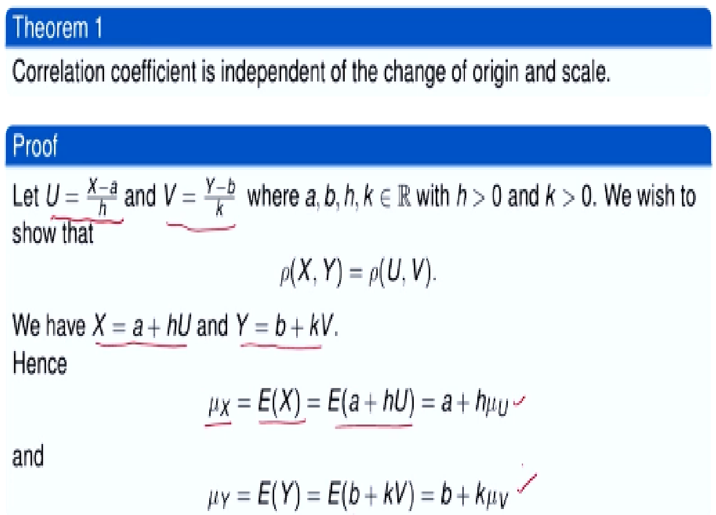
Advanced Engineering Mathematics
Prof. P. N. Agrawal
Department of Mathematics
Indian Institute of Technology - Roorkee

Lecture – 56
Correlation and Regression - II

Hello friends, welcome to my lecture; second lecture on correlation and regression, the correlation coefficient is independent of the change of origin and scale, we shall prove this result.

(Refer Slide Time: 00:46)



So, let us say, let us define the random variable U as x – a / h and the random variable V as y – b over k, where x and y are any w random variables, a, b, h, k are real numbers here, h and k are positive real numbers, okay, we wish to show that the coefficient of correlation of xy is same as the coefficient of correlation of uv that means, it is unaffected by the change of origin, okay and the scale.

So, from x =; from u = x – a/ h, okay, we find x = a + hu and from V = y – b over k, we find that, the random variable y = b + kv, hence the mean of x that is expectation of x = expectation of a + hu and so it is a times expectation of 1; expectation of 1 is 1, so we have a + h times expectation of U that is Mu U. Similarly, expectation of y Mu y = Ey = expectation of b + kV which is b times expectation of 1 which is 1.

(Refer Slide Time: 02:08)

Proof cont...

$$\begin{aligned} Cov(X,Y) &= E[(X-\mu_X)(Y-\mu_Y)] \\ &= E[(a + hU - a - h\mu_U)(b + kV - b - k\mu_V)] \\ &= E(hk(U-\mu_U)(V-\mu_V)) \\ &= hk\, Cov(U,V) \end{aligned}$$

$$\sigma_X^2 = E[(X-\mu_X)^2] = E[(a + hU - a - h\mu_U)^2] = h^2 E[(U-\mu_U)^2]$$

$\sigma_X = h\sigma_U$

and

$$\sigma_Y^2 = E[(Y-\mu_Y)^2] = E[(b + kV - b - k\mu_V)^2] = k^2 E[(V-\mu_V)^2]$$

$\sigma_Y = k\sigma_V$

Hence

$$\rho(X,Y) = \frac{Cov(x,y)}{\sigma_X \sigma_Y} = \frac{hk\, Cov(U,V)}{(h\sigma_U)(k\sigma_V)} = \rho(U,V).$$

