Econometric Modelling

Prof. Rudra P. Pradhan

Department of Management

Indian Institute of Technology, Kharagpur

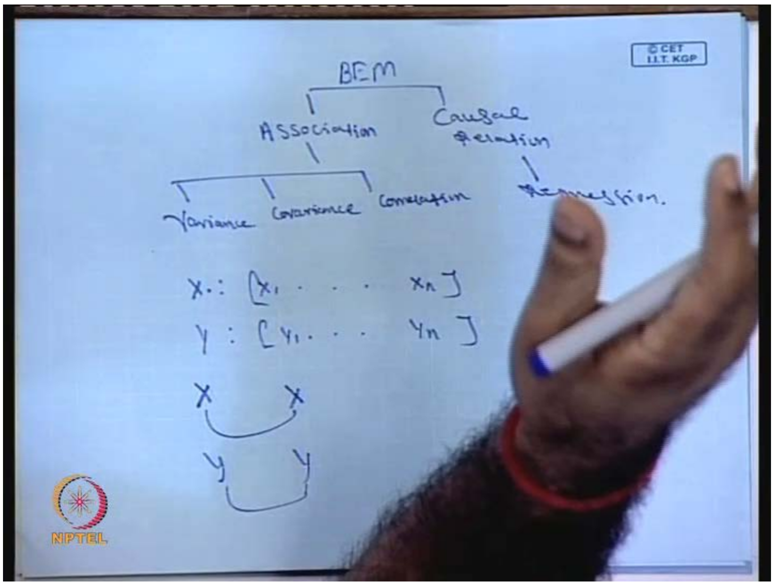
Lecture No. # 05

Bivariate Econometric Modelling (Contd.)

Good afternoon. Welcome to NPTEL project on econometric modeling. This is Rudra Pradhan here. Today, we will discuss bivariate econometric modeling, that is, with respect to regression analysis. In my last lecture, we have discussed the entire structure of bivariate econometric modelling. Basically, it is divided into two parts: first part, association between two variables; and, in the second case, we like to know the association along with the (( ))

Now, in the first case, there are several techniques: variance, covariance and correlation. And, other sides, there is a technical regression. The difference is that in the first case, particularly with respect to variance, covariance and correlation, the objective is to measure the degree of association between the two variables. However, in the case of regression, we are interested for two things: first, the association between two variables; and second, the cause and effective relationship between two variables.

(Refer Slide Time: 02:08)

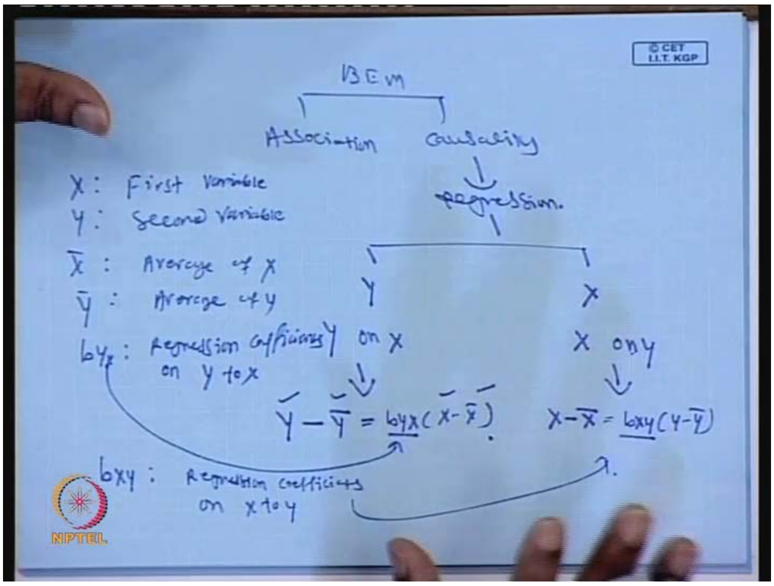


Basically, bivariate econometric modelling divided into two parts: and, that is with respect to degree of association; and, second case it is causal relation. We have standard techniques called as variance, covariance and correlation. Here this standard technique is called as regression. Now, the starting point of this particular structure is that the system must have two variables. Let us take two variables here: X represents X 1 upto X n and Y represents Y 1 upto Y n.

Now, variance and covariance are in fact very similar. Variance means we have to track the association with the same variable. For instance, we have to correlate with X upon x or Y upon y. So, this is what it is called as a variance. Now, if we correlate xy or YX, then it is a covariance. Now, similarly, if we integrate x with y or Y with X, then it is also called as a correlation. There is a small difference between covariance and correlation, but the objective of the particular study is very much similar, because we will try to know the degree of association between the two variables. Here the difference is only with respect to its mathematics only, nothing else.

Now, the case of regression – we like to know the cause and effective relationship between the two variables. In my last lecture, we have discussed the detailed structure and status of covariance and correlation; however, we have not discussed anything about the regression. Today, we will discuss in detail about regression and we have to compare how it is different or advanced to variance, covariance and correlation. One thing is very clear: regression is very much dependent on the variance, covariance and correlation. Until unless you have complete knowledge on variance, covariance and correlation, you cannot go for regression technique. Let me explain this all about regression.

(Refer Slide Time: 05:20)



BEM
Association
Causality
regression.
Y
X
Y on X
X on Y
$\bar{Y}-\bar{\bar{Y}} = b_{YX}(\bar{X}-\bar{\bar{X}})$
$X-\bar{X} = b_{XY}(Y-\bar{Y})$
© CET I.I.T. KGP
NPTEL

X : First variable
Y : Second variable
$\bar{X}$: Average of X
$\bar{Y}$: Average of Y
b_{YX} : Regression coefficients on Y to X
b_{XY} : Regression coefficients on X to Y

Now, I have already discussed the structures, bivariate econometric modeling: this is here association rule; and, this is here causality rule. Now, we will take the case causality. Now, causality is nothing but its technique called as a regression. Now, the first question is what is all about regression? Regression is to predict or forecast a particular variable with respect to a given variable. In other words, it is the average association between two variables keeping in mind two objectives: first objective – degree of association between the two variables; and, second objective is to know cause and effective relationship between the two. So, that means, if there are two variables whether X causes Y or Y causes X, in fact, in the time series modeling, we have very interesting component (( ))

There are three different options. If X causes Y and vice versa is not true, it is called as a unidirectional causality. Again, Y causes X and not Y X versa, then again it is called as unidirectional causality. If X causes Y and Y causes X, then it is called as bidirectional causality. However, we are not going in detail about the time series modelling right now; we will discuss in detail in the later part. So, in the mean time, we have to discuss what the entire structure of regression is; means, we like to know what the basics of regression modelling are.

