INDIAN INSTITUTE OF TECHNOLOGY ROORKEE

NPTEL

NPTEL ONLINE CERTIFICATION COURSE

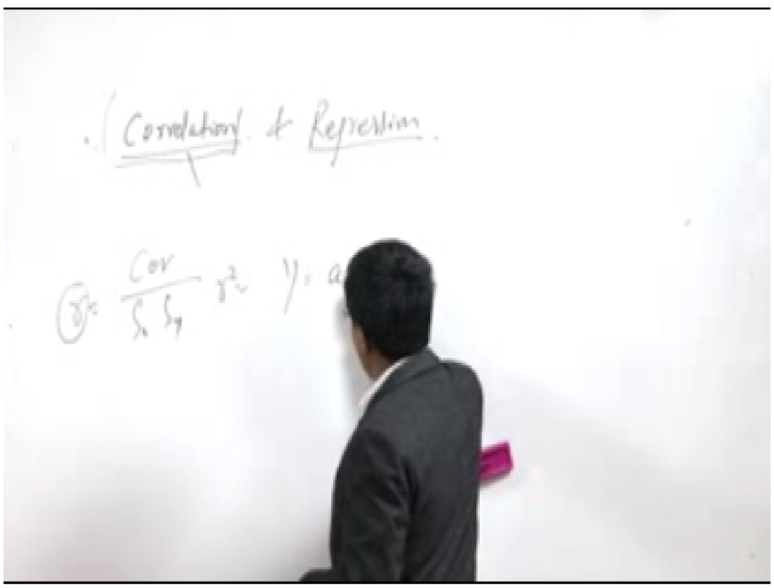
Marketing Research

Lec -24
Regression

Dr. Jogendra Kumar Nayak
Department of Management Students
Indian Institute of Technology Roorkee

Welcome friends to the section of correlation regression well we are continuing from the last section where we did with correlation right so in correlation we understood about what is, correlation basically it is a standardized covariance model basically right.

(Refer Slide Time: 00:43)



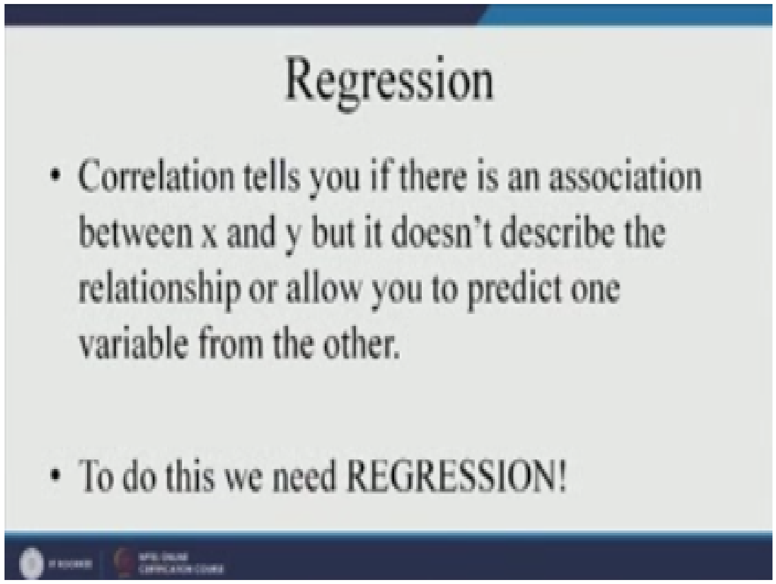
So where we are saying that when the covariance which is basically the we say the covariance right correlation we said is a covariance right r= r which is the correlation covariance divided by standard deviation of x by standard deviation of y so this is we said right and it is very helpful in several ways to understand the relationship between x and y right x and y being the two variables independent and depended okay.

But we are also talking about something called r^2 right what is this r^2 is nothing but the square of the correlation value right now the square of the correlation r^2 means that when suppose our r= let us say 0.9 right in our last case it was something around 0.93 or something so that step r is one that means the relationship between when two variables is 90% right, our explanation power that means r^2 is = 81% or 0.81this says that the explanation power of the independent variable in the dependent variable is 0.81 in this case right.

So that is a very important analyzes is very useful as an analyses and I would tell also that one thing that through a correlation also helps in checking hypothesis right correlation also helps you in testing a hypotheses and through the same method we will see, so this r^2 that you just saw is called as the coefficient of determination please remember it, today we will go into in this section we will discuss about something which is highly connected with the correlations right.

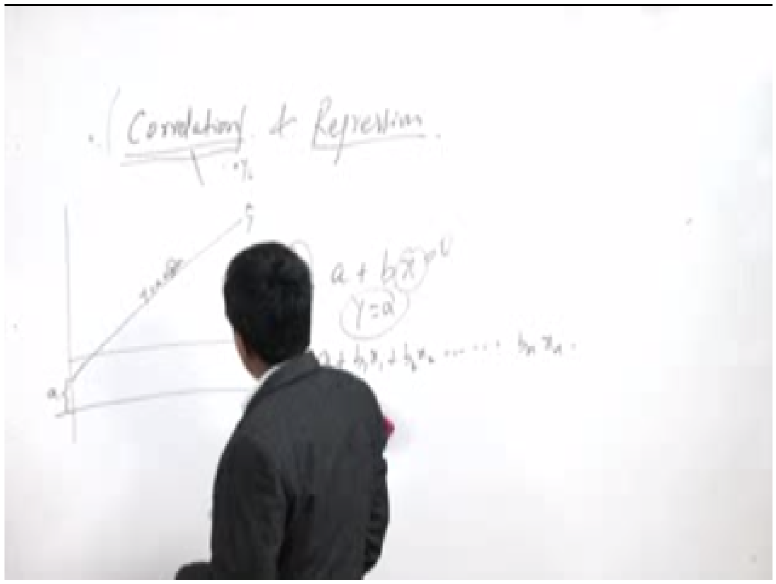
So that is why in most of the books or any place you see it is correlation and regression in correlation we were trying to measure the relationship between x and y right what is the relationship between x and y but what if we want to know key how much is change coming in y because of the change in x suppose you want to create a relationship you know relationship so to do that we would use something called a regressing equation right.

(Refer Slide Time: 03:08)



So regression basically if you can see correlations tells you that there is an association between a and y right but it does not describe the relationship okay it does not describe the relationship but that is what we do in the regression so regression equation is equal to is a

(Refer Slide Time: 03:19)



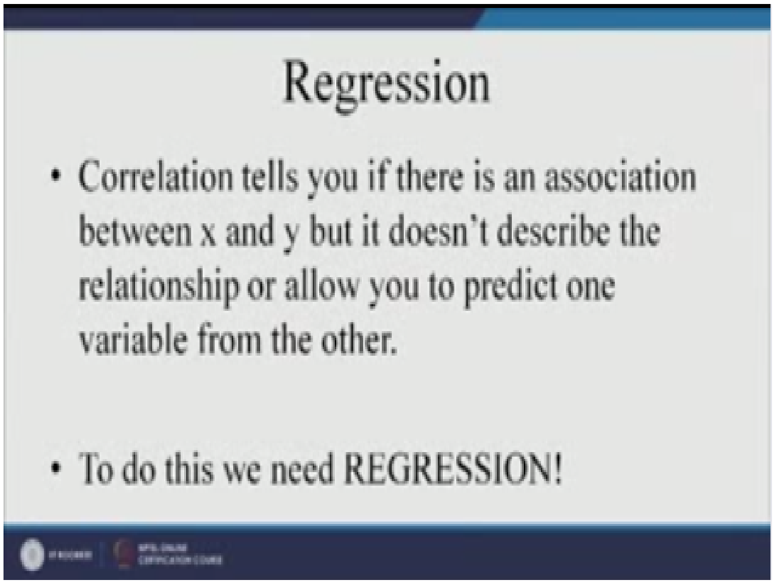
Y = y is the dependent is = a + bx right these are very simple regression suppose you have more than one variable the y + b_1x_1 + b_2x_2 + goes on till b_nx_n right so the change in y due to the change in the variables independent variables x, x1, x2, x3, x4, xn right is what is of for most importance to any researcher right so y helps in understanding it is helping and predicting a lot of things in life okay.

