}{\sqrt{\left[\sum X^2 - n\bar{X}^2\right]\left[\sum Y^2 - n\bar{Y}^2\right]}}$	$b_{YX} = \frac{\sum XY - n\bar{X}\bar{Y}}{\left[\sum X^2 - n\bar{X}^2\right]}$	$b_{XY} = \frac{\sum XY - n\bar{X}\bar{Y}}{\left[\sum Y^2 - n\bar{Y}^2\right]}$
$r_{YX} = r_{XY} = \frac{\sum XY - n\bar{X}\bar{Y}}{\sqrt{nS_X^2 nS_Y^2}}$	$b_{YX} = \frac{\sum XY - n\bar{X}\bar{Y}}{nS_X^2}$	$b_{XY} = \frac{\sum XY - n\bar{X}\bar{Y}}{nS_Y^2}$
$r_{YX} = r_{XY} = \frac{n\sum XY - \sum X \sum Y}{\sqrt{\left[n\sum X^2 - (\sum X)^2\right]\left[n\sum Y^2 - (\sum Y)^2\right]}}$	$b_{YX} = \frac{n\sum XY - \sum X \sum Y}{\left[n\sum X^2 - (\sum X)^2\right]}$	$b_{XY} = \frac{n\sum XY - \sum X \sum Y}{\left[n\sum Y^2 - (\sum Y)^2\right]}$
$r_{YX} = r_{XY} = \frac{n\sum XY - \sum X \sum Y}{\sqrt{nS_X^2 nS_Y^2}}$	$b_{YX} = \frac{n\sum XY - \sum X \sum Y}{nS_X^2}$	$b_{XY} = \frac{n\sum XY - \sum X \sum Y}{nS_Y^2}$
$r_{UV} = r_{VU} = \frac{n\sum UV - \sum U \sum V}{\sqrt{\left[n\sum U^2 - (\sum U)^2\right]\left[n\sum V^2 - (\sum V)^2\right]}}$	$b_{UV} = \frac{n\sum UV - \sum U \sum V}{\left[n\sum U^2 - (\sum U)^2\right]}$	$b_{VU} = \frac{n\sum UV - \sum U \sum V}{\left[n\sum V^2 - (\sum V)^2\right]}$
$r_{XY} = \frac{n\sum D_X D_Y - \sum D_X \sum D_Y}{\sqrt{\left[n\sum D_X^2 - (\sum D_X)^2\right]\left[n\sum D_Y^2 - (\sum D_Y)^2\right]}}$	$b_{YX} = \frac{n\sum D_X D_Y - \sum D_X \sum D_Y}{\left[n\sum D_X^2 - (\sum D_X)^2\right]}$	$b_{XY} = \frac{n\sum D_X D_Y - \sum D_X \sum D_Y}{\left[n\sum D_Y^2 - (\sum D_Y)^2\right]}$
$r_{XY} = \pm\sqrt{b_{YX} b_{XY}}$	$b_{YX} = r\frac{S_Y}{S_X}$	$b_{XY} = r\frac{S_X}{S_Y}$
$r_{YX} = r_{XX} = \frac{S_{XY}}{S_X S_Y}$	$\hat{Y} - \bar{Y} = r\frac{S_Y}{S_X}(X-\bar{X})$	$\hat{X} - \bar{X} = r\frac{S_X}{S_Y}(Y-\bar{Y})$
$r_{YX} = r_{XY} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{nS_X S_Y}$	$\hat{Y} - \bar{Y} = \frac{S_{XY}}{S_X^2}(X-\bar{X})$	$\hat{X} - \bar{X} = \frac{S_{XY}}{S_Y^2}(Y-\bar{Y})$
	$\hat{Y} - \bar{Y} = b_{YX}(X-\bar{X})$	$\hat{X} - \bar{X} = b_{XY}(Y-\bar{Y})$

Q.1: Compute the regression equation of "Y on X" and "X on Y" from the following data using normal equations. Also find coefficient of correlation.