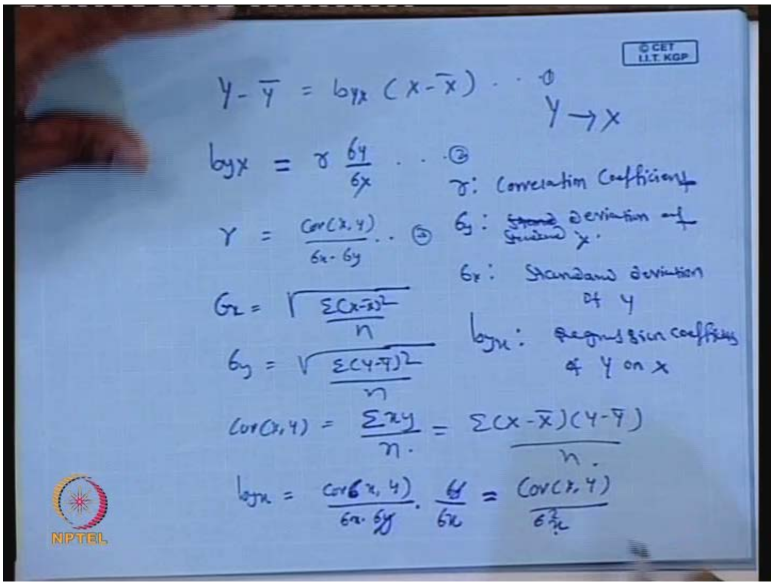
Now, regression is the average relationship between the two variables. Now, the starting point of the game must have two variables. So, let us take two variables here. So, now, regression can be obtained if there are two variables say Y and X. Now, if there are two variables, then we are very much interested for cause and effect; that means, there are two different situations here: Y on X and X on Y. Now, let us start with Y on X; then, here this side X on Y. So, this is the component Y on X, is called as a regression line – Y on X; and, X on Y means regression line from X to Y. So, now, if it is Y on X or X on Y, how is the setup or structures. Let us start with here (Refer Slide Time: 08:06). If it is Y on X, what is the step of regression? Under step of regression, simply mathematically, we can represent Y minus Y bar is equal to byx into X minus X bar. In the case of X on Y, the equation is like this, X minus X bar equal to bxy upon Y minus Y bar.

Now, certain things here you have to be very clear. First thing, we know (Refer Slide Time: 08:38) X is a first variable; this X is the first variable and Y is the second variable. Now, I have not mentioned whether first variable… it means may be dependent, may be independent. Second full – it may be dependent, it may be independent. Now, if I will say Y on X, then obviously, it is Y dependent and X independent. The moment I will say X on Y, then obviously, it is X dependent and Y independent. So, now X and Y are two variables. X bar is the average of X – mean of X; then, Y bar is nothing but average of average of Y. So, now, we get to know Y, Y bar, X, X bar; then, we have no idea about byx and bxy.

byx represents (Refer Slide Time: 09:37) a regression coefficient; it is represented as regression coefficient – coefficients on Y to X. So, this is for byx. Similarly, we have bxy; bxy represents regression on regression coefficients on X to Y. This leads to here only. So, now, the situation is very clear. So, we have two regression coefficients: first is X on Y; second is Y on X. So, now, if it is X on Y, the regression equation is X minus X bar into bxy into Y minus Y bar; and, if it is Y on X, then Y minus Y bar is equal to byx into X minus X bar. So, there are two regression lines. So, obviously, we have two regression equations.

Now, we like to know what is this structure and setup of byx and bxy. bxy is a mathematical coefficient; it is called as regression coefficient. And, bxy is also regression coefficient from X to Y; and, for byx, it is Y to X. So, now, we like to know what is exactly byx and what is exactly bxy; that means, we like to know what is the mathematics or statistics inside bxy and byx. Let me highlight here what is all about this issue.

(Refer Slide Time: 11:26)



© CET I.I.T. KGP

$$y - \bar{y} = b_{yx} (x - \bar{x}) \quad \cdots \textcircled{1}$$

$y \rightarrow x$

$$b_{yx} = r \frac{\sigma_y}{\sigma_x} \quad \cdots \textcircled{2}$$

r: Correlation Coefficient

σ_y: ~~Stand~~ Standard deviation of x.

$$r = \frac{Cov(x, y)}{\sigma_x \cdot \sigma_y} \quad \cdots \textcircled{3}$$

σ_x: Standard deviation of y

$$\sigma_x = \sqrt{\frac{\Sigma (x - \bar{x})^2}{n}}$$

$$\sigma_y = \sqrt{\frac{\Sigma (y - \bar{y})^2}{n}}$$

b_{yx}: Regression coefficient of y on x

$$Cov(x, y) = \frac{\Sigma xy}{n} = \frac{\Sigma (x - \bar{x})(y - \bar{y})}{n}$$

$$b_{yx} = \frac{Cov(x, y)}{\sigma_x \cdot \cancel{\sigma_y}} \cdot \frac{\cancel{\sigma_y}}{\sigma_x} = \frac{Cov(x, y)}{\sigma_x^2}$$

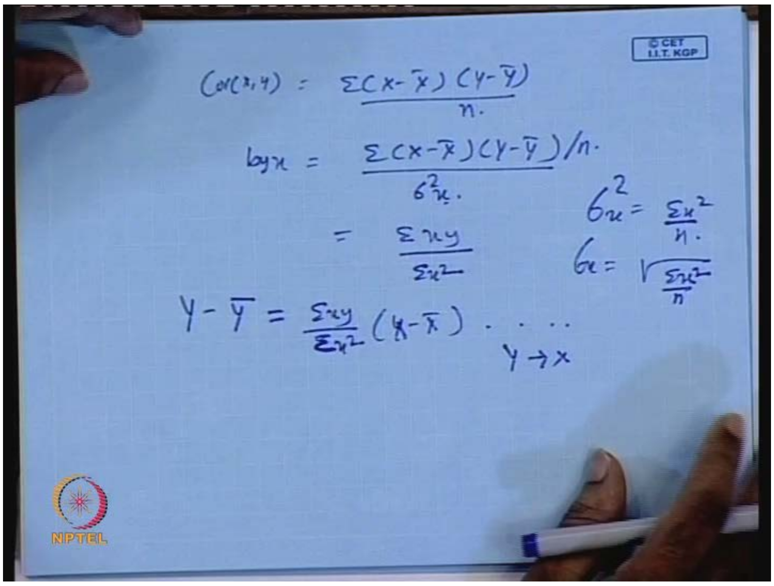
NPTEL

Now, let us start with one equation, Y minus Y bar is equal to byx into X minus X bar. So, this is regression on Y upon X. Now, here byx represents r sigma y by sigma x. What is r here? r here represents correlation coefficient and sigma y represents standard deviation of X variables; then, sigma x represents standard deviation of Y. This is standard deviation of X and this is standard deviation of Y. byx – obviously, we have already represented; this is regression coefficient of Y on X.

In the last lecture, we have discussed what is r. r is basically correlation coefficient, which is again derived through proper structures. Now, here this is the first equation; this is the second equation. Now, the third is r equal to covariance of X, Y by sigma x into sigma y. This is conditional equation – third. Now, we have the original regression equation, is Y minus Y bar equal to byx into X minus X bar followed by bxy equal to r upon sigma y by sigma x. And again, r equal to covariance of X, Y by sigma x and sigma y.

Now, we like to know what is a sigma x, what is sigma y, and what is covariance of X, Y. In fact, we have already discuss all these details. So, now, sigma x is nothing but square root of summation X minus X bar whole square divide by n; and, sigma y represents square root of summation Y minus Y bar whole square divided by n. So, now, we have covariance. So, covariance of X, Y is equal to summation xy by n. It is nothing but summation X minus X bar into Y minus Y bar divided by n. So, this is the regression coefficient; this is (Refer Slide Time: 14:13) correlation coefficient; this is standard deviation of X; this is standard deviation of Y; and, this is covariance of X, Y. So, now, if we summarize all these details, then obviously, ultimately, byx is nothing but covariance of X, Y by sigma x into sigma y into sigma y sigma x. So, now, sigma y, sigma y gets canceled; so, it is nothing but covariance of X upon Y divided by sigma square x; that is nothing but variance of X.