As I was speaking about stoke let us say the share value we will share value change due to an announcement we will share value change due to an announcement of new product let us a change I the companies as chairmanship or let us say another variable is the fight between the company employs right will those effete the final outcome of the share value now e want we do not know we can check that okay and the coefficients that we get the coefficients are basically the you can understand as the slopes basically the coefficients the β coefficient basically what comes is a that tell us okay how much is each independent variable having the effect on the dependent variable okay.

Now to before we go that let us go to the last diagram that I have drawn right which is very important now what was that we where explaining about something called the variation right now we saw that there are two things one this is the mean the why the mean of the value right, then we have something called the this is the you know something the predicted value or predicted this is the predicted y^ we estimated or predicted right so this is the y = a + bx now what is this b the b is nothing but this slope this slope right, this slope is the b what is this a then a is the intercept, now intercept means intercept is a constant a right so this constant tells us k.

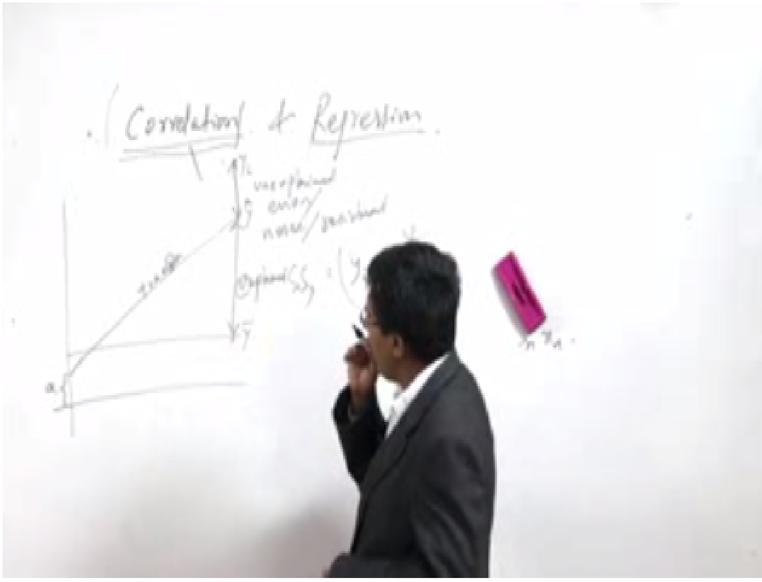
That means what if you understand this way suppose a x = 0 then y = a right so this constant is that value when the independent variable does not exist okay, so to do and to understand okay.

(Refer Slide Time: 06:02)



Yeah from here then we said key if there is let say y let say this is I now we said there are two three things that we need to take right.

(Refer Slide Time: 06:14)

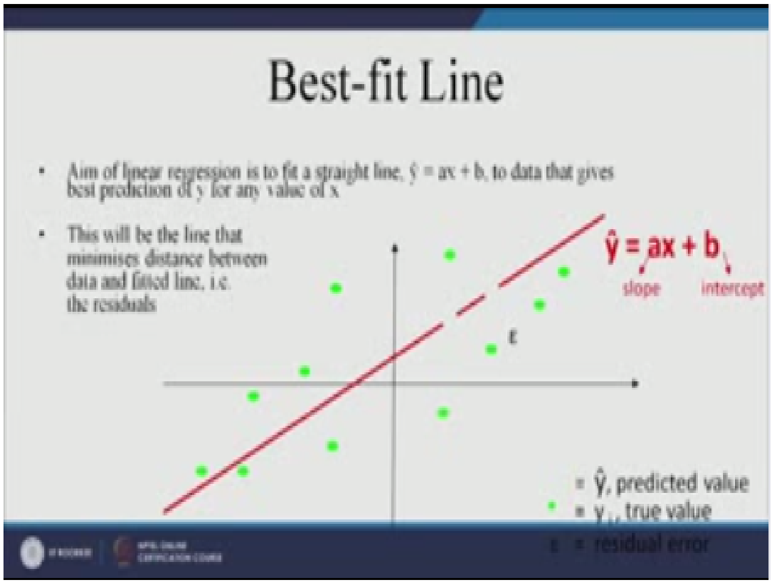


What are these the total variation the total variation in this case is equal to ssy we say is equal to yi – y bar right square okay this part I had explained is called the unexplained or the unexplained error or sometimes it is called as noise whatever way you can explain right, so this is something this part is our explained right or we say so this is a explained part right, so now the total variation so new when we are trying to say how the when we are trying to measure any you know regression equation.

We are interested were interested to understand okay how much of explained variance is there how much of unexplained variance is there right, so what we are doing is basically we are trying to find out we are actually finding out the explained variance the explained variance upon the total variance, we require we are interested generally in the explained variance and the total variance so sometimes you can also make it like this okay that means what total – unexplained or error right.

Divided by total okay so let us see this aim of this is very important the aim of the regression line why is this why the regression line is also called as the best fit line right, if you see this.

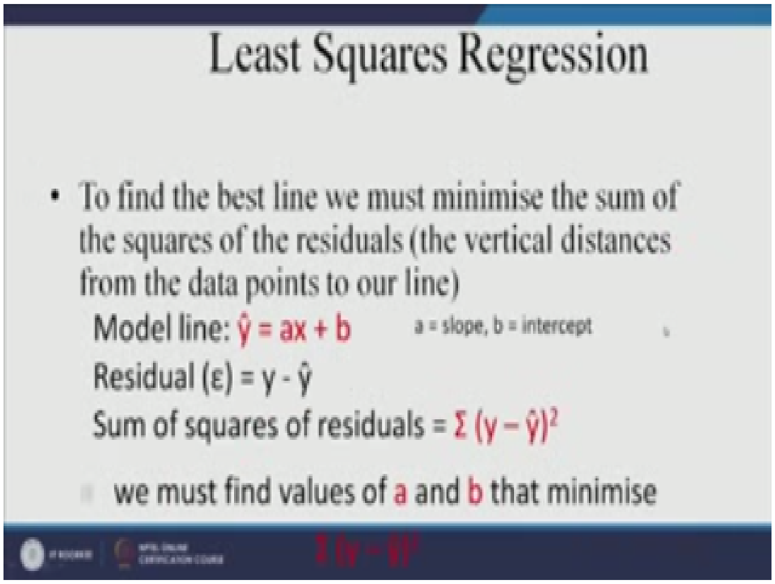
(Refer Slide Time: 08:08)



Now this is called the best fit line now best fit line because if you draw you know any line there could be infinite lines right you could be drawing infinite lines but the point is why only this line is called the regression line because let say there could b infinite you know once so but y only this bold one right why this bold 1 because this line has the minimum deviation that means the if you take the dots are the values the observed values and find out a standard the variance then you will see that this line has it is the minimum variance right.