X	25	30	40	50	65
Y	6	5	4	8	7

Solution

X	Y	X^2	Y^2	XY
25	6	625	36	150
30	5	900	25	150
40	4	1600	16	160
50	8	2500	64	400
65	7	4225	49	455
$\sum X$ =210	$\sum Y = 30$	$\sum X^2 = 9850$	$\sum Y^2 = 190$	$\sum XY = 1315$

Regression equation "Y on X"

$Y = a + bX$

It has two other normal equations

$\sum Y = na + b\sum X$ (i)

$\sum XY = a\sum X + b\sum X^2$ (ii)

Putting the values of equation (i) and (ii)

$30 = 5a + 210b$ (i)

$1315 = 210a + 9850b$ (ii)

Multiplying equation (i) by 42 then subtracting equation (ii)

$1260 = 210a + 8820b$ (i)

$1315 = 210a + 9850b$ (ii)

$-55 = -1030b$

$b = \frac{55}{1030} = 0.0533$ Or $b_{YX} = 0.05$ regression coefficient or slope line

Put the value of "b=0.0533" in eq. (i) than we get the value of "a"

$30 = 5a + 210(0.0533)$

$30 = 5a + 11.2136$

$5a = 30 - 11.2136 = 18.80$

$a = \frac{18.80}{5} = 3.8$ $a_{YX} = 3.8$ "Y-Intercept"

Hence the estimated line

$\hat{Y} = 3.8 + 0.053X$

Regression equation "X on Y"

$X = a + bY$ Or $X = c + dy$

It has two other normal equations

$\sum X = na + b\sum Y$ (i)

$\sum XY = a\sum Y + b\sum Y^2$ (ii)

Putting the values of equation (i) and (ii)

$210 = 5a + 30b$ (i)

$1315 = 30a + 190b$ (ii)

Multiplying equation (i) by 6 then subtracting equation (ii)

$1260 = 30a + 180b$ (i)

$1315 = 30a + 190b$ (ii)

$-55 = -10b$

$b = \frac{55}{10} = 5.5$ Or $b_{XY} = 5.5$ regression coefficient or slope line

Put the value of "b=5.5" in eq. (i) than we get the value of "a"

$210 = 5a + 30(5.5)$

$210 = 5a + 165$

$5a = 210 - 165 = 45$

$a = \frac{45}{5} = 9$ $a_{XY} = 3.8$ Or "X-Intercept"

Hence the estimated line

$\hat{X} = 9.0 + 5.5Y$

Correlation coefficient "r"

$r = \pm\sqrt{b_{YX} b_{XY}} = \sqrt{0.053(5.5)} = 0.54$

Q.4: Show that the sum of errors equals to zero and the sum of squares of the errors equal to 1.1.

X	1	2	3	4	5
Y	1	1	2	2	4

Solution:

X	Y	X^2	XY	$\hat{Y} = -0.1 + 0.7X$	$(Y - \hat{Y})$	$(Y - \hat{Y})^2$
1	1	1	1	0.6	0.4	0.16
2	1	4	2	1.3	-0.3	0.09
3	2	9	6	2.0	0	0
4	2	16	8	2.7	-0.7	0.49
5	4	25	20	3.4	0.6	0.36
$\sum X = 15$	$\sum Y = 10$	$\sum X^2 = 55$	$\sum XY = 37$		$\sum(Y - \hat{Y}) = 0$	$\sum(Y - \hat{Y})^2 = 1.1$

Regression line "Y on X"

$Y = a + bX$

$$b_{YX} = \frac{n\sum XY - \sum X \sum Y}{\left[n\sum X^2 - (\sum X)^2\right]} = \frac{5(37) - (15 \times 10)}{5(55) - (15)^2} = \frac{35}{50} = 0.70 \qquad \overline{X} = \frac{\sum X}{n} = \frac{15}{5} = 3$$

$$a = \overline{Y} - b\overline{X} = 2 - 0.7(3) = 2 - 2.10 = -0.1 \qquad \overline{Y} = \frac{\sum Y}{n} = \frac{10}{5} = 2.0$$

$\hat{Y} = -0.1 + 0.7X$

$\sum(Y - \hat{Y}) = 0$ Hence proved

$\sum(Y - \hat{Y})^2 = 1.1$ Hence proved

Q.6: A market research firm wishes to develop a model to predict purchases of tennis balls by city, based on the number of tennis courts in a city. A simple random of 50 cities developed the following data:

X= number of courts in a city

Y= Thousands of tennis balls sold in the city

$\overline{X} = 235$ $\overline{Y} = 375$ $\sum XY = 4435650$ $\sum X^2 = 2780850$

What is the equation of the estimated regression line that you would use to predict "Y from X".