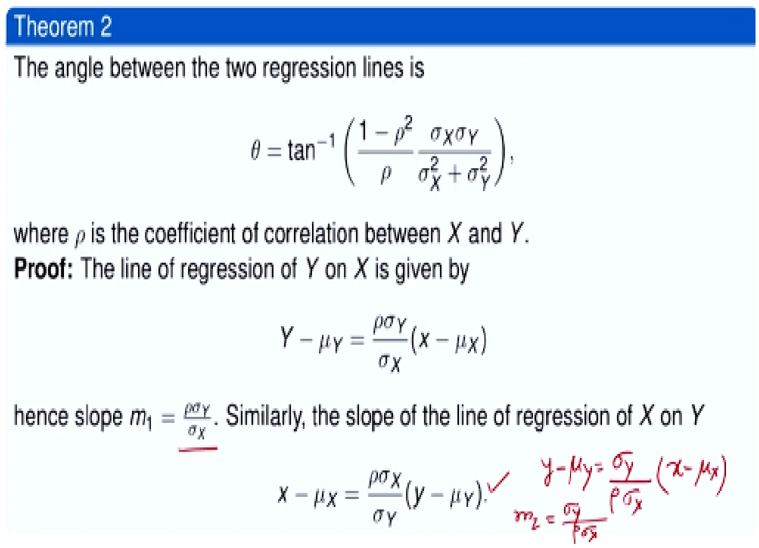
So, b + k times expectation of b that is Mu V, okay, so we get My y = V + k times Mu V, now, covariance of XY, covariance of XY by definition is expectation of X – Mu x * Y – Mu y, X = a + hu, Mu x we have just now seen is a + h Mu u, okay, so we put the value of Mu x here and y = b + kV Mu y is b + k * Mu v, so let us put their values of Mu x and Mu y, what we get here, this a will cancel with this a, b cancels with this b and what we get?

H times u – Mu u k times b – Mu v, okay, so expectation of hk times u – Mu u * v – Mu v, hk is a constant, so it will come out and we get hk times covariance of u,v okay. Now, sigma x square is expectation of x – Mu x whole square, x = a + hu, Mu x = a + h * Mu u, so this a and this a cancel and we get expectation of h square times u – Mu u whole square, so this is h square times expectation of u – Mu u whole square.

And similarly, sigma y square is expectation of y – Mu y whole square which is = k square * expectation of V – Mu v whole square, okay, so this b cancels, so this gives you k square times expectation of V – Mu v whole square. Now, rho X Y by definition is covariance of xy divided by sigma x sigma y, covariance of xy is hk * covariance of uv and sigma x square = h square times expectation of U – Mu u whole square.

So, sigma x = h times sigma u, okay and here sigma y square = k square times sigma b square, so sigma y = k times sigma b, okay, so we put their values, so sigma x = x sigma u, sigma y = k sigma v, so this hk cancels with this hk here and we get covariance of Uv divided by sigma u sigma b, so covariance of xy is same as so, rho XY is same as rho UV that is the coefficient of correlation is not affected by the change of origin and scale.

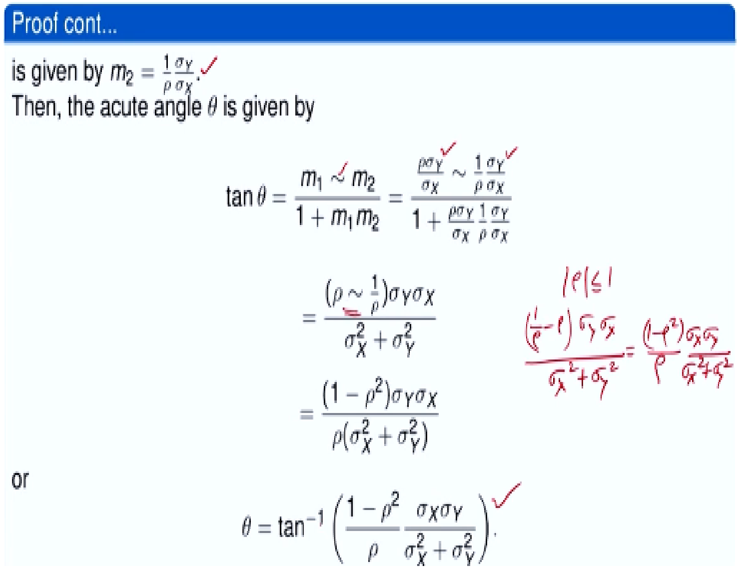
(Refer Slide Time: 04:37)



Now, let us find the angle between the 2 regression lines, the regression line of y on x and the regression line of x on y, if theta is the acute angle between the regression line of y on x and the regression lines of x on y, then theta is given by 10 inverse 1 – rho square/ rho sigma x sigma y over sigma x square sigma y square, where rho is the coefficient of correlation between x and y. Now, the line of regression of y on x we know is given by y – Mu y = rho sigma y over sigma x, x – Mu x.