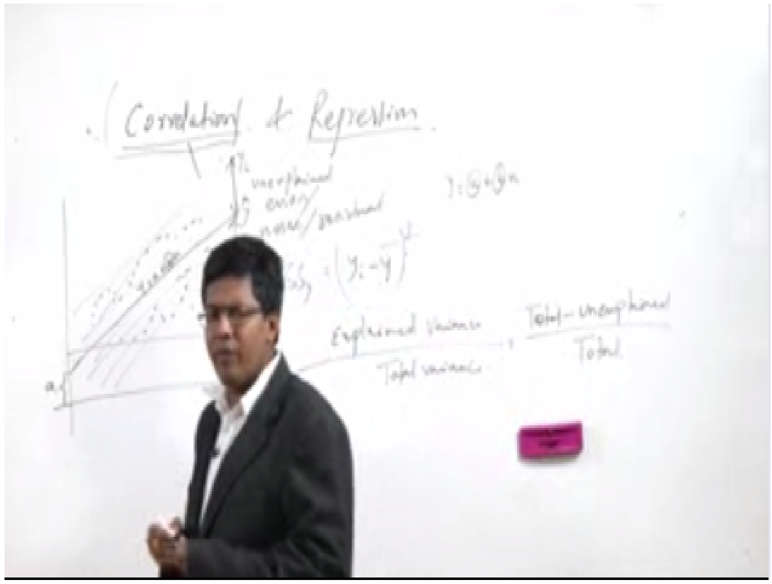
The variation of the value the summation the summated values of all the data's the minimum variation comes in when you use this line right, so that is why it is also called as the best fit line right so y estimated if you can see is the predicted value yi is a true value which is the observed value I has said and e if it is visible or not e is my residual error right okay.

(Refer Slide Time: 09:27)



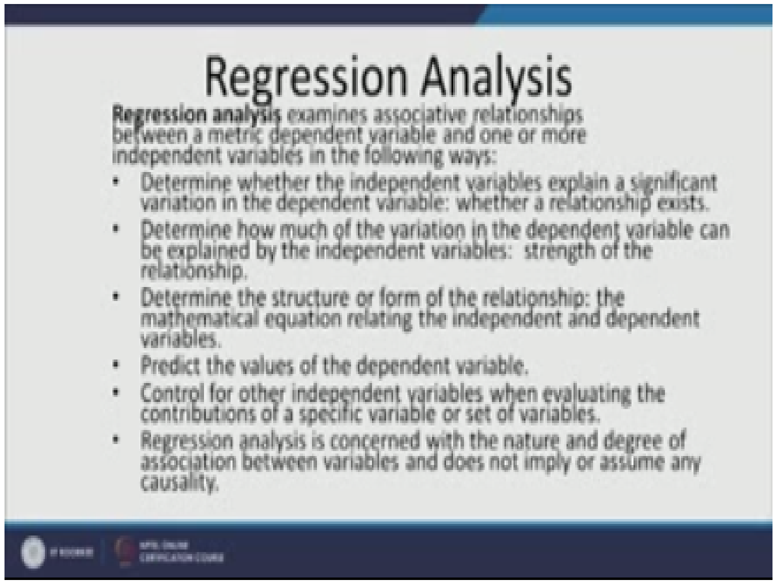
So the question is if I am talking about the regression equation.

(Refer Slide Time: 09:37)



So y = a + bx right so that means what I want two things to know so for a given value of x how much will be the y to do that we need two different values that is the a and b right so how much is our slope and what is the constant if we know that then for any change in x w can know a y right so the prediction becomes much easier, okay.

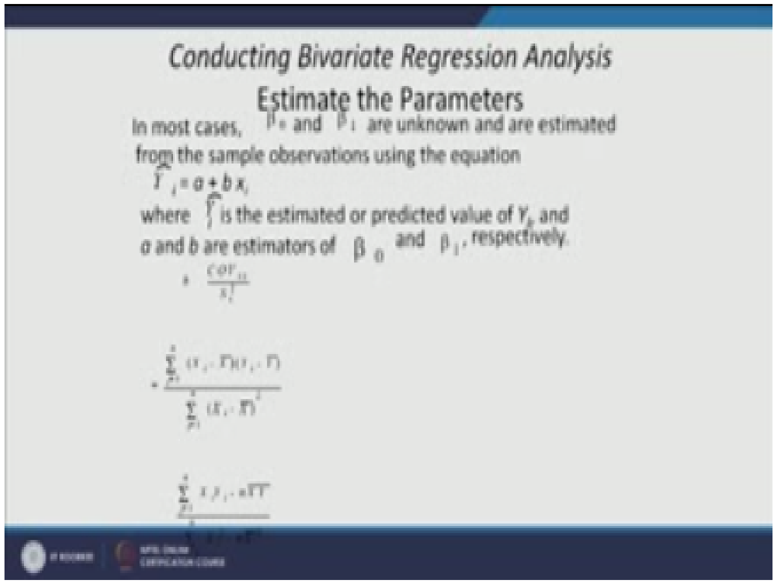
(Refer Slide Time: 10:09)



So let us see this, let us go to the regression analysis so there are few things that are very important it examines associated relationship between metric dependent variables and also one or more independent variables okay, right. So it whether a relationship exits or not first of all. Next it says, how much of the variance variation in the dependent variable is explained by the independent variables that means the strength of the relationship so if you have several let us say the coefficient which I said, the coefficient of the independent variable the higher the coefficient, higher is the explanation power right.

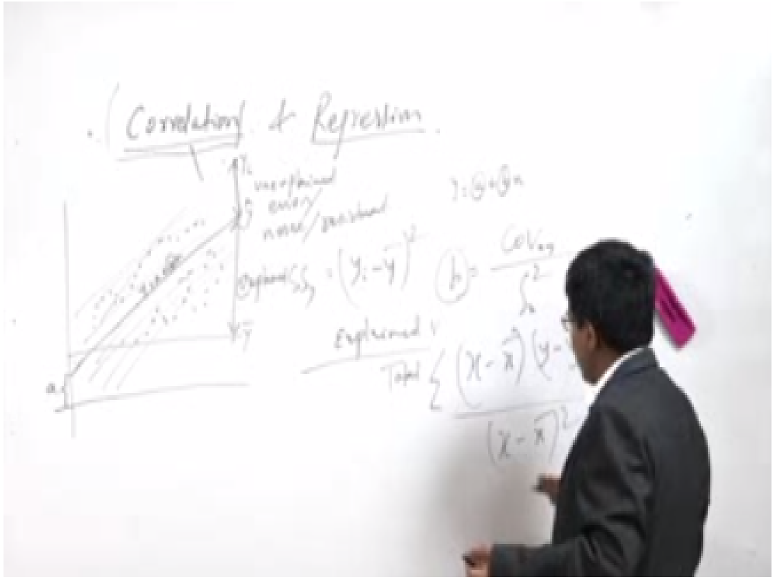
So if the estimate if you look at the estimates basically say beta coefficient the estimates then we say the higher the stronger the beta coefficient the stronger is the explanation power, right. So these are the few things, these are very important okay.

(Refer Slide Time: 11:11)



So let us go to the how do cover that, how to find out right that b, so as I was saying so b or the slope b.