Solution:

$n = 50$ $\overline{X} = 235$ $\overline{Y} = 375$ $\sum XY = 4435650$ $\sum X^2 = 2780850$

$$b_{YX} = \frac{\sum XY - n\overline{X}\overline{Y}}{\left[\sum X^2 - n\overline{X}^2\right]} = \frac{4435650 - 50(235)(375)}{2780850 - 50(235)^2} = \frac{29400}{19600} = 1.50$$

$a = \overline{Y} - b\overline{X} = 375 - 1.5(235) = 375 - 1.5(235) = 22.5$

$\hat{Y} = 22.5 + 1.5X$

Q.8: For 9 observations on supply (X) and price (Y) the following data was obtained

$\sum(X - 90) = -25$ $\sum(X - 90)^2 = 301$ $\sum(Y - 127) = 12$

$\sum(Y - 127)^2 = 1006$ $\sum(X - 90)(Y - 127) = -469$. Obtained the line of regression of "X on Y" and estimate the supply when the price is Rs.125

Solution:

Regression line "X on Y"

$\hat{X} = a + bY$

$\sum D_X = \sum(X - 90) = -25$ $\sum D_X^2 = \sum(X - 90)^2 = 301$ $\sum D_Y = \sum(Y - 127) = 12$

$\sum D_Y^2 = \sum(Y - 127)^2 = 1006$ $\sum D_X D_Y = \sum(X - 90)(Y - 127) = -469$

$$b_{XY} = \frac{n\sum D_X D_Y - \sum D_X \sum D_Y}{\left[n\sum D_Y^2 - (\sum D_Y)^2\right]} = \frac{9(-469) - (-25)(12)}{9(1006) - (12)^2} = \frac{-3921}{8910} = 0.44$$

$$\overline{X} = A + \frac{\sum D_X}{n} = 90 + \frac{-25}{9} = 87.22$$

$$\bar{Y} = A + \frac{\sum D_Y}{n} = 127 + \frac{12}{9} = 128.33$$

$$a = \bar{X} - b\bar{Y} = 87.22 - (-0.44)(128.33) = 87.22 + 56.47 = 143.70$$

$$\hat{X} = 143.70 - 0.44Y$$

Estimate the supply when the price is Rs.125

$$\hat{X} = 143.70 - 0.44(125) = 88.70$$

Q.9: Given the following information, estimate

i) The value of "X" when "Y=30" ii) The value of "Y" when "X=55"

The mean value of "X=54". The mean value of "Y=28". The regression coefficient of "X on Y" is "-0.2". The regression coefficient of Y on X is "-1.5".

Solution:

$\bar{X} = 54$ $\bar{Y} = 28$ $X = 55$ $Y = 30$ $b_{YX} = -1.5$ $b_{XY} = -0.2$

Regression line "X on Y"

$$\hat{X} - \bar{X} = b_{XY}(Y - \bar{Y})$$

$$\hat{X} - 54 = -0.2(Y - 28)$$

$$\hat{X} = 54 - 0.2Y + 5.6 = 59.6 - 0.2Y$$

$$\hat{X} = 59.6 - 0.2Y$$

The estimated value of "X" when "Y=30"

$$\hat{X} = 59.6 - 0.2Y = 59.6 - 0.2(30) == 53.6$$

Regression line "Y on X"

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

$$\hat{Y} - 28 = -1.5(X - 54)$$

$$\hat{Y} = 28 - 1.5X + 81 = 109 - 1.5X$$

$$\hat{Y} = 109 - 1.5X$$

The estimated value of "Y" when "X=55"

$$\hat{Y} = 109 - 1.5X = 109 - 1.5(55) = 26.5$$

Example: From the following information

Arithmetic mean of "X-series=25" Arithmetic mean of "Y-series=18"

Standard deviation of "X-series=3.01" Standard deviation of "Y-series=3.03"

Sum of product of deviation from mean of "X and Y" series=12. Number of pairs of observations of "X and Y" series=15. Compute the regression line "Y on X" and estimate the value of "Y" for "X=20".