So, here slope of this regression line of y on x is rho sigma y over sigma x which we denote by m1 and the similarly the slope of the regression line of x on y we can find, the regression line of x on y is given by x – Mu x = rho times sigma x over sigma y * y – Mu y and this can be written as y – Mu y = sigma y divided by rho sigma x * x – Mu x, okay, so here m2, the slope of this regression line of x on y = sigma y divided by rho times sigma x, okay.

(Refer Slide Time: 06:12)

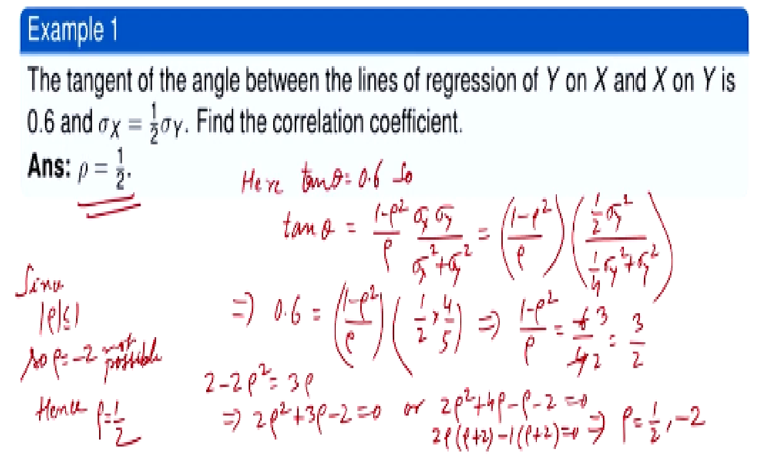


So, m1 is rho sigma y over sigma x, m2 is sigma y over rho sigma x, okay, so this is m2, now theta is given by 10 inverse m1 – m2, now this symbol means we consider the difference m1 – m2, if m1 is > m2 and we consider the difference m2 – m1, if m2 is bigger than m1, so that means we will always consider the positive sign here okay, so 10 theta = m1 – m2 or m2 – m1 divided by 1+ m1 m2, you put the values of m1 and m2 here okay, what we get?

It simplifies to this rho, now rho is mod of rho is < 1 <= 1, so we shall write it as 1/rho – rho, okay, this will be written as 1/rho – rho sigma y sigma x divided by sigma x square + sigma y square and this gives you 1 – rho square divided by rho sigma x, sigma y divided by sigma x square + sigma y square, so theta = 10 inverse 1- rho square divide by rho * sigma x sigma y divided by sigma x square + sigma y square.

Now, from here we can see if the random variable x and y are uncorrelated that is rho = 0, then theta = pi/2, 10 inverse infinity is pi/2, so that means the regression lines of y on x and the regression lines of x on y will be perpendicular to each other.

(Refer Slide Time: 07:54)