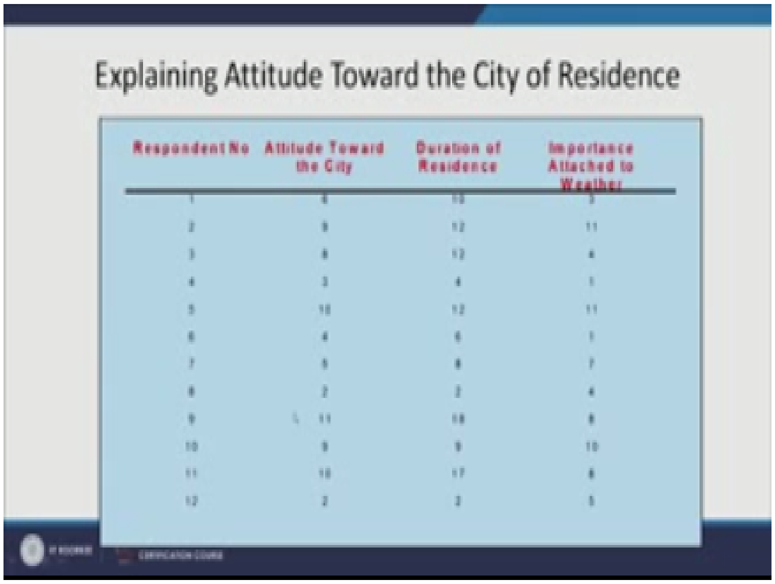
(Refer Slide Time: 11:21)



b= the nothing but the covariance that means the covariance between x and y divided by the x^2 right, it is what it is saying b=covariance of covariance divided by xy divided by Sx^2, Sx^2 is nothing but the variance in the x value right, the variance right, the it is the square the standard deviation square of x basically which is the independent variable, okay. Now if we take this we will get the slope the b value right.

Now let us break it up, now as you know how will it look like let us say, so $x-\bar{x}$ I think you must be remembering by into $y-\bar{y}\sum$ right, divided by so what was $(x-\bar{x})^2$ correct, so this is what we want so if we have this if we get this then automatically we are finding out we can find out the value of b right, so using this formula you can find out the value of b, right.

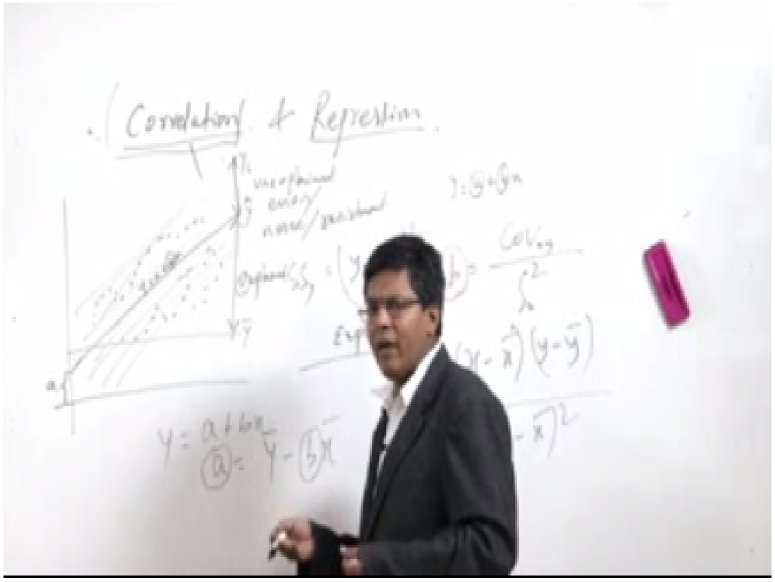
(Refer Slide Time: 12:40)



Respondent No	Attitude Toward the City	Duration of Residence	Importance Attached to Weather
1	6	10	3
2	9	12	11
3	8	12	4
4	3	4	1
5	10	12	11
6	4	6	1
7	5	8	7
8	2	2	4
9	11	18	8
10	9	9	10
11	10	17	8
12	2	2	5

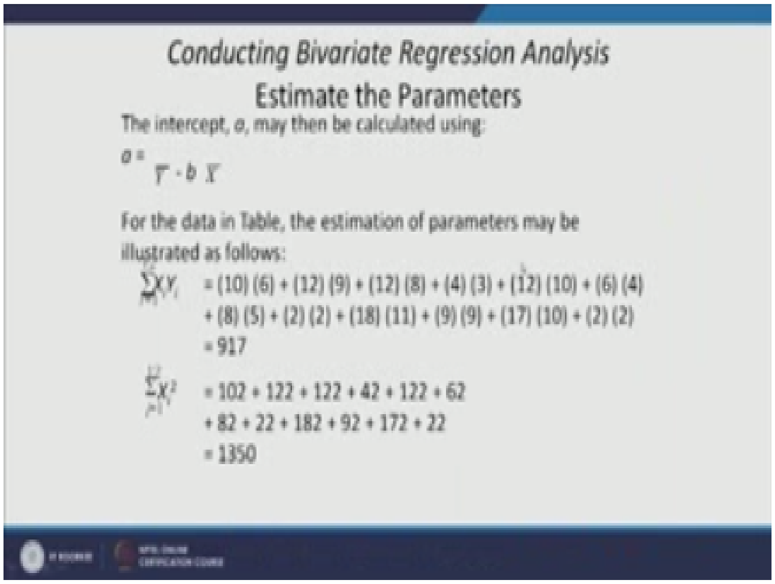
So let us take this there is a case, but before that once you have found the value of b you can also find the value of a, now how much is a then.

(Refer Slide Time: 12:50)



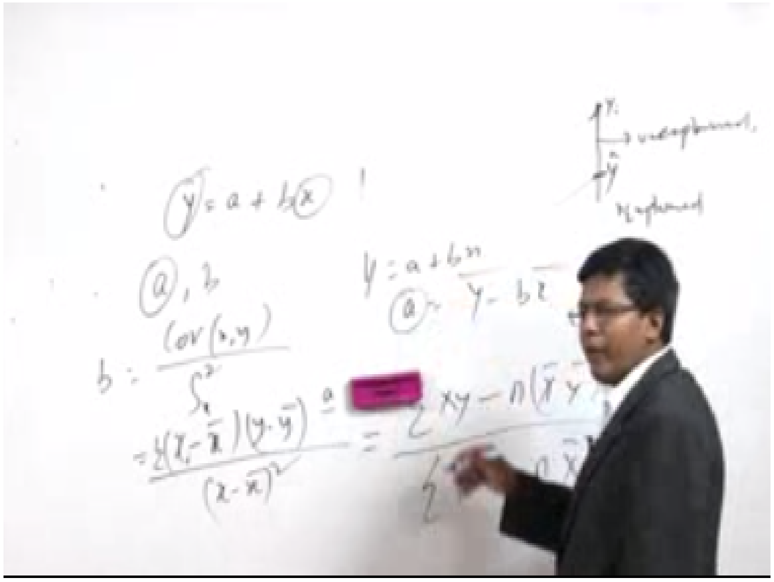
Now we said y=a+bx, so a is equal to is nothing but y-bx so to do that when you have already got the value of y what should you do, you just have to take the y mean and the x mean and use it for calculating a, and once you get the value of a and then you have got already the value of b then what we can do is, we can just you know measure any value of y after that, right. So let us take a case the same example we are holding right, so we said that the respondents is the respondents the attitude towards the city is our dependent variable, the duration of stay in a particular city is my independent variable x1, and this is my importance to the, whether which is x2. Now I maybe using 1 or 2 let me see.