(Refer Slide Time: 15:06)



© CET I.I.T. KGP

$$Cov(x,y) = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{n}.$$

$$b_{yx} = \frac{\Sigma(x-\bar{x})(y-\bar{y})/n.}{\sigma_x^2.}$$

$$= \frac{\Sigma xy}{\Sigma x^2}$$

$$\sigma_x^2 = \frac{\Sigma x^2}{n.}$$

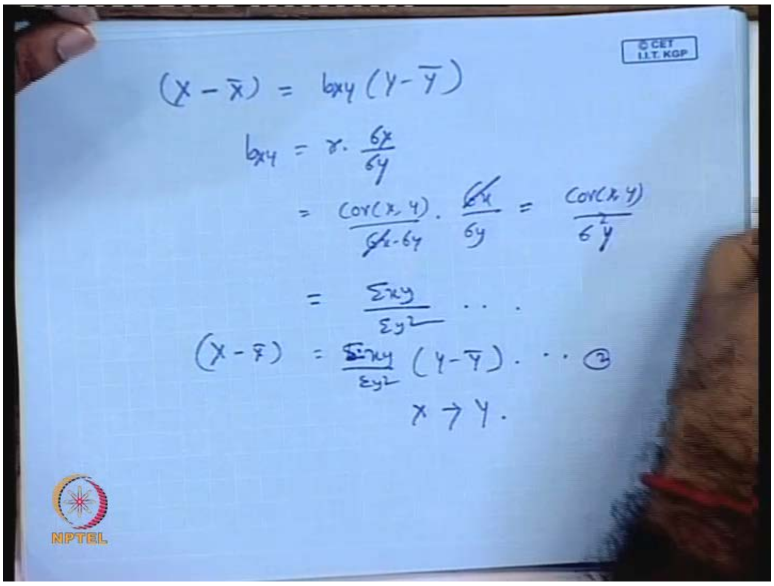
$$\sigma_x = \sqrt{\frac{\Sigma x^2}{n}}$$

$$y - \bar{y} = \frac{\Sigma xy}{\Sigma x^2}(x-\bar{x}) \; . \; . \; . \; .$$

$y \rightarrow x$

Now, this covariance of x, y is nothing but summation X minus X bar into Y minus Y bar divide by n. Now again, byx is equal to summation X minus X bar into Y minus Y bar by sigma square x. So, it is sometimes written as summation xy by summation x square. This is divided by n (Refer Slide Time: 15:43). So, obviously, summation xy by… n, n cancels. At the moment, you will take sigma x square, because sigma x square equals to summation x square by n; that means, standard deviation of x is nothing but square root of summation x square by n. So, now, this is byx. So, now, if we simplify, then it is nothing but Y minus Y bar equal to summation xy by summation x square into X minus X bar. So, this is the question of Y to X – regression equation.

(Refer Slide Time: 16:38)



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

$$b_{xy} = r \cdot \frac{\sigma_x}{\sigma_y}$$

$$= \frac{\mathrm{cov}(x,y)}{\cancel{\sigma_x}\cdot\sigma_y} \cdot \frac{\cancel{\sigma_x}}{\sigma_y} = \frac{\mathrm{cov}(x,y)}{\sigma^2_y}$$

$$= \frac{\Sigma xy}{\Sigma y^2} \;\ldots\;.$$

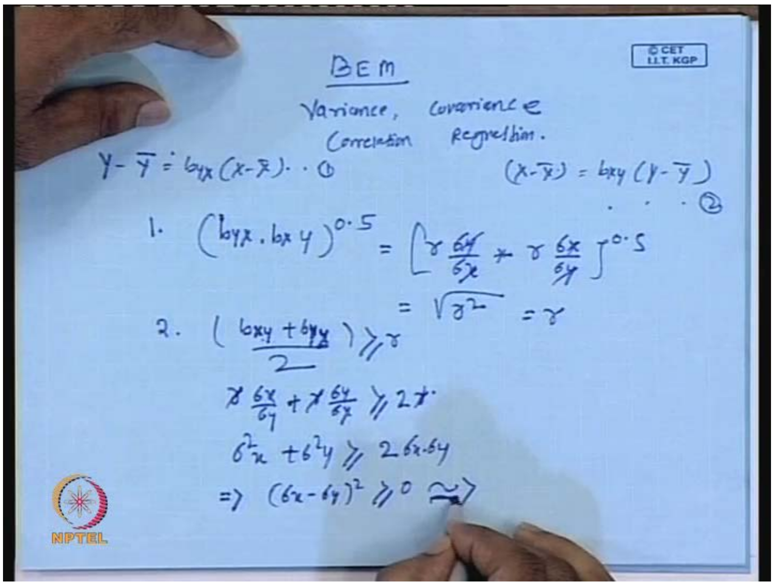
$$(X-\bar{X}) = \frac{\Sigma xy}{\Sigma y^2}(Y-\bar{Y}) \;.\;.\;.\; \textcircled{2}$$

$$X \to Y.$$

Now, come down to other part of this problem, that is X on Y. Now, for X on Y, the regression equation is nothing but X minus X bar is equal to bxy into Y minus Y bar. So, as usual, bxy is equal to r sigma x upon sigma y. Now, it is nothing but covariance of x upon y by sigma x into sigma y multiplied by sigma x by sigma y. So, sigma x, sigma x cancels; ultimately, covariance of x, y divided by sigma square y. Now, if we further simplify, then it is something – summation xy by summation y square. So, this is the final coefficient for bxy. So, ultimately, regression equation will be X minus X bar is equal to summation xy by summation y square into Y minus Y bar. So, this is the second equation of X on Y.

So, we have to variables; corresponding to two variables: Y and X, we have two regression equations: Y on X and X on Y. For Y on X, the regression equation is Y minus Y bar into byx into X minus X bar; and, for X on Y, it is nothing but X minus X bar equal to bxy upon Y minus Y bar. So, now, byx and bxy are the regression coefficients. So, now, we have to see how these two regression coefficients are integrated to each other and how it is very useful or very structure in the bivariate econometric modelling. So, let me explain here.

(Refer Slide Time: 18:43)



BEM

Variance, Covariance
Correlation Regression.