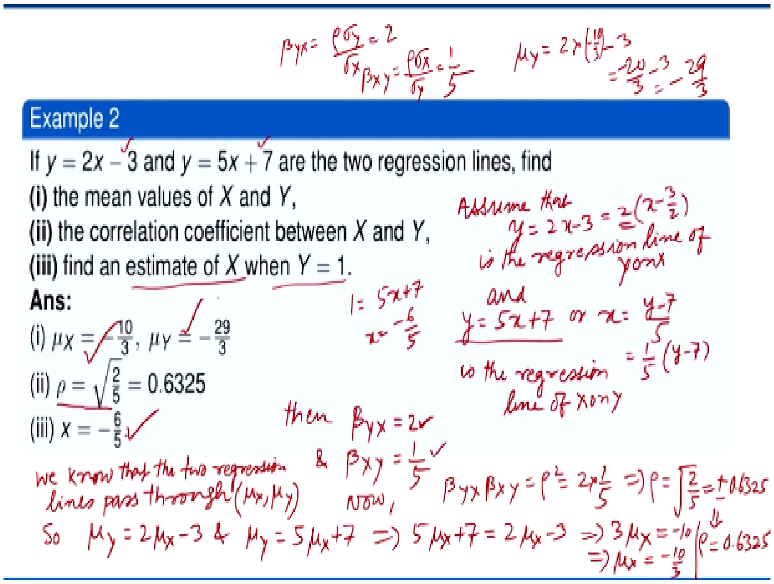


So, when x and y are uncorrelated, rho = 0 that is the 2 regression lines are perpendicular to each other, when x and y are perfectly correlated that is rho = + - 1, what we get here; 1 – rho square becomes 0, so theta = 0 that is the 2 regression lines coincide. Now, let us do this problem, the tangent of the angles between the lines of regression of y on x and x on y is 0.6, okay. So, here 10 theta = 0.6, okay, so 10 theta = 1- rho square divided by rho sigma x sigma y over sigma x square + sigma y square give you 1 – rho square divided by rho * sigma x = 1/2 sigma y.

So, 1/2 sigma y square and divide by 1/4 sigma y square + sigma y square and this is =; this gives you 10 theta 0.6, so 0.6 = 1 – rho square divided by rho and here we get 1/2 * 4/5, okay, so this implies 1 – rho square divided by rho = 6/4, okay, so 3/2, so we get 2 – 2 rho square = 3 rho, which implies 2 rho square + 3 rho – 2 = 0 and this can be written as 2 rho square + 4 rho – rho – 2 = 0, so we get 2 rho times rho + 2 – 1 times rho + 2.

So, we get the 2 roots as rho = 1/2 an d -2, okay, now since – rho is bounded by 1, okay, mod of rho is <= 1, so rho =-2, not possible and hence rho must be = 1/2 okay, so the value of rho is 1/2, okay.

(Refer Slide Time: 10:31)



Now, let us take another example, if y = 2x – 3 and y = 5x - 7 are the 2 regression lines, find the mean value of X and Y, the correlation coefficient between X and Y, find and estimate of X, when Y = 1, okay. Now, sometimes they are given the problem where the; we do not know which one is regression line of y on x and which one is the regression line of x on y, so we can choose any line as the regression line of y on x.

Then the other will be regression line of x on y and then we shall find the rho from there, okay, if the value; we shall find the beta xy and beta yx, the product of beta xy and beta yx should be = rho square and rho is <=1, so if rho square <= 1, we get; then our choice is correct, if rho square comes out to be > 1 then we will have to our assumption is wrong, so we will have to then take the other line as the regression line of y on x and (()) (11:41).

So, we will change the choice of regression line of y on x, so here we are not told which one is the regression line of y on x and which one is the regression line of x on y, so let us assume that y = 2x – 3 is the regression line of y on x and y = 5x + 7 or we can say x = y – 7/5 that is to say 1/5 times y – 7, is the regression line of x on y, okay then beta xy; beta xy is the regression coefficient of y on x, okay, so this will be = 2, okay.

Y = 2 times, I can write it as 2 times x – 3/2, okay, so this 2 gives me beta xy and similarly beta yx, no this is the regression coefficient of y on x, so I will write it as beta yx, okay beta yx is 2 and beta xy, okay, x – x bar, x – Mu x= beta xy * y – Mu y, okay, so 1/5 is beta xy, so now we know that beta yx * beta xy = rho square, okay. So, here what do we get; 2 * 1/5, so this is 2 * 1/5, okay, which is < 1, okay.