(Refer Slide Time: 13:45)



So the intercept for example in this, now what did we say now if you go back this is also written as so if you use this formula right, you can use through you can calculate through this method I do not like to remember things but if you want you remember you can do remember so there is another way of understanding remembering this formula so as I have saying so to have predictability.

(Refer Slide Time: 14:18)



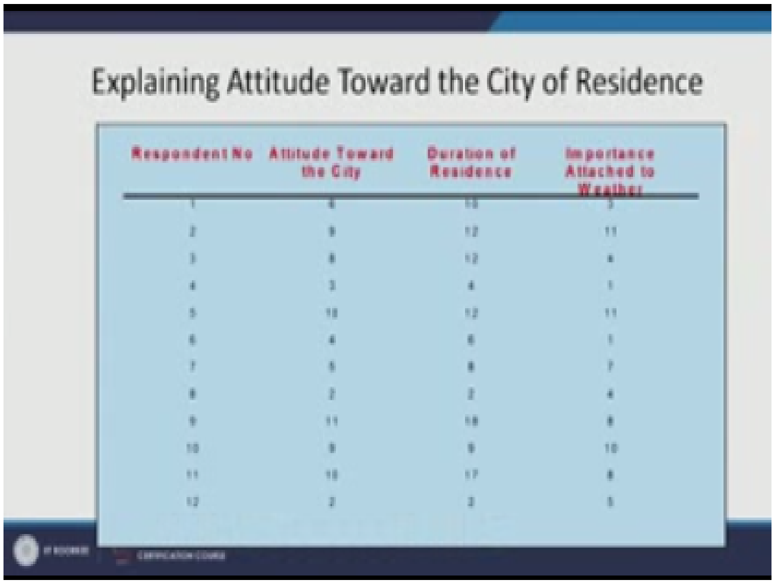
Y = a + bx right where x being the independent and this is the dependent so we need two thing the A and B okay. So A is as I explain if you remember right so suppose this is my y mean this is the estimated regression right estimate the y value and this is the y = a + b x the regression line right we are also discussed why it is call the best fit line right so and this is let us say my y the true value or the observe value.

Now we have two things this is the total right this my explained part and this is up to this is my unexplained part which I had already written right so to this is the interest we are talking about and the slop right so the researcher needs to understand to find out to predict a the value of y he needs to find out also the value of a and the value of b okay, now what are they say the b value the slop the slop is nothing but the co variants x y of x and y / the only the standard deviation or the variants this of x the variants of x right.

Now if you look at this now what does it becoming now let us see so if I write this can I write this like this now co variants of x and y is equal to x – x bar right into y – y bar right divided by the x – x bar square right so summation of this will automatically tell me what is my slop right. So if I have got my slop right and then calculate my a right so let us see from here even you can write this formula as something like this because in books you will find then many a times like this so this is something like summation of x y – n x bar y bar right divided by summation x^{2} – n x bar square right.

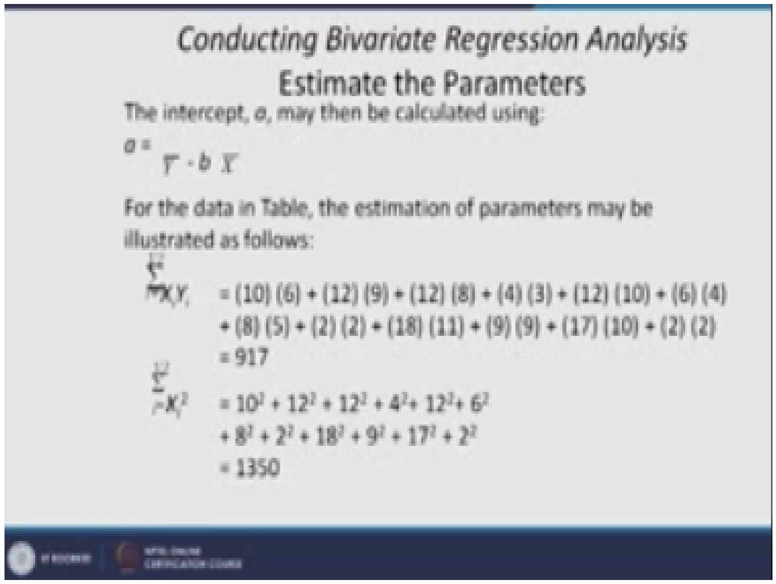
So this is the same formula between results two right so let us look at this problem which there we have already done right.

(Refer Slide Time: 17:00)



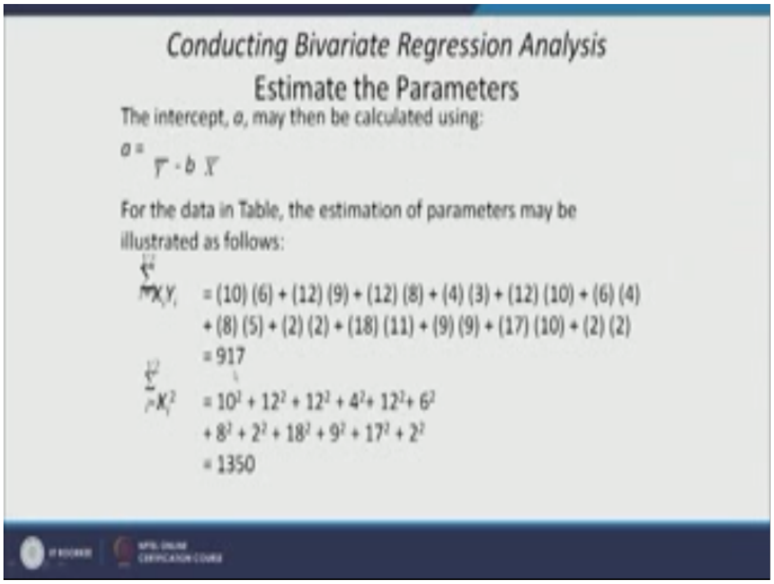
So let us look at this so what does it say now attitude towards the city is are dependent variable and that the duration of residence is my independent variable and I want to predict what is the attitude of the residence for a city by taking in to note or account the amount of time is spent in the city is lived that in the city right. So what is done is a is this is little upside because of my adjustments not been able to do well.

(Refer Slide Time: 17:34)



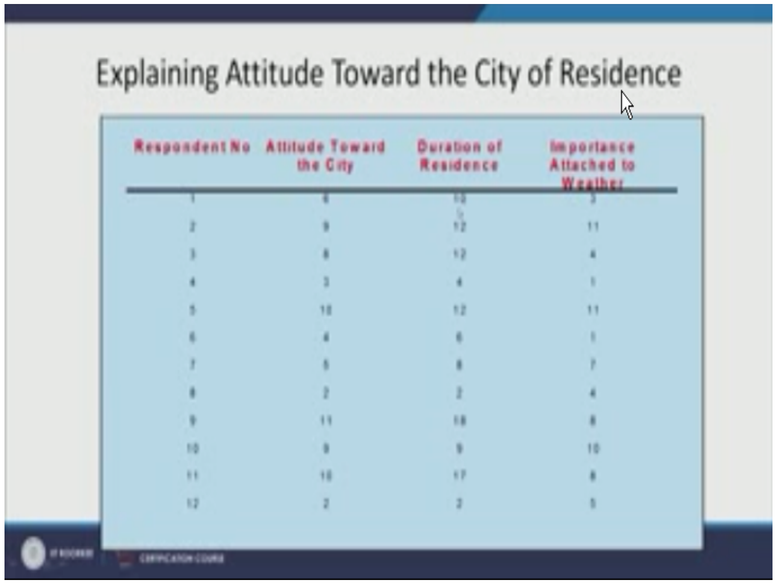
This is a = y – bx now what did I saying once you have calculated this b right then we also need to calculate a now how do we calculate the a let us say y = a + bx so a = y – bx right but what value of y and x we are taking we are taking the mean the average of y and x right, so if you take if you do this then you can calculate the value of a now using the value of A and B now you can predict any value of y for any given value of x okay so let us look at this case right so what is doing it is what did I said it is first calculate the x y as given in our formula.

