Data Analysis and Decision Making - I
Prof. Raghu Nandan Sengupta
Department of Industrial & Management Engineering
Indian Institute of Technology, Kanpur

Lecture – 22
Simple Linear Regression

A warm welcome to my dear friends, a very good morning, good afternoon, good evening to all of you and welcome to this Data Analysis and Decision Making- I under the NPTEL MOOC series. And this is with the total course duration as you know which I repeat beginning of each class, this is for 30 hours course, 12 weeks and each week we have 5 lectures and each lectures is of half an hour. And we are on the 22nd lecture and I am Raghunandan Sengupta from IME department IIT, Kanpur.

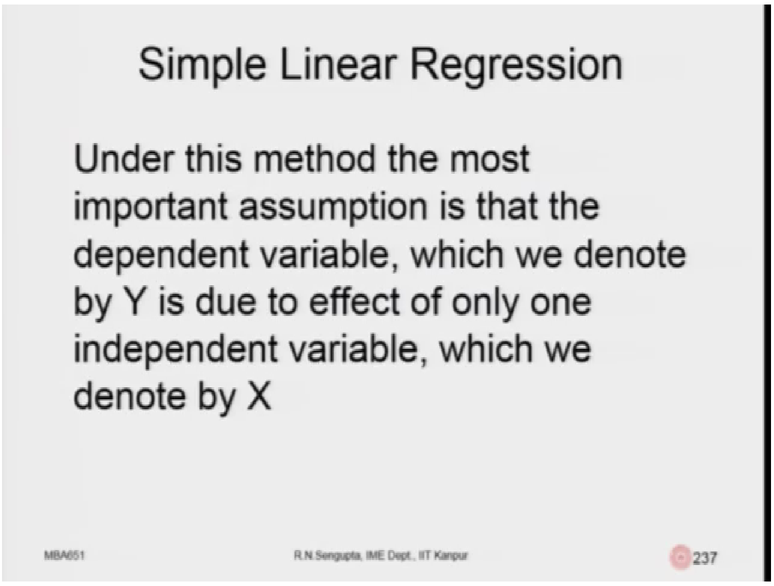
So, if you remember the last slide which were discussing before we left I will repeat that again when I am going to come, we have just started the concepts of multiple linear regression. And later on we will discuss something to do with the forecasting methods very briefly and then go into other models and multivariate statistics methods. So, in multiple linear regression, what we want to do is that with repetition I am saying because I was already said that, but I did mention the last class 21st lecture the last slide I would be coming back to this discussions time and again.

So, in multiple linear integration what the idea is basically you want to regress, that means, go back use the information about the information which is given by the independent variables consider there are k number of independent variables in many of books it is given as either small k or small p. And using this k or small p number of independent variables which are denoted from by x 1 to x k or x 1 to x p, you want to basically predict forecast the yth variable which is the dependent one. And as you are doing it, there would be many different type of assumptions which we need to understand such that it will be able to give you the maximum set of information for the prediction of y using x’s.

Now, multiple regression or generally if technically we should basically start with the simple linear regression. In simple linear regression, there is only one x and given one x is your basically y to predict. So, with all the information which you have about x is subsumed as such that x is able to predict y to the maximum possible way value.

Now, it does not mean that when you are using one x or multiple x's, you are able to predict y always with 100 percent accuracy. There would be an error, so that error of the white noise would have some properties. And I did mention some of the simple properties when you are basically discussing what are the assumptions based on which multiple linear regression works. I will come to that again, but let us first start with simple linear regression and come back to multiple linear regression in all its details.