Solution:

n=15 $\bar{X} = 25$ $\bar{Y} = 18$ $X = 20$ $S_X = 3.01$ $S_Y = 3.03$ $\sum (X - \bar{X})(Y - \bar{Y}) = 12$

$$b_{YX} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{nS_X^2} = \frac{12}{15(3.01)^2} = 0.088$$

$$a = \bar{Y} - b\bar{X} = 18 - (0.088)(25) = 18 - 2.207 = 15.80$$

$$\hat{Y} = 15.80 + 0.088X$$

The estimated value of "Y" when "X=20"

$$\hat{Y} = 15.80 + 0.088X = 15.80 + 0.088(20) = 17.56$$

Q.19: From the following table, compute the coefficient of correlation by Karl Pearson's method. Arithmetic mean of "X and Y" series are 6 and 8 respectively.

X	4	6	?	2	8
Y	8	9	5	11	7

Solution: First we find missing value. Let "A" be a missing value

X	Y	X^2	Y^2	XY
4	8	16	64	32
6	9	36	81	54
A=10	5	100	25	50
2	11	4	121	22
8	7	64	49	56
$\sum X$ =30	$\sum Y$ =40	$\sum X^2$ =220	$\sum Y^2$ =340	$\sum XY$ =214

$$\bar{X} = \frac{\sum X}{n} = \frac{20 + A}{5}$$

$$6 = \frac{20 + A}{5}$$

$$30 = 20 + A$$

$$A = 30 - 20 = 10$$

$$r_{YX} = r_{XY} = \frac{n\sum XY - \sum X \sum Y}{\sqrt{\left[n\sum X^2 - (\sum X)^2\right]\left[n\sum Y^2 - (\sum Y)^2\right]}}$$

$$r_{YX} = r_{XY} = \frac{5(214) - (30)(40)}{\sqrt{5(220) - (30)^2\left[5(340) - (40)^2\right]}} = \frac{-130}{\sqrt{(200)(100)}} = -0.92$$

Q.20: Given the following information
Number of pairs of observations of "X and Y" series=15
Arithmetic mean of "X" series=25
Standard deviation of "X" series=3.01
Arithmetic mean of "Y" series=18
Standard deviation of "Y" series=3.03
Sum of products of deviations from means of "X and Y" series=122
Compute coefficient of correlation between "X and Y".
Solution:

n=15 $\bar{X} = 25$ $\bar{Y} = 18$ $S_X = 3.01$ $S_Y = 3.03$ $\sum (X - \bar{X})(Y - \bar{Y}) = 122$

$$r_{YX} = r_{XY} = \frac{\sum \left(X - \bar{X}\right)\left(Y - \bar{Y}\right)}{nS_X S_Y} = \frac{122}{15(3.01)(3.03)} = 0.89$$

Q.21: Compute coefficient of correlation between "X and Y" from the following data
Sum of deviations of "X" series = 5
Sum of deviations of "Y" series = 4
Sum of square deviations of "X" series = 40
Sum of square deviations of "Y" series = 50
Sum of the product of deviations of "X and Y" series = 32
Number of pairs of observations of "X and Y" series=10
Solution:

$n = 10$ $\sum D_X^2 = 40$ $\sum D_Y = 4$ $\sum D_Y^2 = 50$ $\sum D_X D_Y = 32$

$$r_{XY} = \frac{n\sum D_X D_Y - \sum D_X \sum D_Y}{\sqrt{\left[n\sum D_X^2 - (\sum D_X)^2\right]\left[n\sum D_Y^2 - (\sum D_Y)^2\right]}} = \frac{10(32) - (5)(4)}{\sqrt{\left[10(40) - (5)^2\right]\left[10(50) - (4)^2\right]}} = \frac{300}{\sqrt{(375)(484)}} = 0.704$$

Q.22: i) Coefficient of correlation between two variates "X and Y" is 0.8. The variances of "X" is 16 and $S_{XY} = 20$. Find the standard deviation of "Y" variate.
ii) If the coefficient of correlation between "X and Y" is -0.75, the standard deviation of "Y" series is 5 and $\sum \left(X - \bar{X}\right)\left(Y - \bar{Y}\right) = 15n$ .What will be standard deviation of "X" series.
iii) From the following information, calculate the number of items for which $r = 0.5$ $\sum \left(X - \bar{X}\right)\left(Y - \bar{Y}\right) = 120$, standard deviation of "Y" series=8 and $\sum \left(X - \bar{X}\right)^2 = 90$
Solution:

i) $r = 0.8$ $S_X^2 = 16$ $S_{XY} = 20$ $S_Y = ?$

$$r_{YX} = r_{XX} = \frac{S_{XY}}{S_X S_Y}$$

$$0.80 = \frac{20}{4S_Y}$$

$$S_Y = \frac{20}{4(0.8)} = 6.25$$

ii) $\sum \left(X - \bar{X}\right)\left(Y - \bar{Y}\right) = -15n$ $r = -0.75$ $S_Y = 5$ $S_X = ?$

$$r_{YX} = r_{XY} = \frac{\sum\left(X-\bar{X}\right)\left(Y-\bar{Y}\right)}{nS_X S_Y}$$

$$-0.75 = \frac{-15n}{nS_X 5}$$

$$-0.75 = \frac{-3}{S_X}$$

$$S_X = \frac{-3}{-0.75} = 4.0$$

iii) From the following information, calculate the number of items for which $r = 0.5$ $\sum\left(X-\bar{X}\right)\left(Y-\bar{Y}\right) = 120$, standard deviation of "Y" series=8 and $\sum\left(X-\bar{X}\right)^2 = 90$

$$S_Y^2 = \frac{\sum\left(Y-\bar{Y}\right)^2}{n}$$

$$\sum\left(Y-\bar{Y}\right)^2 = nS_Y^2 = 64n$$

$$0.5 = \frac{120}{\sqrt{90(64n)}}$$

Squaring both sides

$$(0.5)^2 = \left[\frac{120}{\sqrt{90(64n)}}\right]^2$$

$$0.25 = \frac{14400}{5760n}$$

$$n = \frac{14400}{5760(0.25)} = 10$$

Q.23: In order to find the correlation coefficient between two variables "X and Y" the following results were obtained $n = 10$ $\sum X = 500$ $\sum Y = 1100$ $\sum X^2 = 28400$ $\sum Y^2 = 134660$ $\sum XY = 61800$. It was however later discovered at the time of checking that two particular sets of observations, namely X=57 and 35, Y=121 and 114 was wrongly taken, the correct values being X=67 and 15, Y=112 and 124. Compute the correct value of the correlation coefficient.

Solution:

First we find corrected Sum and Sum of squares

$$\sum X(corrected) = \sum X(in\ \ corrected) - (incorrected\ values\ of\ \ X) + (corrected\ \ values\ of\ \ X)$$

$$\sum X(corrected) = 500 - (57+35) + (67+15) = 490$$

$$\sum Y(corrected) = \sum Y(in\ \ corrected) - (incorrected\ values\ of\ \ Y) + (corrected\ \ values\ of\ \ Y)$$

$$\sum Y(corrected) = 1100 - (121+114) + (112+124) = 1101$$

$$\sum X^2(corrected) = \sum X^2(in\ \ corrected) - (incorrected\ values\ of\ \ X)^2 + (corrected\ \ values\ of\ \ X)^2$$

$$\sum X^2(corrected) = 28400 - (57^2+35^2) + (67^2+15^2) = 28640$$

$$\sum Y^2(corrected) = \sum Y^2(in\ \ corrected) - (incorrected\ values\ of\ \ Y)^2 + (corrected\ \ values\ of\ \ Y)^2$$

$$\sum Y^2(corrected) = 134660 - (121^2+114^2) + (112^2+124^2) = 134943$$

$$\sum XY(corrected) = \sum XY(in\ \ corrected) - (incorrected\ values\ of\ \ XY) + (corrected\ \ values\ of\ \ XY)$$

$$\sum XY(corrected) = 61800 - (57\times121) - (35\times114) + (67\times112) + (15\times124) = 60277$$

$$r(corrected) = \frac{n\sum XY - \sum X\sum Y}{\sqrt{\left[n\sum X^2 - (\sum X)^2\right]\left[n\sum Y^2 - (\sum Y)^2\right]}} = \frac{10(60277) - (490\times1101)}{\sqrt{\left[10(28640) - (490)^2\right]\left[10(134943) - (1101)^2\right]}}$$

$$r(corrected) = \frac{63280}{\sqrt{[46300][137229]}} = \frac{63280}{797100.11667} = 0.794$$