$Y - \bar{Y} = b_{yx}(X - \bar{X}) \cdots ①$

$(X - \bar{X}) = b_{xy}(Y - \bar{Y}) \cdots ②$

1. $(b_{yx} \cdot b_{xy})^{0.5} = \left[r \frac{\sigma_y}{\sigma_x} * r \frac{\sigma_x}{\sigma_y} \right]^{0.5}$

$= \sqrt{r^2} = r$

2. $\left(\frac{b_{xy} + b_{yx}}{2} \right) \geq r$

$r \frac{\sigma_x}{\sigma_y} + r \frac{\sigma_y}{\sigma_x} \geq 2r$

$\sigma^2_x + \sigma^2_y \geq 2 \sigma_x \cdot \sigma_y$

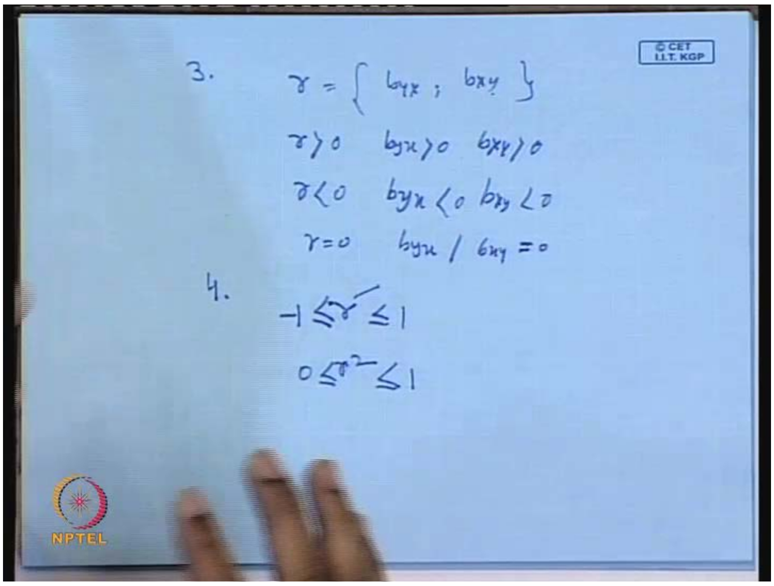
$\Rightarrow (\sigma_x - \sigma_y)^2 \geq 0$

There are various properties here, which is associated with the regression coefficient, correlation coefficient, covariance and variance. Ultimately, in this bivariate data analysis or bivariate econometric modelling, we are very much interested about variance, covariance, correlation and regression. So, we have two series: Y minus Y bar equal to byx into X minus X bar. And, another side, X minus X bar equal to bxy upon Y minus Y bar. So, now, this is equation 1; this is equation 2. So, now, we like to know how they are integrated to each other. So, that means, is there any relationship between these two equations or two regression coefficients? And, how these two regression coefficients are integrated with correlation coefficients; that to variance, covariance and structure?

Now, let us start with here bxy. So, one standard property is that the geometric mean of two regression coefficients is equal to correlation coefficients. What is geometric mean? Now, byx into bxy – these two regression coefficients will be like this – 0.5. So, what is byx and what is bxy? It is already mentioned. So, byx is nothing but r sigma y by sigma x multiplied by r sigma x upon sigma y. So, sigma x, sigma x cancels; sigma y, sigma y cancels. This is to the power 0.5 (Refer Slide Time: 20:38). So, it is simply r square upon square root. So, this means r. So, now, the physical interpretation is that the geometric mean up to regression coefficient is the correlation coefficients. So, that means, if we have two regression coefficients, then we can get to know the correlation coefficient. That is simply the geometric mean of byx and bxy.

Now, this is the second property. The arithmetic mean of the two regression coefficients is greater than two correlation coefficients. What is arithmetic mean? Now, for arithmetic mean, bxy and byx is nothing but bxy plus byx by 2 greater than equal to correlation coefficient. So, now how is the structure? It is nothing but r sigma x by sigma y plus r sigma y by sigma x greater than equal to 2r. So, now, r, r, r cancels. So, sigma square x plus sigma square y greater than equal to 2 sigma x into sigma y. So, this implies sigma x minus sigma y whole square should be greater than equal to 0. So, it is a meaningful statement. So, that means, we can justify that the arithmetic mean of two regression coefficients should be always greater than equal to correlation coefficient; by any chance, it cannot be less than that. So, this is the second issue of the association between regression coefficients and correlation coefficients.

(Refer Slide Time: 22:42)



3. $r = \{ b_{yx} ; b_{xy} \}$

$r > 0 \quad b_{yx} > 0 \quad b_{xy} > 0$

$r < 0 \quad b_{yx} < 0 \quad b_{xy} < 0$

$r = 0 \quad b_{yx} / b_{xy} = 0$

4. $-1 \leq r \leq 1$

$0 \leq r^2 \leq 1$

Third property here – you know, r depends upon byx and bxy. So, correlation coefficient simply represents or functional association between byx and bxy. Now, it is very interesting. If r greater than 0, then byx is obviously greater than 0; bxy greater than 0. So, that means, for if byx and bxy are positive, then r must be positive. Then, if r is less than 0, then byx less than 0, bxy less than 0; or, if r equal to 0, then byx or bxy is equal to 0. So, that means, both regression coefficients and correlation coefficients are usually same signed, by any chance it cannot be different. For instance, if regression coefficients are negative, then obviously, correlation coefficient will be negative, because it is the geometric mean of the two. So, obviously, both should be positive, so that we will get the positive correlation coefficient; and, both should be negative to get the negetive correlation… So, in one instance, bxy is positive and in other instance, bxy should be positive; it cannot be other way around. So, this is how the third property is all about between regression coefficients and correlation coefficients.

Now, you must be very much concerned about the coefficient of correlation and the coefficient of regression. We have already mentioned that r, correlation coefficient always lies between minus 1 and plus 1. Now; obviously, r square lies between 0 to 1. So, this is correlation coefficient (Refer Slide Time: 25:06) and this is the square of correlation coefficient. In fact, in (( )) analysis, when we go deep into the regression, obviously, the r square component is a very important vector; means, it has lots of beautiness. So, we will discuss in detail what is all about r square and how it is very useful for entire regression issues. So, in the mean time, I like to know it is simply r square; that means, square of correlation coefficient. And, if we go into deep, then the r square is represented as the coefficient of the determination and that is the measure of goodness fit test.

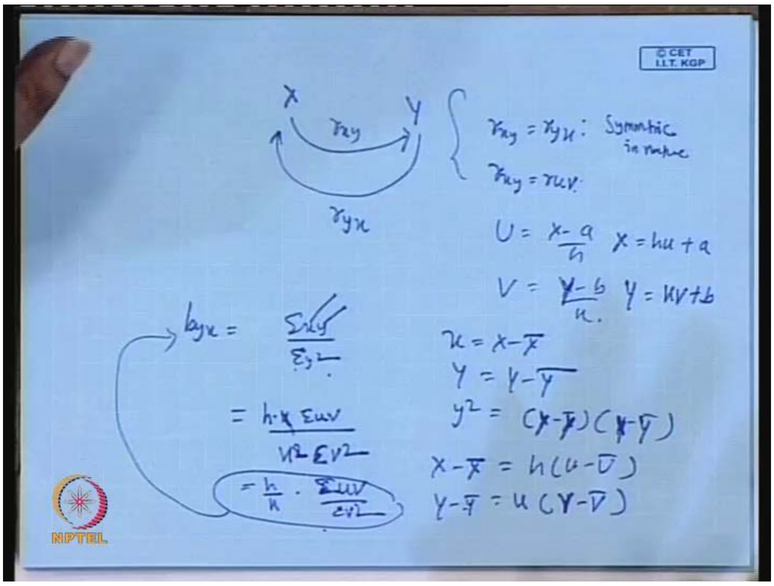
In the first lectures when I mentioned about the structure of econometric modeling, I have discussed the entire setup – how you start with econometric modelling and how you end with the econometric modeling. In the very beginning, I have mentioned, the first starting point is to define the problem that you have to borrow from the theory; then, you have to transfer the theory into mathematical form of the model; then, you have to transfer the mathematical form of the model to statistical form of the model; then, we have to investigate that models. And, for that, you need to have information; that is what we call data. So, the moment you have data, then you have to apply the statistical technique tools for computational process and you the estimated (( )). Now, what you have estimated model in your front, then obviously, the first assignment is to check the reliability before you like to go for forecasting or other issues.