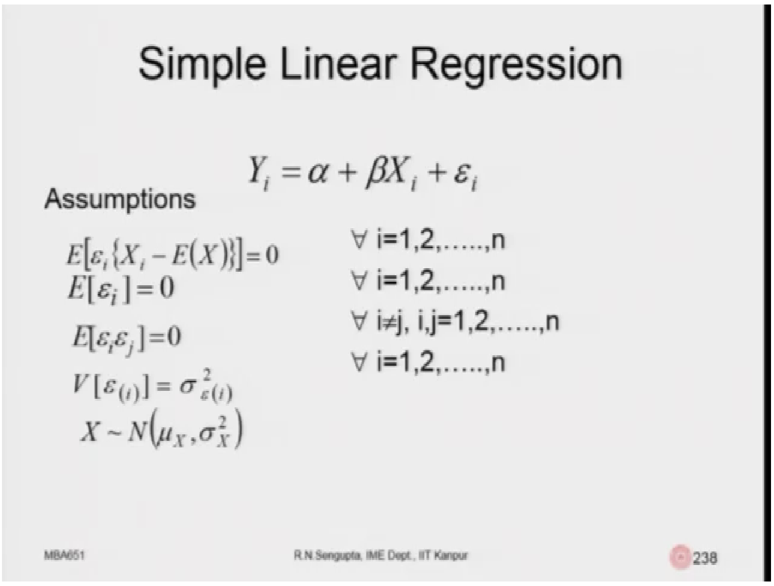
(Refer Slide Time: 03:30)



So, under this method the most important assumption is that the dependent variables which we denote by in I am sorry then so we are we should have been basically X as bold in order to denote is a vector, but or matrix what for the time being will consider it is a simple linear regression. So, the important assumption is that the dependent variable which we denote by Y capital Y is due the affect of only one independent variable, in that case obviously, X is a would be denoted not with the bold and the nomenclature would change according I will come to that later on.

So, you will basically have 1 X, 1 the X would basically have the values and we will use those two basically predict the future values of Y, and obviously there would be error when we will try to tackle how the errors are handled and what are the properties. So, generally is the for the simple linear regression again what we were discussing over the multi linear regression I am going to come to that later on please bare with me.

(Refer Slide Time: 04:34)



For the simple linear regression Y i is given by alpha; and this known as suffix for alpha plus beta is no suffix for beta X i plus epsilon i. So, this is basically means the reading number. So, if you have 10 readings, I would be 1, 2, 3, 4, 5, 6, 7, 8, 9, 10. Now, here are what the actual variables are, Y with the suffix are the set of dependent variables starting from 1 to n considering, they are n number of readings. Considered if it is time period, if it will be for January, February, March, April, May, June and so on and so forth; if it is basically for week it would be first week second third week and all the values.

Now, alpha is basically the constant value technically constant value would basically means that even if the X's values are 0, Y does have a value which is alpha. And if you compare this equation with m x plus c, you will find out some similarity there. Beta is basically regression coefficients and we will see later on it gives you the rate of change of the value of Y with one unit change in X. So, consider that is the regression coefficients. And later on when there are the multiple linear regression we will consider that they are the regression coefficients that the partial regression coefficients. Why, I am using the word partial you will also understand that.

And if you if you all of us know that the concept of partial differentiation what they actually do is that we keep all other variables fixed, and basically find out the rate of change of the function with respect only one variables. So, partial differentiation is used for that case, and we will use the partial regression coefficients for in that context only. I will come to that when we do the multiple linear regression.

So, you have betas than you have basically multiplied by X i's, X i's X as I mentioned just before this slide these are the independent variables and the number of reading are 1 to n. And epsilon are the errors, so the errors also are numbers 1 to n. Now, you may be thinking that what do we do with this equation; this basically we are trying to utilize the X values for 1 to n reading, Y values from 1 to n reading. And trying to find out the values of alpha and beta which are actually the population some the population which you have picking up.