Q.26: The following summary statistics were recorded $n=20$ $\bar{X}=25$ $\bar{Y}=35$ $\sum(X-\bar{X})^2=80$ $\sum(Y-\bar{Y})^2=170$ $\sum(X-\bar{X})(Y-\bar{Y})=-100$

Show that the coefficient of correlation is the geometric mean of regression coefficient.

$$r_{YX}=r_{XY}=\frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2\sum(Y-\bar{Y})^2}}=\frac{-100}{\sqrt{80(170)}}=-0.86$$

Regression line "Y on X"

$$Y=a+bX$$

$$b_{YX}=\frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(X-\bar{X})^2}=\frac{-100}{80}=-1.25$$

$$a_{YX}=\bar{Y}-b\bar{X}=35-(-1.25)(25)=35+31.25=66.25$$

$$\hat{Y}=66.25-1.25X$$

Regression line "X on Y"

$$X=a+bY$$

$$b_{XY}=\frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(Y-\bar{Y})^2}=\frac{-100}{170}=-0.60$$

$$a_{XY}=\bar{X}-b\bar{Y}=25-(-0.6)(35)=25+20.60=45.60$$

$$\hat{X}=45.60-0.60Y$$

Regression coefficient "Y on X"

$$b_{YX}=-1.25$$

Regression coefficient "X on Y"

$$b_{XY}=-0.60$$

The geometric mean of two regression coefficient

$$G.M=-\sqrt{(-1.25)(-0.60)}=-0.866$$

Hence proved that the geometric mean of two regression coefficient are equal to correlation coefficient

$$r_{XY}=\pm\sqrt{b_{YX}b_{XY}}$$

-0.866=-0.866

Proved

Q.27: The equation of two regression lines of "Y on X" and "X on Y" respectively are given from paired observations of two variables "X and Y".

8X-10Y+66=0 40X-18Y-214=0

i) Compute and interpret the coefficient of correlation

ii) Compute the values of $\bar{X}$ and $\bar{Y}$

Solution:

The equation of regression lines can be written as

$$10Y=66+8X$$

$$Y=\frac{66+8X}{10}=\frac{66}{10}+\frac{8}{10}X=6.6+0.8X$$

$Y=6.6+0.8X$ Regression line "Y on X" and regression coefficient $b_{YX}=0.8$

$$40X=214+18Y$$

$$X=\frac{214+18Y}{40}=\frac{214}{40}+\frac{18}{40}Y=5.35+0.45Y$$

$X=5.35+0.45X$ Regression line "X on Y" and regression coefficient $b_{XY}=0.45$

$r_{XY}=\pm\sqrt{b_{YX}b_{XY}}=\sqrt{0.80(0.45)}=0.60$ There is strong correlation between "X and Y"

Compute the values of $\bar{X}$ and $\bar{Y}$

$10Y-8X=66$ As we know that by the property of arithmetic mean

$10\bar{Y}-8\bar{X}=66$ (i)

$40X-18Y=214$

$-18\bar{Y}+40\bar{X}=214$ (ii)

Equation (i) multiplying by 5 and adding in (ii)

$-18\bar{Y}+40\bar{X}=214$

$50\bar{Y}-40\bar{X}=330$

$32\bar{Y} = 544$

$\bar{Y} = \frac{544}{32} = 17$

Putting the value of $\bar{Y} = 17$ in equation (i) then we get the value of $\bar{X}$

$10\bar{Y} - 8\bar{X} = 66$

$10(17) - 8\bar{X} = 66$

$170 - 66 = 8\bar{X}$

$8\bar{X} = 104$

$\bar{X} = \frac{104}{8} = 13$

Q.31: If the mean weight of 200 fathers is 140 pounds with standard deviation of 6 pounds and the mean weight of their youngest sons is 142 pounds with standard deviation of 8 pounds. The coefficient of correlation between them is 0.9. Estimate the two regression equation.