I clearly mentioned during that times that so far as the reliability check is concerned, we have three different specifications. See three different (( )) that is, goodness fit test, then specification test and out of sample prediction test. And, goodness fit test is one of the issues here only. So, now, what we are talking about (Refer Slide Time: 27:04) r square is nothing but the goodness fit test. So, the goodness fit of that particular model depends upon the value of r square. If the r square value is very close to 1, then obviously, the fit of the model is better. So, if the r square value is close to 0, then the model cannot be better fitted; that means, if the goodness fit is not reliable, it cannot give any positive indication, obviously, we cannot go for forecasting; that means, we have to apply, go back to the second stage. So, the way we have structure or given a detailed structure about the flowchart, accordingly, we have to proceed. So, now, by any chance if r square is close to 1, then that means, model is reliable, then you can go for forecasting; and, that is observed or that can be done with the basis of only goodness fit test.

Now, here the point is that (Refer Slide Time: 28:00) if r is in between minus to 1, then obviously, coefficient determinant limit is 0 to 1. Now, if the value of correlation coefficient is minus 1, it is positive; means, perfectly negatively associated with each other. And, if it is equal to 1, then it is positively associated to each other. And, that is perfect positive correlation; and, this is perfect negative correlation. So, in between 0 must be there; 0 means no correlation between the two. So, that means, if the r value is 0, then obviously, the causality factor will not come into the picture, because if we take any regression equation: Y on X or X on Y, then obviously, byx and bxy are the factors. So, the moment you have r coefficient 0, then obviously, the entire issue – byx and bxy are also 0. So, as a result, regression equations: Y on X and X on Y cannot be observed. So; that means, from a correlation coefficient itself it will give indication whether there is any cause and effect relationship, because it is the essential point whether you have to proceed further; that means, if you are staring from the variance, covariance, correlation and regression, then it is just like a step-by-step process.

Now, if the correlation gives 0 results, then obviously, no point to go for regression. The region is that because regression is the advance technique; it is very time taking; and also, mathematically very complex. So, the moment you will get r equal to 0, then you can stop there; there is no point to discuss about cause and effect relationship, because the relation itself has no meaning at all. Now, if the correlation coefficient minus 1 to 1, then obviously, coefficient of determination is 0 to 1. So, accordingly, the goodness of fit will give the forecasting results. If it is close to 1, then it is better forecasted; if it is close to 0, then there is no question of forecasting, but if it is close to 0, then it is less reliable for forecasting.

(Refer Slide Time: 30:35)


X
Y
r_{xy}
r_{yx}
© CET I.I.T. KGP
NPTEL

$r_{xy} = r_{yx}$: Symmetric in nature

$r_{xy} = r_{uv}$

$U = \frac{X-a}{h}$ $X = hu + a$

$V = \frac{Y-b}{k}$ $Y = kV + b$

$x = X - \bar{X}$

$y = Y - \bar{Y}$

$y^2 = (Y-\bar{Y})(Y-\bar{Y})$

$X - \bar{X} = h(u - \bar{u})$

$Y - \bar{Y} = k(V - \bar{V})$

$b_{yx} = \frac{\Sigma xy}{\Sigma y^2}$

$= \frac{h \cdot k \, \Sigma uv}{k^2 \Sigma v^2}$

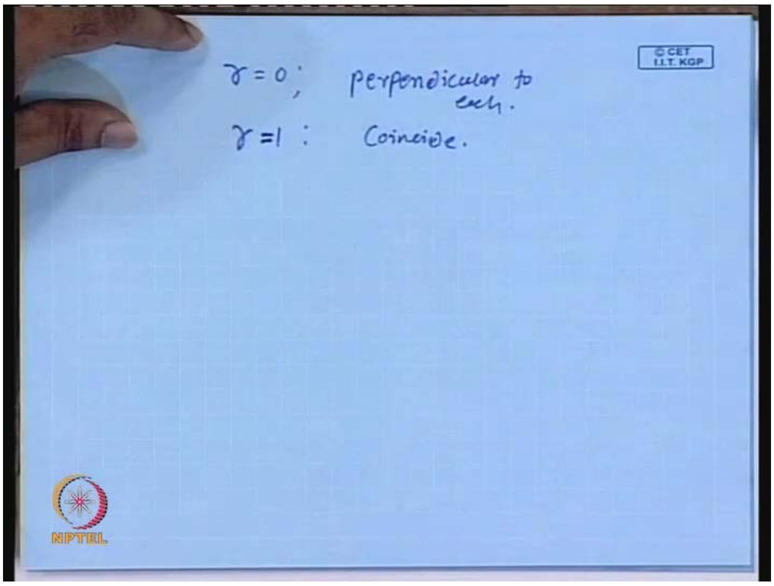
$= \frac{h}{k} \cdot \frac{\Sigma uv}{\Sigma v^2}$

Now, corresponding to this particular setup, we have another property associated with the regression coefficient, correlation coefficient. Structure is that here we define the correlation structure; the correlation is very symmetric in nature; so that means, r yx equal to r xy. For instance, like this, we have two variables X and Y; that means, we can correlate X upon Y or we can correlate Y upon X. If will correlate X on Y is called as r xy; and, if will be correlate Y on X, then it is called as r yx; that means, the fundamental theorem is that r xy is equal to r yx; that means, it is simply symmetric in nature. And, one of the other important point is that r xy is equal to r uv; u and v are other variables, which is other way representation of xy. For instance, u can be X minus a by h and v can be X minus b by k; that means, it is represented as the (( )) coefficient is independent of change of origin (( )).

However, in the case of regression, it will not be an issue; in the case of regression, it is change of origin, but, not this scale (Refer Slide Time: 31:41). The reason is that byx is simply summation xy by summation y square. So, now, if x represents here X minus X bar and y represents here Y minus Y bar and y square represents Y minus Y bar into Y minus Y bar. So, now if we simplify here, then X equal to hu plus a and here v equal to Y minus b… In fact, Y equal to kv plus b. So, now, X minus X bar is nothing but h into u minus u bar and Y minus Y bar represents k into v minus v bar. So, now, if it is simplified, the expression bxy is nothing but h into k summation uv divided by summation X square – means it is simply k square into summation v square.

Now, this k, k cancels. Then, ultimately, we have the factor h by k into summation u v by summation v square; so, that means, clear cut indication is that regression coefficient is independent of origin, but not the scale. Similarly, you can make a verification for bxy. So, the point is that correlation coefficient is independent of change of origin and scale; whereas regression coefficient is change of origin, but not the scale. So, this is the it case of regression issue and correlation issue.