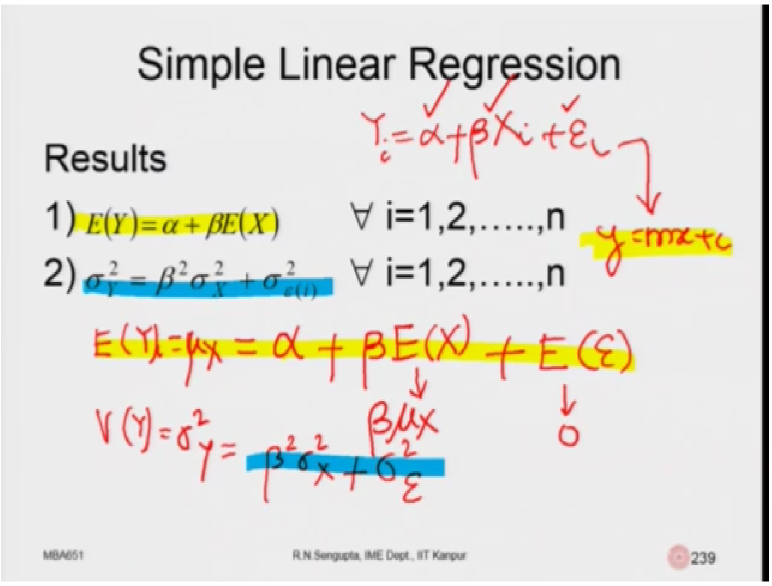
And trying to find an estimate of those population parameters such that when we predict for the nth plus 1 reading for Y considering that n number of readings are there and for the nth plus 1 reading we do not have y, but we do we do have x that means, X suffix n plus one reading is there we will use those values of X suffix and plus one in order to basically predict the nth plus 1 reading for Y. And once the actual value of Y is obtained we will try to easily find out that what is the error which is the difference between Y actual value for the nth plus 1 reading and Y predicted for the nth plus one reading. So, what I am saying that all are in words I mean I basically write those equations you will understand.

Now, obviously, there would be assumptions. So, what are the assumptions the first assumption is the covariance of the relationship of the interrelationship existing between the errors and the dependent variable independent variables is 0 for all the readings Y 1 to n that means the errors do not affect the in independent variable X. And that were also be carried forward when you are more than 1 x that means, p x's or k x's. The second one is that the errors would have a distribution technically we consider the distribution aware as is normal with 0 mean and 1 standard deviation in the simplest case. So, the second bullet point when the expected value epsilon 1, i is 0 it actually means that.

The third one is basically the errors do not effect each other. The hence the covariance existing between the errors from time period one to time period two, time period two to time period three, are all independent. Hence, this is 0; which means that we do not have the concept of heteroscedasticity in the problem such that the errors by themselves are not changing, they are changing, but they are not changing in the sense that the distribution are not changing if they are parameter values.

The fourth information is that the variance has a value of the errors as a value of sigma square suffix epsilon. We I did mention that in simplest case we take the variance as 1, but in general sense we will take the variance of the errors to be some fix value sigma square. And we consider the distribution of X to be normally distributed with the mean value of mu suffix x and variance of sigma square suffix x. And as X's are normally distributed as epsilon a normally distributed X's are independent to each of the with respect to epsilon, and obviously, alpha being a constant value. And as per the properties are normal distribution Y would also have a normal distribution with a certain mean and a certain standard deviation. We are going to come to that.

(Refer Slide Time: 10:11)



Now, let us find out the expected value of Y and similarly the variance of Y. So, the first part is for the expected value. So, when as if you remember that I mentioned that during the discussion the equation Y is equal to alpha plus beta X i plus epsilon. This is something like Y is equal to m x plus c. Now, this you will find out comes out to be right when we solve the equation. So, what we are doing it I am going to come to that.

So, let us take this Y has also a normal distribution with certain mean and certain standard deviation. Alpha is a constant, X is basically also a normal distribution with a certain mean and certain standard deviation, epsilon is also an error with the normal distribution with the 0 mean and sigma square, the variance. So, let us find out the expected value.

So, if I find out the expected value of Y, so that is this technically for the population. So, expected value of alpha is fixed is alpha this is so this becomes beta into epsilon X plus 0 because the expected value let me write it, it is better, wait, this is the error, and this becomes basically equal to 0. So, if you see it the expected value of Y, here the expected value of Y is equal to alpha plus beta into expected value of X which is mu X plus 0 because this is. So, this equation is obtained from here.

Now, we need to find out the variance of Y. And I will draw also in order to make you understand. So, once I finish, I will erase on this slide only and draw it. So, let us go to the variance. So, the variance of Y, consider this is equal to sigma square suffix Y. Now, we need to find out the variance for three terms which is the first one, the second one and third one and the covariance existing between the first and the second, second and third, and first and the third. Now, let us go one by one.