So, our choice is correct, okay, so this is regression line of y on x, this is regression line of x on y and this also gives me the value of rho, rho = square root 2/5, okay, so this is the value of rho, rho = square root 2 over 5. Now, let us find the mane values of X and Y, okay, so mean values of x and y, Mu x we have to determined, okay, we know that the 2 regression lines both pass through the point Mu x, Mu y, okay.

So, Mu x Mu y, so Mu y = 2 Mu x – 3 and Mu y = 5x; 5 Mu x + 7, oaky, so this will give you what; let us solving this 2 equations, what we will have; 5 Mu x + 7 = 2 Mu x – 3, okay, so this will give you 3 Mu x = 10, okay, this will become 3 Mux = 10, okay, so this gives me Mu x = 10/3, okay and Mu y is; Mu y is then, what will be Mu y; Mu y = 2 Mu x, so 2 * 10/3 – 3, this comes out to be 20/3 – 3, so 20 -9; 11/ 3.

What is wrong, oh, okay, okay, okay, let me do it again, we have to take rho to be; now see we have to take beta yx, okay, this is square root 2/ 5, a square root 2/5 is 0.6325, okay, this was + - 0.6325, okay. Now, our problem is whether we should take positive sign or negative sign, okay, so beta yx =; beta yx when written in terms of rho, what is the formula for that; beta yx = sigma, rho sigma y over sigma x, yeah, rho sigma y over sigma x, okay.

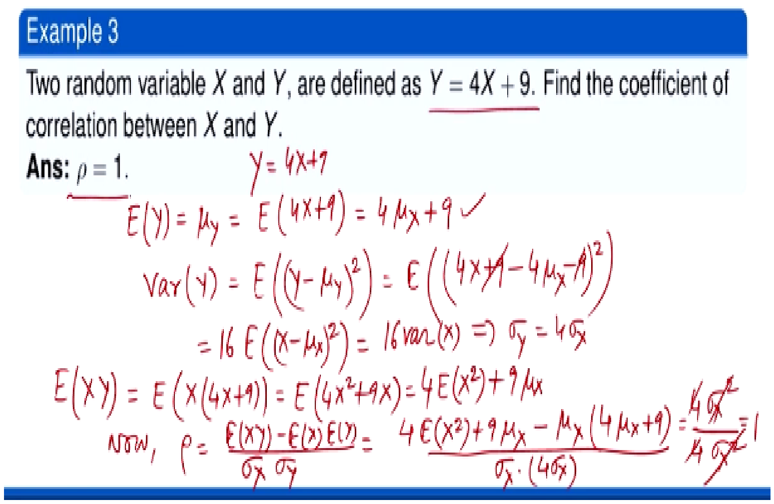
And so here, since beta yx = 2, okay, so rho must be positive okay, beta xy also, beta xy = rho times sigma x over sigma y, okay and this is the given to be = beta yx is given to be 2, okay, beta yx given to be 1/5, okay, so since beta xy and beta yx are both positive, so rho must be positive and therefore, we take the value of rho as 0.0, 0.6325, okay, we choose this sign as positive, because beta yx and beta xy are positive, okay.

Now, we know that the 2 regression lines are pass through Mu x, Mu y, so this gives you Mu y = 2 Mu x -3 and here Mu y = 5x; Mu x + 7, so 5 Mu x + 7 is 2 Mu x – 3, 5 Mu x + 7 = 2Mu x – 3, so 3 Mu x = - 10, this is -10 here, so Mu x = -10/3, okay, so Mu x + -10/3 and Mu y = 2 * -10/3 =

- 20/3 – 3, so that gives me -29/3, okay, so this is Mu y. Now, let us determined, find and estimate of x, when y = 1.

So, we have to find and estimate of y, estimate of x when y is given that means we use the regression line of x on y, regression line of x on y is this, okay, y = 5x + 7, so when y =1, okay, we have y = 1 = 5x + 7, so x = -6/5, okay, so we get x = -6/5, the value of x when y is given is found from the regression line of x on y.