(Refer Slide Time: 33:50)

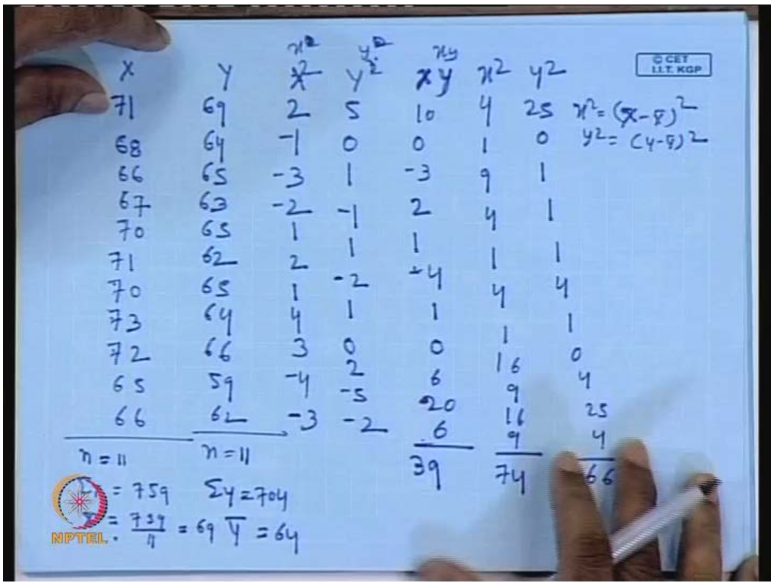


$r = 0$; perpendicular to each.

$r = 1$: Coincide.

Now, there is another point you can hear and note down on this. When r equal to 0, then obviously, the two regression lines are perpendicular to each other. So, if r equal to 1; means, if the correlation coefficient is equal to 1, then it means two lines coincide; so; that means, if r equal to plus minus 1, then two regression equations are usually different and there is an exact relationship between the two. If regression coefficients are not plus minus 1 or it is equal to 0, then obviously, there is no relationship between the two. So, that means, there are three different situations altogether. If it is plus minus 1, then there is a perfect relationship, perfect association between the two. If it is less than that, then there is relation, but it is not perfectly related to each other. However, if it is equal to 0, then there is no question of association and also there is no question of causality. So, this is the basic background of the regression analysis.

(Refer Slide Time: 35:12)



X	Y	x	y	xy	x^2	y^2
71	69	2	5	10	4	25
68	64	-1	0	0	1	0
66	65	-3	1	-3	9	1
67	63	-2	-1	2	4	1
70	65	1	1	1	1	1
71	62	2	-2	-4	4	4
70	65	1	1	1	1	1
73	64	4	0	0	16	0
72	66	3	2	6	9	4
65	59	-4	-5	20	16	25
66	62	-3	-2	6	9	4
$n = 11$	$n = 11$			39	74	66

$x^2 = (X - \bar{X})^2$

$y^2 = (Y - \bar{Y})^2$

$\Sigma X = 759$ $\Sigma y = 704$

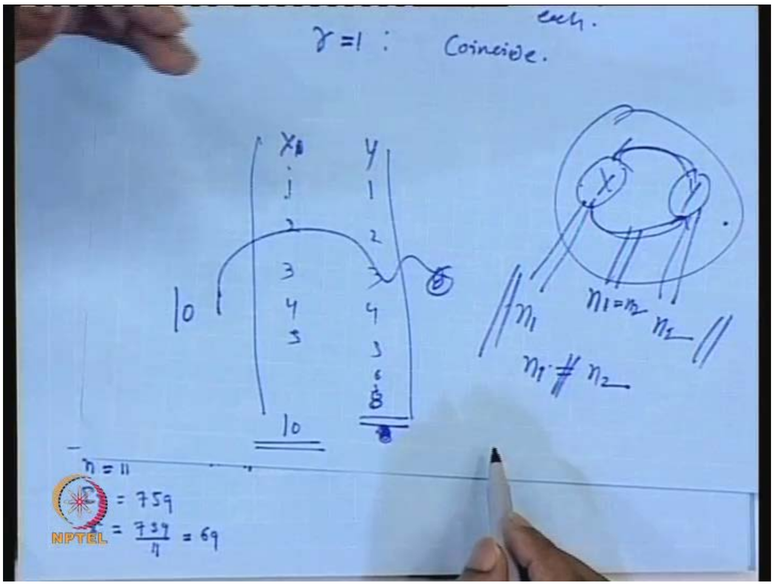
$\bar{X} = \frac{759}{11} = 69$ $\bar{Y} = 64$

© CET I.I.T. KGP

Now, we will explain with the h-v table examples here. So, how it is exactly structured as far as the regression equation is concerned. Let me take example here. This is X series and this is Y series. So, (( )) 71; then, this is 68; then 66, then 67, then 70, 71, 70, 73, 72, 65, then 66. So, 1 2 3 4 5 6 7 8 9 10 11; so, that means, here n equal to 11. Now, it will make sum; then, this is sum X equal to 759. So, obviously, X bar is equal to 759 divided by 11, which is nothing but 69.

Now, come to Y structure. For 71, Y is 69; for 68, Y is 64; then, for 66, it is 65; then, for 67, it is 63; then 65; then, this is 62; 65, 64; then, 66, 59, then 62. So, obviously, n is here 11, because in the very beginning, I have mentioned (( )) which will go for variance, covariance, correlation and regression. The essential condition is that the sample observations most be same and unique. If the sample observations are not same, then you cannot make any association or you cannot correlate; you cannot regress; you cannot covariate. This is in order of the case when you go for univariate setup.

(Refer Slide Time: 37:17)



For instance, within this particular structure like this, we are just… This is X within this system and this is Y within this system. We are doing just our assignment Y to X or X to Y either to regress or to correlate. But, this is what? Bivariate. Within the bivariate, if we are interested about X or we are interested about Y, then there is a no point if this particular structure we call it n 1 and this particular structure sample observation we call it n 2, there is a no point that n 1 should be exactly equal to n 2. But, if we correlate between these two, in that case, it is mandatory, n 1 should be equal to n 2. But, in this case – this particular case and this particular case, n should not be exactly equal to 2; that means, for univariate analysis, any univariate statistics if you look, there is no point or there is no question about the sample observation information, because uniform of sample observation within the system represents the structure of univariate and bivariate and multivariate. Our objective is completely different in the case of bivariate or multivariate; we just like to know how they are related to each other. And, for that, we need to have information about the univariate statistics.

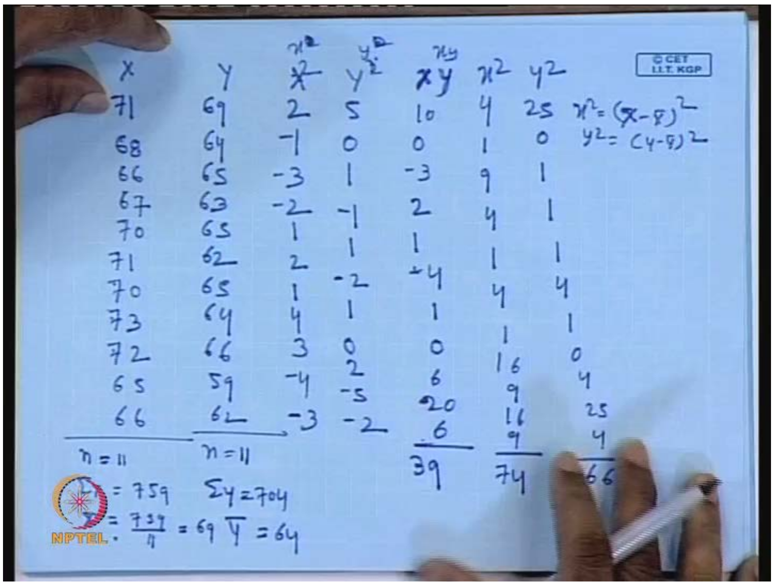
Now, within the setup, if you are looking for univariate statistic of particular variable, then this is completely different. Now, for another variable, the univariate statistics is completely different; that means, altogether they are independent. But, when you like to integrate each other with respect to (( )) correlation (( )) regression, that time this full unit should be very similar in nature. In the first case, it is not mandatory; but, in the second case, it should be mandatory.