Solution: Given that

$n = 200 \qquad \bar{X} = 140 \qquad \bar{Y} = 142 \qquad S_X = 6 \qquad S_Y = 8 \qquad r = 0.9$

Regression line "Y on X"

$\hat{Y} - \bar{Y} = r\frac{S_Y}{S_X}(X - \bar{X})$

$\hat{Y} - 142 = 0.9\frac{8}{6}(X - 140)$

$\hat{Y} - 142 = 0.9\frac{8}{6}(X - 140) = 1.2(X - 140) = 1.2X - 168$

$\hat{Y} = 142 + 1.2X - 168 = -26 + 1.2X$

$\hat{Y} = -26 + 1.2X$

Regression line "X on Y"

$\hat{X} - \bar{X} = r\frac{S_X}{S_Y}(Y - \bar{Y})$

$\hat{X} - 140 = 0.9\frac{6}{8}(Y - 142)$

$\hat{X} = 0.675(Y - 142) + 140$

$\hat{X} = 0.675Y - 95.85 + 140 = 44.15 + 0.675Y$

$\hat{X} = 44.15 + 0.675Y$

Short Question

Q.1: What is meant by regression?

Ans: The dependence of a random variable (dependent variable) upon other non-random variable (independent variable) is called regression.

Example

i) Temperature depends on sunshine

ii) Good yields depend on good seeds

iii) Sale of product depends upon its quality

Q.2: What is meant by simple linear regression?

Ans: If the dependent variable depends on a single independent variable is called simple linear regression modal. Simple linear regression modal is $Y = a + bX + e_i$

Where $Y =$ Dependent variable; $X =$ Independent variable

$a =$ Intercept i.e. average value of "Y" when "X=0"

$b =$ Regression coefficient or coefficient of independent variable Slope of regression line

$e =$ Random error

"Random variation in an observational process by the variables which are not included in the modal"

Q.3: Define simple linear regression modal.

Ans: Simple linear regression modal is $Y = a + bX + e_i$

Where $Y =$ Dependent variable; $X =$ Independent variable

$a =$ Intercept i.e. average value of "Y" when "X=0"

$b =$ Regression coefficient or coefficient of independent variable Slope of regression line
$e =$ Random error
"Random variation in an observational process by the variables which are not included in the modal"
Q.4: Define intercept of straight line or regression line.
Ans: In regression modal the average value of dependent variable when there is no association is called intercept. In simple linear regression modal is $Y = a + bX + e_i$
$a =$ Intercept i.e. average value of "Y" when "X=0"
Q.5: Define regression coefficient or slope line.
Ans: The coefficient of an independent variable in a regression modal is called regression coefficient. In simple linear regression modal is $Y = a + bX + e_i$ it is denoted by b_{xy} *or* b_{yx}
$b =$ Regression coefficient or coefficient of independent variable Slope of regression line
"It measures the average change in dependent variable for a unit change of independent variable. It is free from unit of measurement"
Q.6: Define independent variable in regression modal.
Ans: A variable that provides the basis for estimation is called the predictor or Regressor or explanatory or independent variable. In simple linear regression modal is
$Y = a + bX + e_i$
Where $X =$ Independent variable.
Q.7: Define dependent variable in regression modal.
Ans: The variable that is being predicted or estimated is called the dependent or regressand or predictand or response variable. In simple linear regression modal is
$Y = a + bX + e_i$ Where $Y =$ Dependent variable.
Q.8: Write a short note on a scatter diagram.
Ans: Scatter diagram helps us to find out the relationship between two variables. If we plot "X values on X-axis and Y values on Y-axis" then joint points of (X_i, Y_i) on graph paper is called scatter diagram.
Or
A graph of the pairs of observations on two variables is called scatter diagram.
Q.9: Explain the term linear regression.
Or
Q.9: Differentiate between linear regression and curvilinear regression?
Ans: When dependence of the variables is represented by a straight line, then it is called linear regression otherwise it is called non-linear regression (curvilinear regression).
Q.10: Properties of regression line or least square line.
Ans: Properties of regression line are given below
i) The regression line always passes through the points $(\bar{X}, \bar{Y})$.
ii) Sum of residual equal to zero i.e. $\sum (Y_i - \hat{Y}) = 0$
iii) The sum of square differences between the observed and estimated value minimum.
i.e. $\sum (Y_i - \hat{Y})^2 \rightarrow \min$ *imum or least*
iv) "a and b" are unbiased estimate of "α *and* β"
v) Sum of estimated values= sum of observed values i.e. $\sum Y_i = \sum \hat{Y}$
Q.11: Properties of regression coefficient.
Ans: The properties of regression coefficient of are given below
i) It is independent change of origin but not scale.
ii) "b_{yx} *and* b_{xy}" always contain same sign i.e. "positive or negative"
iii) The Geometric mean of two regression coefficients is correlation coefficient
$r_{XY} = \pm\sqrt{b_{YX} b_{XY}}$
iv) It is asymmetric with respect to "X and Y" i.e. $b_{YX} \neq b_{XY}$
Q.12: Define correlation.
Ans: The interdependence between any two random variables is called correlation. It is denoted by "r".
Examples:
i) The height and weight of children correlated with age.
ii) Supply and demands of goods correlated with price.