(Refer Slide Time: 18:22)



And so earlier I to make a change because there was a slight change so this was looking like a 10*2 which because earlier it was the mistake because of the you know change in these slide so this is basically a square so x^2 so earlier it is looking like 102, so if you must by chance if you are seen you do not get confused so 10^2.

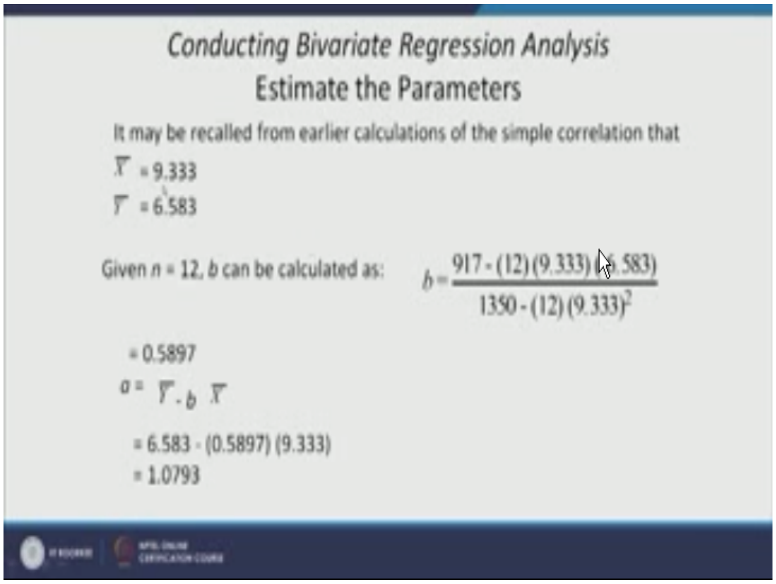
(Refer Slide Time: 18:45)



Respondent No	Attitude Toward the City	Duration of Residence	Importance Attached to Weather
1	6	10	3
2	9	12	11
3	8	12	4
4	3	4	1
5	10	12	11
6	4	6	1
7	5	8	7
8	2	2	4
9	11	18	8
10	9	9	10
11	10	17	8
12	2	2	5

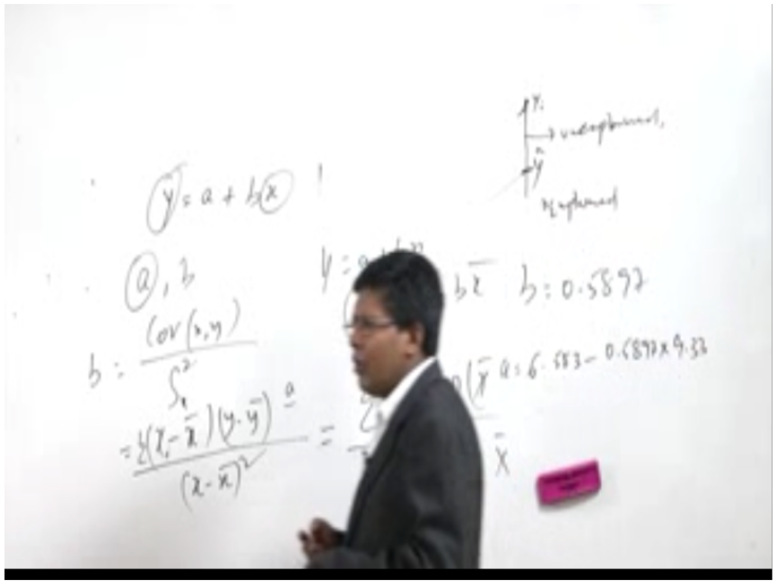
Now look at this value $10^2+12^2+12^2$ up to 2^2 now this totally 1350 okay so we have got all our values now x bar y bar x bar y bar is how much now x bar if you remember we are calculated earlier so x bar was or we can calculate from here 9. Something 9.5 was something it was I do not know let me see yeah so x bar is 9.333 and y bar is 6.583.

(Refer Slide Time: 19:10)



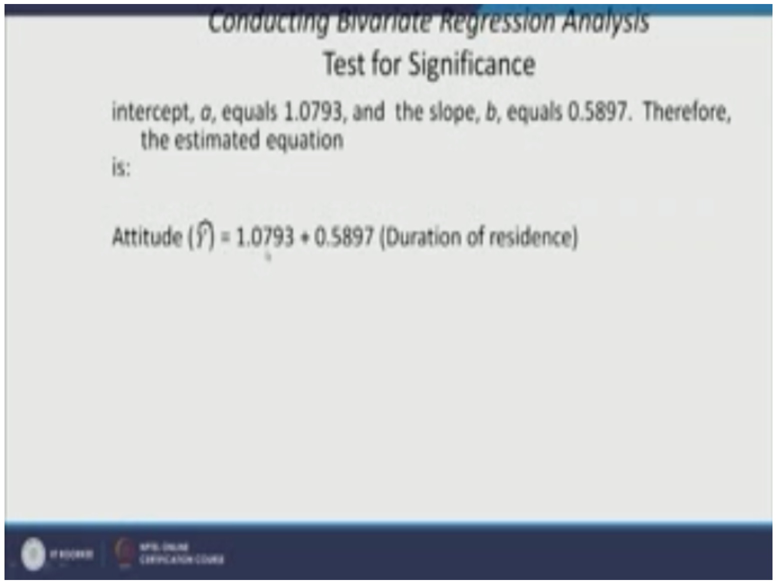
So n=12 so if everything is given to us now you can surely calculate the even calculate from here only let us not go back so b=917 which is our this x y value-n n is 12 so x bar* y bar/ x^2 which is 1350 right -12*x bar right so this value has come to 0.5897 so by getting this value of 0.5897 that means b is equal to 0.5897 we obviously as I said we can put the value of y, y bar was how much y bar was 6.5 so y bar is 6.583.

(Refer Slide Time: 20:07)



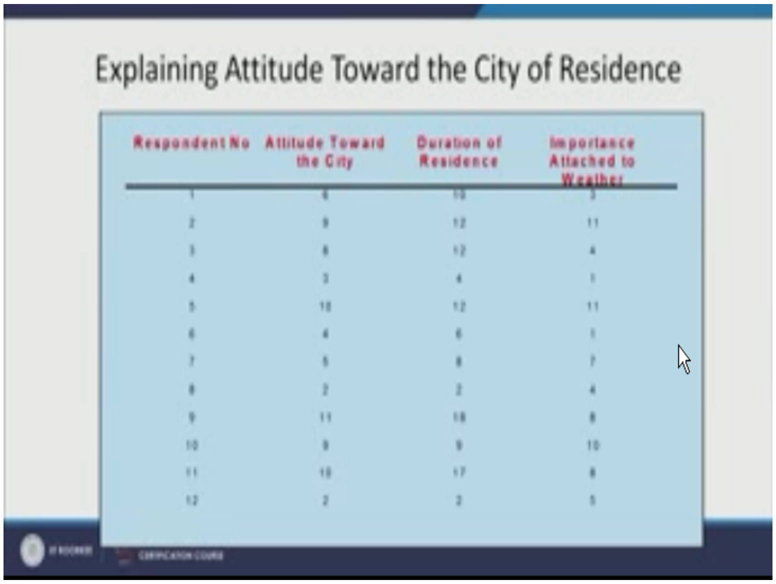
So 6.583- a=y-bx so b was 0.5897*x bar is 9.333 something right so by doing this we have calculated the value of a=1.0793 so that means if x=0 still the constant value of the constant is y=1.0973 0793 that mains okay.