Yes, there is certain issues here is; if (Refer Slide Time: 39:24) we have X variable and Y variable, here 1, 2, 3, 4, 5 up to 10 observations. Another case, we have 1, 2, 3, 4, 5, 6 up to 8. So, now, the observations 6, 7 or take this as 8. This is 8. So, now, what we have to do, in that case, so far as the univariate statistics is concerned, you can do an analysis here; you can do analysis here (Refer Slide Time: 39:55), but when we apply bivariate modeling here, then in that case, the system is totally inconsistent, because this is the sample observation of 8 and this is the sample observation of 10. So, now this is nothing but inconsistency. So, to solve this particular problem or to handle the particular issue, what we have to do? We have to artificially create uniform sampling; so, that means, either you can reduce full size or increase the sample size to 10; means 10 is already there for X, but in the case of Y, it is not there. So, it is only 8. So, we can extend 9 and 10 further.

You have to extend 9 and 10 further. For that, either we have to explore that information is available; if it is so, then your task is very easy, you can go head. But, sometimes in the real world business, you may not have information, but there is a standard mathematical technique through which you can fill that gap also. Here one of the standard techniques is called as interpolation and extrapolations. If we apply interpolation and extrapolation, then the sample unit 8 can be extended to 10. So, in that case, you will get the a uniformity in the structures. Then, of course, you can go ahead with the solutions. But, every time you can apply interpolation and extrapolation.

However, there are certain problems associated with the interpolation and extrapolation. It can solve the problem; it can get the model splited; you can go for forecasting; you can go for anything, etcetera, but it will affect the liability part of the model. So, the moment you will go for interpolation and extrapolation to enhance the sample size or to get the uniformity in the sample, then one of the standard problem you can face is that called as a correlation, which is very complex, very serious and very interesting also. We will discuss in detail when we go for the autocorrelation modeling. So, right now, it is not an issue here. So, in the first hand, we just want to know how we have to solve this particular problem. Later on, when there is additional problem or additional complexity, so far as the reliability check is a concerned, that time we have some other tricks how you have to eradicate that problem. So, we will discuss in detail when we go for that.

(Refer Slide Time: 42:42)



X	Y	x	y	xy	x^2	y^2
71	69	2	5	10	4	25
68	64	-1	0	0	1	0
66	65	-3	1	-3	9	1
67	63	-2	-1	2	4	1
70	65	1	1	1	1	1
71	62	2	-2	-4	4	4
70	65	1	1	1	1	1
73	64	4	0	0	16	0
72	66	3	2	6	9	4
65	59	-4	-5	20	16	25
66	62	-3	-2	6	9	4
n = 11	n = 11			39	74	66

$x^2 = (X-\bar{X})^2$

$y^2 = (Y-\bar{Y})^2$

$\Sigma X = 759$ $\Sigma Y = 704$

$\bar{X} = \frac{759}{11} = 69$ $\bar{Y} = 64$

© CET I.I.T. KGP
NPTEL

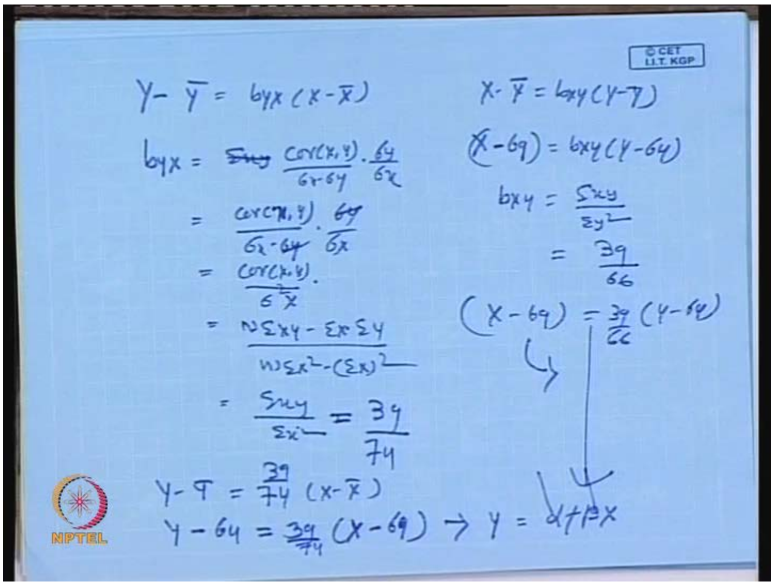
In the mean time, we have two observations: X and Y. X contains this much of information; Y contains this much of information. Now, in the first hand, sample observations are similar; that means we can proceed further for the analysis. So, origin or (( )) In fact, sometimes when there is series of observations. This is 11 – by look, we can say that there is inconsistence. And, there are two variables only; then, you can say that there is inconsistence. But, when there is multiple variables and multiple sample points, that time it is very difficult to observe. Yes, we have standards softwares. So, we just enter the data, then we crosscheck it. For all these variables, the moment you will put the descriptive statistics, it will give you indication what is the observation n for all the variables and what is the mean of all the variables, what is standard deviation, variance – all these descriptive statistics it will give you in detail.

Now, with the available information X and Y, we need to find out X squares, we need to find out Y squares, we need to find out X Y; then, we have to proceed for the regression coefficient. To simplify further or the structure because of its simplicity, we can go for small x square small y square and small xy. So, here (Refer Slide Time: 44:11) small x square is represented as X minus X bar whole square and small Y square represents Y minus Y bar whole square. X bar is 69 here. So, corresponding to this Y, summation Y is equal to 704. So, n is 11 So, obviously, Y bar equal to 704 by 11, which is nothing but 64.

Now, if we transfer it, then (Refer Slide Time: 44:44) every item has to be transferred into X minus X bar; that means, for first case, it should be X minus 69. So, now, if we transfer, then this structure will come to minus 1, minus 3, minus 2, then 1, 2, 1, 4, 3, minus 4, minus 3. Similarly, in the second case, this is in fact, x; this is in fact x (Refer Slide Time: 45:19). So, obviously, we will go for X square and Y square. So, for Y, it is nothing but 5, 0, 1, minus 1, 1, minus 2, 1, 0, 2, minus 5, then minus 2. Now, this is X deviation format and Y deviation format. So, now, we need XY. So, XY is 10 and 0, minus 3, 2, 1, minus 4, then 1, 0, then 6, then 20, then 6. Similarly, I will get x square and you will get y square. x square represents 4, 1, 9, 4, 1, 4, 1, 16, then 9, then 16, then 9. Then, corresponding Y, we have y square 25, 0, 1, 1, 1, 4, 1, 0, 4, 25, 4. So, now, we have explain x square, we have y square, and we have xy; of course, it is in deviation format. So, now, summation xy is equal to here 39; then, summation x square is equal to 74; and, summation y square is here 66.

Now, we have to see how the regression coefficient is, how the correlation coefficient is. So, now given information with (( )) detail about its statistics: invariate statistic and bivariate statistics, that is, with respect to variance and covariance, we have to proceed further for regression and also its correlation coefficient.