Q.13: Define correlation coefficient.
Ans: The numerical value of interdependence between any two random variables is called correlation coefficient or coefficient of correlation. It is given as

$$r_{YX} = r_{XY} = \frac{\sum\left(X-\bar{X}\right)\left(Y-\bar{Y}\right)}{\sqrt{\sum\left(X-\bar{X}\right)^2 \sum\left(Y-\bar{Y}\right)^2}}$$

Q.14: Distinguish between positive and negative correlation.
Ans: If the movements of two random variables are in the same direction then it is called **positive correlation**.
Example:
i) The length of iron bar increases as temperature increases.
ii)

X	1	2	3
Y	4	5	8

If the movements of two random variables are in the opposite direction then it is called **negative correlation**.
Example:
i) The demand of items increases as price of items decreases.
ii)

X	1	2	3
Y	3	2	1

Q.15: Differentiate between perfect positive and perfect negative correlation.
Ans: If "$r = +1$" then it is said to be perfect positive correlation. This occurs when both variables move in same direction exactly at the same rate.
If "$r = -1$" then it is said to be perfect negative correlation. This occurs when both variables move in opposite direction exactly at the same rate.
Q.16: Write down the properties of the correlation coefficient.
Ans: Important properties are given below
i) The correlation coefficient is symmetric with respect to "X and Y" i.e. $r_{xy} = r_{yx}$
ii) The correlation coefficient is Geometric mean of two regression coefficients
$r_{XY} = \pm\sqrt{b_{YX} b_{XY}}$
iii) The correlation coefficient lies between "-1 and +1" i.e. $-1 \le r \le +1$
iv) The correlation coefficient is without unit of measurement
v) The correlation coefficient is independent change of origin and scale. $r_{XY} = r_{UV}$
vi) If "X and Y" are independent then $r_{XY} = 0$
Q.17: Differentiate between regression and correlation?
Ans: Regression and correlation is based on quantitative variables which are measurable. In regression we use the dependence of one variable upon another independent variable. In correlation both variable are random.
Q.18: Differentiate correlation and association?
Ans: Correlation is based on quantitative variables which are measurable. In correlation both variables are random. Association deals with qualitative variables which are non-measurable.
Q.19: Define perfect correlation.
Ans: If all the points on scatter diagram are exact on the line then it's called perfect correlation.
Perfect positive correlation $= +1$
Perfect negative correlation $= -1$
Q.20: Define linear correlation.
Ans: If all the points on scatter diagram are not exact on the line then it's called linear correlation.
Linear positive correlation $= 0 < r < 1$
Linear negative correlation $= -1 < r < 0$
Q.21: Discuss the method of lest squares.
Ans: A method which gives the sum of square residual from the fitted line is minimum or least is called method of least square. i.e. $\sum\left(Y_i - \hat{Y}\right)^2 \to \min imum$
Q.22: Write down the aims regression and correlation analysis.
Ans: i) Regression analysis provides estimates of the dependent variable for given values of independent variable.

ii) Regression analysis provides measures of the errors that are likely to be involved in using the regression line to estimate the dependent variable.
iii) Regression analysis provides an estimate of the effect on the mean value of "Y" of a unit change in "X".
iv) Correlation analysis provides estimates of how strong the relationship is between the two variables.

Q.24: What is meant by residual?

Ans: In regression analysis the difference between observed value" Y_i " and its estimated value " $\hat{Y}$ " is called residual. Sum of residual equal to zero i.e. $\sum\left(Y_i - \hat{Y}\right) = 0$

Q.25: When the two variables are said to uncorrelated?

Ans: When " $r = 0$ " the variables are said to be uncorrelated. It means that there is no linear relationship.

Q.26: Define Zero correlation.

Ans: Two random variables "X and Y" are said to be uncorrelated or no correlation if and only if "X and Y" are independent. i.e. $r = 0$