(Refer Slide Time: 20:41)



So by doing this now you can predict suppose attitude y estimated y=a 1.0793+0.5897 given into multiply by the number of years the person has take so in this case if the person has state for let say so 1.793+b is 0.5897 suppose the person is state for 30 years in the city so we can say y=that means how much so 1.0793+35 almost this I almost 18 almost 18 not 18 you can deduct 1 may be 17 something around 18.

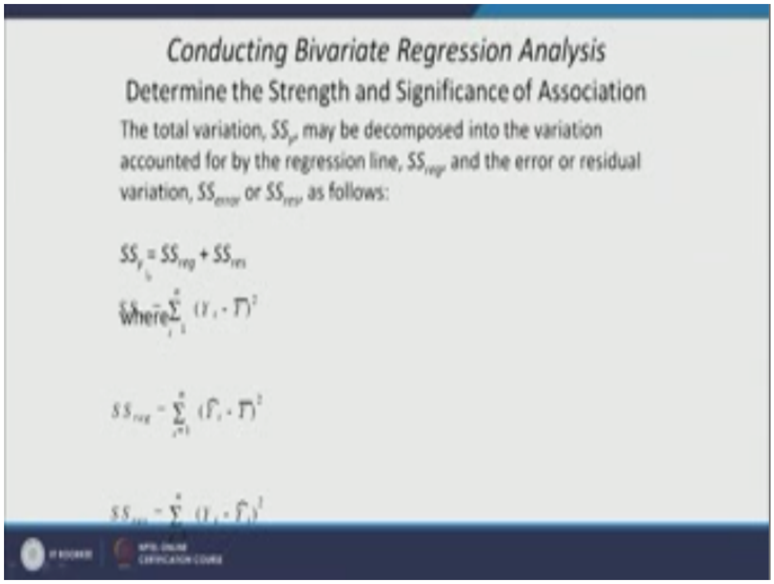
(Refer Slide Time: 21:32)



Respondent No	Attitude Toward the City	Duration of Residence	Importance Attached to Weather
1	6	10	3
2	9	12	11
3	8	12	4
4	3	4	1
5	10	12	11
6	4	6	1
7	5	8	7
8	2	2	4
9	11	18	8
10	9	9	10
11	10	17	8
12	2	2	5

So the attitude will be round 18 something 17 18 now let us go back to the table and see so if you look t the attitude the attitude is going to improve the very high if I am taking the duration of residence this is very strong relationship that we have find this is good strong relationship that we have finding right so this is let us take one more problem right.

(Refer Slide Time: 21:49)

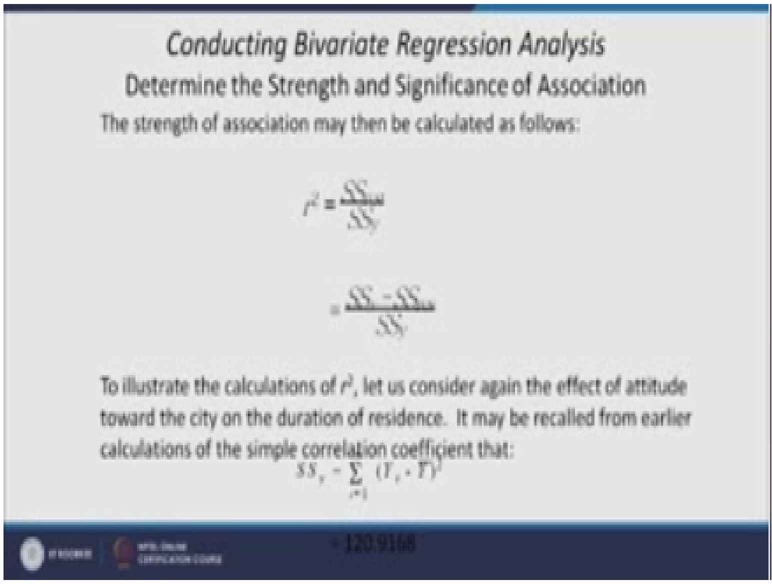


So as I am explaining this part kindly check this so as I am saying if you remember this total y total regression total the total variation right the total variation now this was the total not this one this is the total so this was my total variation y bar so this is my let say total variation so ssy is my total this is the total right and this is my, you can say this is my ssr the regression so why doing any regression the researcher is interested in this three variances right.

So ss regression is nothing but my estimated y estimated y was this one –y bar right and similarly you can find the regression also residual what do you mean by residual as I have already explained that means this is something which you are not able to explain the researcher is unable to explain now why he is unable to explain that is again the apart of the study that one needs to understand may be he has missed something very important.

Or the way he has collected the data has been something some anomaly or could have been something which we do not know at the moment and this explanation and whatever the test you do=, there will be slight errors will always be there, but if there are unexpected variance delivers the explained variance, the whole purpose of the study will be last okay.

(Refer Slide Time: 23:26)

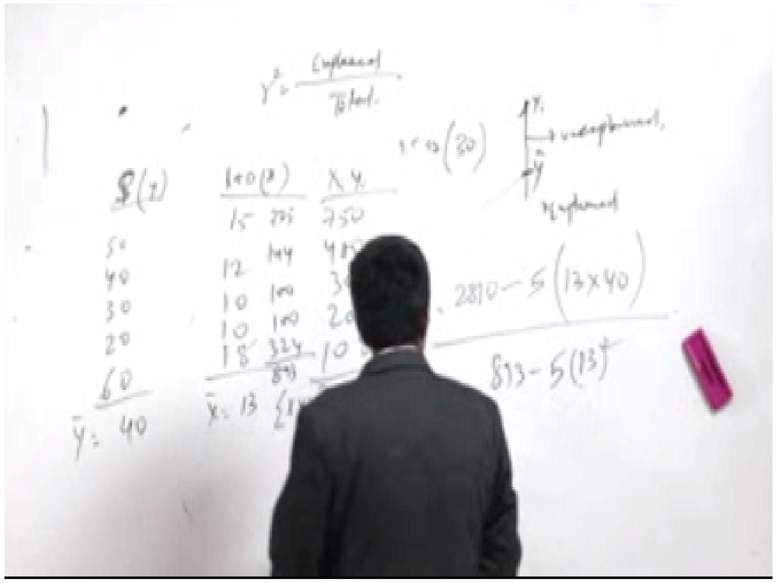


So if you see this r^2 which I have already explained in the coefficient of the determination, r^2 is nothing but it says SSr/ SSy right r^2 is nothing but this given by the explained variance/ by the total variance, I hope you remember what I have said r^2 = this coefficient of the determination the explained/ total. Now what part of the total am I able to explain; now this is what the coefficient of the determination says.