(Refer Slide Time: 47:24)



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

$$b_{yx} = \cancel{S_{xy}}\ \frac{cov(x,y)}{\sigma_x \cdot \sigma_y} \cdot \frac{\sigma_y}{\sigma_x}$$

$$= \frac{cov(x,y)}{\sigma_x \cdot \sigma_y} \cdot \frac{\sigma_y}{\sigma_x}$$

$$= \frac{cov(x,y)}{\sigma^2_x}$$

$$= \frac{N\Sigma xy - \Sigma x \Sigma y}{N\Sigma x^2 - (\Sigma x)^2}$$

$$= \frac{S_{xy}}{\Sigma x^2} = \frac{39}{74}$$

$$Y - \bar{Y} = \frac{39}{74}(X - \bar{X})$$

$$Y - 64 = \frac{39}{74}(X - 69) \rightarrow Y = \alpha + \beta X$$

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

$$(X - 69) = b_{xy}(Y - 64)$$

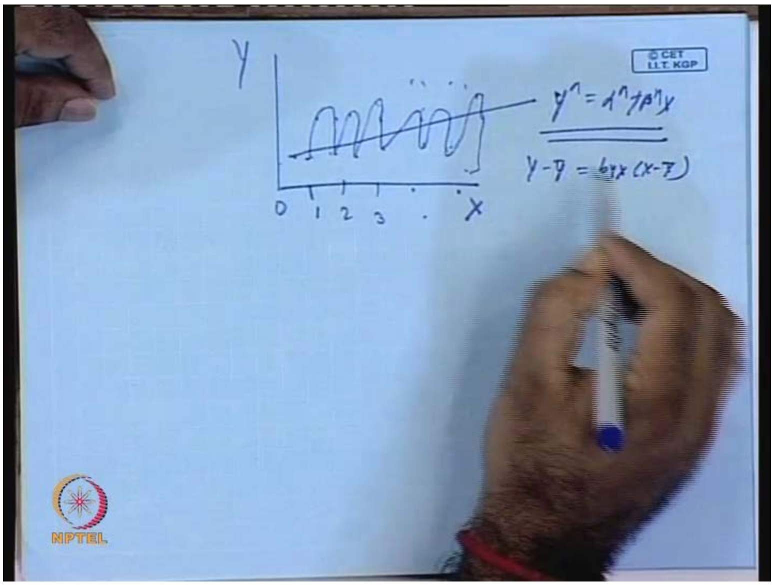
$$b_{xy} = \frac{S_{xy}}{\Sigma y^2} = \frac{39}{66}$$

$$(X - 69) = \frac{39}{66}(Y - 64)$$

Now, we have two different equations: Y minus Y bar equal to byx into X minus X bar; and, other side, X minus X bar is equal to bxy into Y minus Y bar. Now, first of all, we calculate what is byx. byx is equal to covariance of x, y by sigma x into sigma y into sigma y by sigma x. So, sigma y sigma y cancels. So, covariance of x, y by sigma square x. To the simplify, it is nothing but N summation xy minus summation x into summation y divide by N summation x square minus summation x whole square. If we further simplify, then it is something summation xy by summation x square; that means, summation xy here is 39. So, 39 divided summation x square – is 74 here. So, this is what the regression coefficient is.

Now, the equation will be Y minus Y bar is equal to 39 by 74 into X minus X bar. In other words, Y minus Y bar – is 64 is equal to 39 by 74 into X minus X bar, that is, 69. Now, if we simplify, then this will be simply in the format of Y equal to alpha plus beta X; alpha and beta – supporting components. Similarly, in the case of X minus X bar, it is nothing but X minus 69 into bxy Y minus 64. So, what is bxy? bxy is equal to summation xy upon summation y square. So, this is nothing but 39 by summation y square is 66. So, X minus 69 into 39 by 66 into Y minus 64. If we simplify again further, you will get in the format of Y equal to alpha plus beta X. Now, here alpha and beta are very supporting factors; beta is the slope of this particular line. So, beta is the real structure, where it means it is the main regression coefficient, what we call it byx and bxy. Alpha will just give you the indication where the line exactly starts.

(Refer Slide Time: 50:52)


Y
0
1
2
3
X
$\hat{y} = \hat{\alpha} + \hat{\beta} x$
$y - \bar{y} = b_{yx}(x - \bar{x})$
© CET
I.I.T. KGP
NPTEL

Let us take a case here. Whatever information we have since we are going for Y on X or X on Y, then obviously, the moment is like this. If we put it here X and Y and if we plot all these points, then we get to know how is the set. Usually if we (( )) then the structure will be this. It will be like this. So, now, for every sample units 1, 2, 3, 4, 5 like this, then obviously, there is some Y observations. So, now, here the moment will be like this. It will connect each and every point and the moment will be like this. So, within the moments, we like to know how is the path; that means, this path is called as the line of the best fit. This is what we call it as Y head equal to alpha head plus beta head X. This is the estimated equation, which we derive from Y into Y bar equal to byx into X minus X bar. The detail calculation procedure we get to know when we will go for the exact econometric modelling and regression modeling. We are not discussing the detail issue about the structure and setups; we are just briefing what is all about regression analysis. Once we enter to this – the structure of bivariate and multivariate in a research angle or practical problem angle, then obviously, you can get to know how complex it is, how it is derived really. So, (( )) structures.

Now, we will summarize this entire concept – what is all about bivariate regression modeling; what is the structure about variance, covariance, correlation and regression. So, the basic of objective behind bivariate modelling is that we like to know what the association between the two variables is; that means, the fast condition is that in a particular system problem setup we must have two variables. This is the first condition. And, for particularly covariance, correlation and regression, then the second important point is that both the variables have same number of observations. If one variable exceeds or less than that of other variable, then obviously, the structure is inconsistent; then, you cannot proceed further. So, the first condition is that you must have two variables in the system and both the variables have same number of observations.

Then, we like to know what the degree of association between the two variables is. For that, you can apply covariance, you can apply correlation, but correlation is better than covariance, because it is unitless measurement while covariance is not at all a unitless measurement. So, obviously, correlation is better choice than the covariance although the equation of correlation is little bit complex. So, now, if your objective is to know the degree of association between the two variables, then correlation is the best technique for that. However, if you like to know what is the cause and effective relationship between the two variables, then of course you have to go for regression analysis.

Regression analysis basically gives you an indication whether it is X influence Y for Y influence X. So, in that, we have two standard equations. For Y on X, it is Y minus Y bar into byx upon X minus X bar. And similarly, for X on Y, X minus X bar into bxy upon Y minus Y bar. So, this is to know the regression coefficient, because it will give you indication how the path is all about, because the moment you will get regression coefficients byx and bxy, then it will give you the indication of what is the value of correlation coefficient, the square of correlation coefficient, that is, coefficient determination. And, that will give you the weightage of that particular relationships. If the value of that r square is very high and close to 1, then they are perfectly related to each other or their degree of association is very high. And, if it is very high, then obviously, the prediction and forecasting structure is very (( )). So, now, the r will give you the indication about the moment between these two variables in their association, also its causality.

Now, with this, we have to end this particular class. Next class, we will discuss in detail about the basis statistic before you enter into the econometric modeling. So, that is the case of probability and hypothesis testing. So, in the next lectures we will discuss the probability and hypothesis; then, we will proceed to the multivariate econometric modeling. With this, we will close this class.

Thank you very much. Have a nice day.