(Refer Slide Time: 19:14)



Now, let me go to another question, 2 random variables X and Y are defined as Y = 4X + 9, find the coefficient of correlation between X and Y, okay, so coefficient of correlation Y = 4X + 9, okay, so let us first find the expected value of Y, okay, expected value of Y is Mu y, so this = expected value of 4X + 9, okay, so this is = 4 times expected value of X that is Mu x + expected value of 9, which is 9, okay.

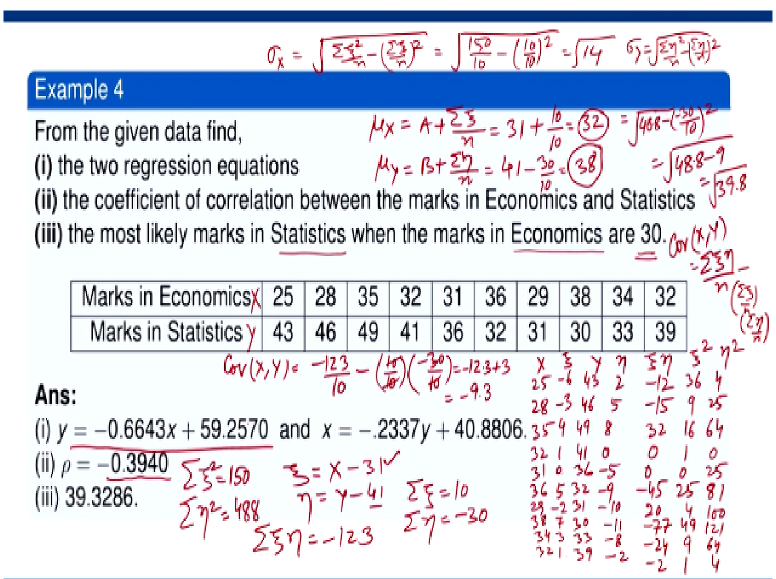
So, expected value of y is 4 Mu x + 9, okay, variance of Y, okay, let me write directly the variance of Y, so variance of Y, okay, variance of Y is expectation of Y – Mu Y whole square, okay, so expectation of Y = 4X + 9, okay – Mu Y, Mu Y = 4 Mu X + 9, so we get this, okay, so we cancel, this will cancel with this and we get 4 square, 4 square means 16, so 16 times variance of expectation of X – Mu X whole square.

So, 16 times variance of X, thus we get Mu Y = 4 times Mu X + 9 are variance of Y = this, so this implies sigma Y = 4 times sigma X, okay. Now, expectation of we need to find covariance, so expectation of X Y, so expectation of X Y is expectation of X * 4X + 9, so this is 4 times expectation of X squire, so expectation of 4X square + 9X, so this is 4 times expectation of X square + 9 times Mu X, okay. Now, rho = EXY – EX EY divided by rho X sigma X sigma Y, okay.

So, EXY is 4EX square + 9 Mu X - expectation of EX; expectation of X * expectation of Y = 4 times Mu X + 9, 4 times; expectation of X we can write as Mu X, so this is Mu X, okay and then I get here 4 times Mu X + 9, okay divided by sigma X and sigma Y is 4 sigma X, okay, now what is this; so, 9 Mu X will cancel with 9 Mu X, what we get here, this = 4 times EX square – Mu X whole square.

So, we get variance of X that is 4 times sigma X square, okay, 9 Mu X, 9 Mu X cancel and we get here 4/sigma X square, so this cancels with this and this cancels with this and we get 1, okay, so rho = 1.

(Refer Slide Time: 23:29)



Now, let us take the last problem form the given data, find the 2 regression equations; regression lines of y on x and regression line of x on y, the coefficient of correlation between the marks in economics and statistics is are given in this table, the most likely marks and statistics, when the marks in economics are 30, okay, so let us say we are given x and y, okay, x is 25, okay, 28, 35, we have 32, okay, 32, 36, then we have 29, 38, okay, 34 and 32, okay.