So if I have the r value also, the correlation value let say 9 I said here r^2 becomes 81% that 81% you have been to explain that is 0.81 which will come here and the unexplained part which remains is the 0.19 okay 19% okay. Now let us take one more problem, which I have bought. Let us solve this problem. So what I have done is we will take a case where the company, where it wants to calculate and find out this is the relationship between two things.

One is the amount of the expenditure it makes in R and D against the total profit, so what is my profit.

(Refer Slide Time: 24:58)



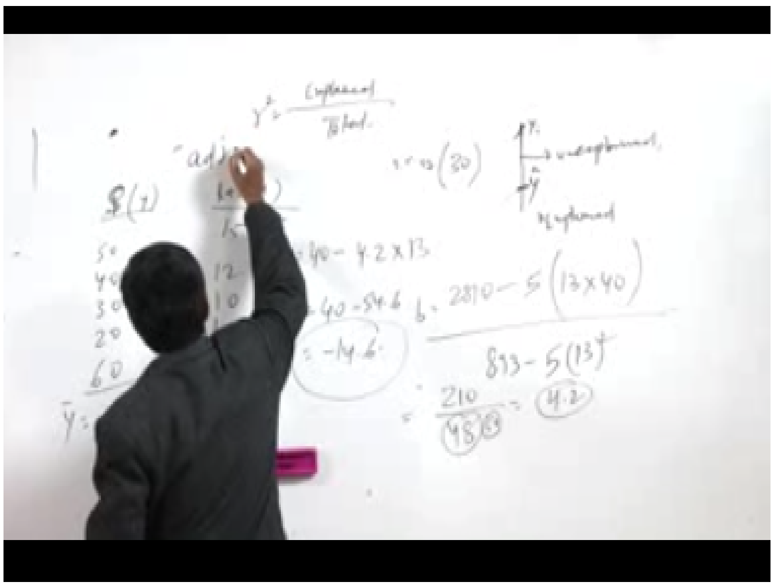
Dependent and r and d expenses become my independent okay, so you want to calculate, can I find out and predict that what would be my future profit, if I change R and D expenditures okay. so do that we have brought in some of a you can say that profit sales, 50 we can say 50 million, 40, 30 and 60 they are randomly brought in, so these are expenditures in 15 million, 12, 10, 10 and last is 18. So I suggest you side wise with me along with me, so my x this is my y. so y bar = how much the total is 200.

So this make 1, 2, 3, 4, 5 so 40, so y bar = 40, x bar = how much it has come around 65, so x bar = 13 okay. What do I need? I need the xy basically so X_i Y_i if I do know how much will come okay, so this is 300, this is 200, this is 1080 right this totals comes to 2810 okay ∑ of xy = 2810 right. What else we require? Now let us try to do it, if I am calculating b I will have 2810 – what was my n = 5, so 5 x what was the next x bar y bar, so x bar is how much 13 x y bar is how much 40/ 893. This is x^2 x bar is 225, 144, 100, 100, 324, so this will become somewhere around 893.

So 893 – 5 x x bar n x x^2 right, x^2 is how much? X^2 = 13. So if you do this you are getting almost the b value is I do not know you can do it with me something around I have already calculated so if 210/48 I do not know my calculation I have done it just it might be wrong also so 2810 the 2l certain was 210/48 okay so something around this so 48 so even if you if I to make it easier let me make it somewhere close to 50 so 4.2 almost 4.2 so my b is almost 4.2 and if my b is 4.2 what is my A now my A I am rubbering this off I have no place so if my b is 4.2 by a is y-4.2.13.

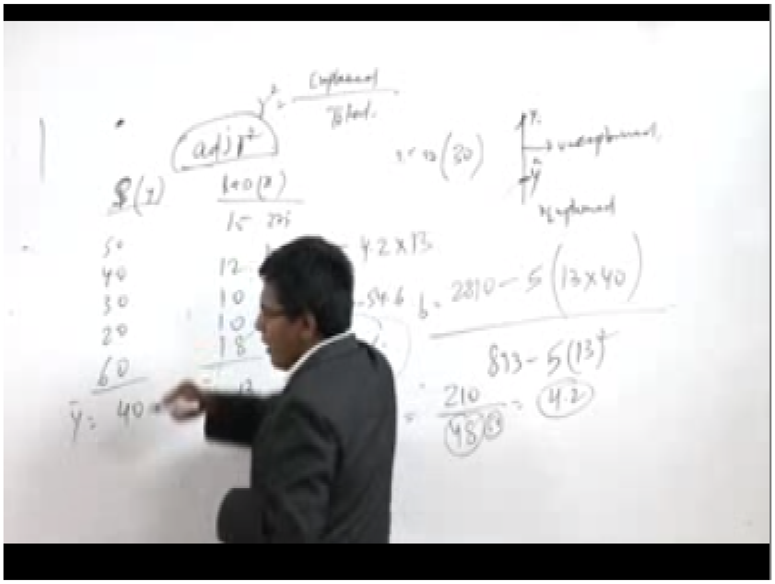
So this is my A so this is equal to this implies a=40-54.6 so something around -14.6 so having this values of a & b we can easily put it into the regression equation and can say if the company changes the value of expenditure in whether he is going to increase his sales or he is going to decrease his sales we do not know we will have to put it because every time we cannot say that the relationship would be directly proportional or positive right so it will go on increasing not necessarily so that is where something which is very important but here before I wind up this session I want to tell you something the r^2 that you saw this r^2 is very important but in research paper sometimes you will see that there is something called an adjusted r^2 also

(Refer Slide Time: 30:04)



Just I want to brief this also to all the listeners you must have heard of something called as adjusted r^2 now what is this adjusted r^2 now ^{r2} as you can see this r^2 will go on increasing when you add up any independent variable right so if you have 3 independent variable and your are adding 4 independent variable the r^2 will increase if you add a 5^{th} one as per will see still increase but some variables actually do not contribute right and the so what happens is just by adding a variable although the r^2 value increases but the actually contribution does not increase so what it does it we have a something called an adjusted $r^{2.}$

(Refer Slide Time: 30:51)



Which is very important what it does it takes in account only till that which if you add there is a growth in explanation power so after certain stage when you add a new variable and the growth in the explanation power does not change much or it even falls down then we stop there so this is the value that you will you will see in many of the research papers which they sight it which is called as the adjusted r^2 right so this is something that I wanted to explain in the regression as I have said it is predicted analysis which is used in almost all fields of science.

And management and everywhere right so where the reached wants to predict by having through an independent variable right so there is a independent variable we are saying it is a single you know regression simple regression suppose you have two or three or n variables we will say it is multiple regression or you know so in those cases the researcher have to understand what is the strength of the coefficient right the strength of the coefficients or the estimates we say basically we will explain how much of how important is that particular independent variable to the final dependent variable right so this is all for the regression session thank you so much.

For Further Details Contact

Coordinator. Educational Technology Cell
Indian Institute of Technology Roorkee
Roorkee 247 667
E-Mail Etcellitrke@gmail.com etcell@itr.ernet.in
Website: www.itr.ac.in/centers/ETC. www.nptel.ac.in

Production Team
Sarath Koovery
Mohan Raj. S
Jithin. K
Pankaj saini
Graphics
Binoy. V.P
Camera
Arun. S

Online Editing
Arun.S
Video Editing
Arun.S

NPTEL Coordinator
Prof B.K Gandhi

An Educational Technology Cell
IIT Roorkee Production