And y values are 33, 46, 49, 41, 36, 32, 31, 30, 33, 39, okay, let us since the coefficient of correlation is independent, it does not change with the change in origin in origin and scale, let us shift the origin, so let me write here, Xi = say x – 31, okay and y =; eta = y – 41, okay, so if we do that then Xi = ; so we are subtracting 31 here, so -6, okay, this will be -31 means -3, 31 we subtract from 35, we get 4, 31 we subtract here, we get 1, here we get 0, here we get 31 subtracted gives 5, here we get -2, here we get 7, here we get 3, here we get 1, okay.

And eta; eta will be when we are subtracting 41, so 2 and then here 5, here we get 8, here we get 0, okay, here we get -5, here we get -9, here we get -10, here we get -11, here we get -8 and here we get -2, okay, so what is sigma Xi, let us find. So, -6, -3, -9, -9 + 4; -5, -5 + 1; -4, -4 + 5; + 1, + 1 -2 is -1, - 1 + 7 is +6, 6 and 3, 9 and 1, 10, so sigma Xi is 10, sigma eta = 2 + 5; 7, 7 + 8; 15, 15 – 5 is 10, 10 -9 is 1, 1 – 10 is -9, -9 -11; -20, -20 – 8; -28 -2; -30, so we get -30, okay.

Now, then we need to get Xi eta, okay, so -6 * 2 is -12, then -3 * 5 is -15, then we get 4 * 8, 32, okay, we get 1 * 0; 0, 0 * -5; 0, 5 * -5; -25, -2 * 10; -10, + 20, 7 * -11; -77 and then we get 3 * -8; -24, then 1 * -2 is -2, okay, we also need to find Xi square, okay, so Xi square is 36, okay, then we have – 3 square is 9, 4 square is 16, 1 square is 1, 0 square 0, then we have 25, then we have 4, then we have 49 and then we have here 9 and then we have 1, okay.

And we need to find eta square, so 4, then we have 25, then we have 64, then we have 0, we have 25, we have 31, okay, - 9 square, -10 square is 100 and then we get 121 and then we get here 64 and here we get 4, okay. Now, what is sigma Xi square, okay, sigma Xi square is 36 + 9; 45, 45 + 16 is 61, okay, so we get 61 then 1, 62, 62 + 25; 62 + 25 means 87, okay, 87 + 4; 91, 91 + 49; so we get here 140, then 10, 150.

So, we get 150 here and sigma eta square we can find, okay, so this is 4 + 25; 29, 29 + 64, so we get here 93, 93 + 25, we get here 118, 118 + 81, we get here 199, okay, then 100, so we get 299, okay, 299 and then we need to add 121, 64, 4; 121 + 64 + 4; 68 we add here, so this gives me

189, okay, so 189, so 189 we add to 299, okay, so 299 we add, okay, this comes out 488, okay, so we have the values here, now sigma Xi eta also we have to find.

So, we have here -12, -15, -27, -27 + 32; + 5, + 5 – 45; - 40, -40 + 20; -20, -20 -77; -97, -97 -24 and also we have -2, so 123, okay, so - 123, okay now let us find; first we find Mu X, okay, Mu X = assumed mean, okay, A + sigma Xi over n, A is the assumed mean, assumed mean = 31, okay, so 31 and sigma Xi = 10; 10/10 because there are 10 values, okay, so this is 32 and Mu Y; Mu Y = assumed mean, assumed mean is 41 here, so B + sigma eta divided by n, okay.

B = 41 and sigma eta = -30, so -30 divided by10, so this is -3 that means -38, okay then we need to find variance sigma X, sigma Y, so sigma X, okay, sigma X is given by square root sigma Xi square divided by n – sigma Xi/ n whole square, okay, so this is square root sigma Xi square, 150/10 – sigma Xi =10, 10 divided by 10 whole square, so this is square root 15 -1 that is 14, okay.

And sigma Y similarly = square root sigma eta square divide by n- sigma eta divided by n whole square, okay, so this sigma Y = then sigma eta square, 488, so 488 divided by 10 that means 48.8 – sigma eta, sigma eta is -30, -30 divided by -3, oh sorry, -30 divided by 10 whole square, so this is square root 48.8 – 9, okay, so this is square root 39.8, okay, so this is the value of sigma Y, okay. Now, we have to find the 2 regression equations.

We have the value of Mu X, Mu Y, okay so, we need to find the value of coefficient of correlation also we need to find and the 2 regression lines are given by beta YX and beta XY, okay, so we need to find, okay, so, okay, E XY, okay, the covariance also we can find, covariance of Xi eta, okay, so covariance of XY, this is also = sigma Xi eta/ n – sigma Xi y/ n * sigma eta yn, okay.

So, this comes out to be covariance of XY sigma Xi eta Yn, sigma Xi eta is -123 divided by 10, okay – sigma Xi Yn, sigma Xi Yn is 10/10 that is 1, okay and sigma eta Y n is -30/10, so what we get, this cancels, and this cancels and we get +3, so here -123, -12.3 and + 3, so how much we get; -9.3, okay, -9.3 that is covariance; covariance of XY divided by sigma X sigma Y, okay.

(Refer Slide Time: 34:56)

$$\rho = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{-9.3}{\sqrt{14}\sqrt{39.8}} = -0.3940$$

Regression line of Y on X is given by

$$Y - \mu_Y = \frac{\rho \sigma_Y}{\sigma_X}(X - \mu_X)$$

$$Y - 38 = (-0.3940)\frac{\sqrt{39.8}}{\sqrt{14}}(X - 32) \checkmark$$

Regression line of X on Y is given by

$$x - \mu_X = \frac{\rho \sigma_X}{\sigma_Y}(y - \mu_Y)$$

$$x - 32 = -\frac{0.3940\sqrt{14}}{\sqrt{39.8}}(Y - 38) \checkmark$$

So, we need to find covariance, so sigma rho = covariance of XY divided by sigma X sigma Y, so -9.3 divided by sigma X sigma Y; sigma X = root 14, sigma Y = root 39.8, so root 14 and root 39.8, okay, it comes out to be = rho = -0.3940, so -0.3940, so this rho, okay. Now, the regression line of y on x is given by Y - Mu Y rho sigma Y over sigma X * X – Mu X, okay, so Y – Mu Y we have found.

Mu Y = 378, okay and Mu X is 32, okay, so 38, rho we have found, rho = -0.3940, okay and what is sigma Y and sigma X, sigma Y = this one, root 39.8, okay and sigma X is root 14, so root 39.8 divided by root 14 and X – Mu X is 32, okay, this Mu X 32, so this is the regression line of y on x and regression line of x on y is given by X – Mu X = sigma rho sigma X over sigma Y Y – Mu Y, okay.

So, we have X – Mu X is 32, okay, this should be small x = rho is – 0.3940 * sigma X is square root 14 divided by square root 39.8, okay * Y – Mu Y which is 38, so these 2 are the regression lines of y on x and x on y, okay, the most likely marks in statistics when the marks in economics are 30, okay. So, marks in economics we have denoted by the random variable X and marks in statistics we have denoted by the random variable Y.

We want to find the most likely marks in statistics that means we want to find the likely marks in Y, when the marks in X are given, so they can be found from the regression line of y on x, okay because the value of Y is to be found, when the value of X is given, so regression line of y on x is this one, okay, this is the regression line of y on x, in this you put x = 30, okay, so you get the value of the likely marks in statistics.

When you put S 30, Y will come out to be -0.6643 * 30 + 59.2570 which will be = 39.3286, so that is how we solve this problem, with that I would like to end my lecture, thank you very much for your attention.