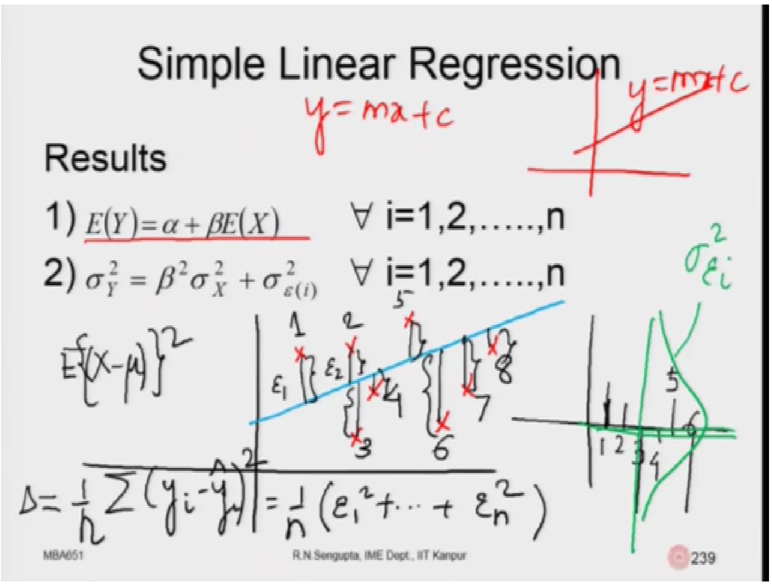
The variance of alpha which is a fixed quantity 0; variance of beta into X i would be beta square into variance of X because the variance of X is given. The third term is epsilon Y epsilon, so hence the variance epsilon as per the assumptions is sigma square suffix epsilon. So, all these three first three parts are done.

Now, let us come to the covariance factors. Now, the covariance if I consider for alpha which is a constant value with respect to any random variable would be 0 which means the covariance existing between the first term and second term is 0, covariance existing between the first term and the third term is 0. Now, let us consider the covariance existing between the second term and the third term. Now, covariance between the second term on third term would be 0, because you have intrinsically assumed before we started solving this problem that the errors do not affect X and vice verse, hence the covariance existing between X and epsilon is 0.

So, what are the terms left, there is no covariance term. First term variance of alpha is gone; only the variance of the second term and third term is there. Second term variance is beta square sigma square x; and the variance of the third term is given by sigma square epsilon. So, if you check the formula, so this is the equation we found out and this matches here. So, both the expected value and the variance for the Y which is the dependent variable have been proved. Now, you would be saying that I have promised that I will try to draw and illustrate that, I will do that.

So, let me erase whatever I have basically done. Now, I will I will missed one thing, I will come back to that later just one minute before I draw.

(Refer Slide Time: 15:22)



So, if you remember these equation and I had said y equals to m x plus c. So, this is the equation which you have done when you are kids. So, this is the equation y is equal to m x plus c this basically matches this. So, this is basically c this, beta is m and x is the expected value. Now, let us draw it. So, I will basically use colour. So, these are the axis and you have the values of x with y as this.

So, now I draw line, consider this is the line. So, if I consider the first point, point 1, 2, 3, 4 and so on and so forth. So, the first point and obliviously this is basically so this equation would have an error. So, actual value is of Y is here, predicted value of Y is the straight line. So, obliviously this is an error. So, this is basically epsilon 1, which is the error for reading one. Second equation actual Y is here, predicted value is Y is here. So, this is epsilon 2. Similarly, this straight line is epsilon 3, epsilon 4, epsilon 5, 6, 7, 8.

Now, if you consider the errors as such, so these are the errors. So, technically if I if I draw the errors, so there is some positive error. So, this is the first one. So, I will mark them as the I will I will basically mark it as 1, the second one is also positive here, third one is negative here, fourth one is negative here, fifth one is here, sixth one is here and so on and so forth. So, if I find out the distribution of these errors and plot it, so let me plot using another different colour, this will be a normal distribution with the 0 mean along the x axis. And the variance which would be there is basically sigma square epsilon i which you have already noted down number 1.

Number 2 is that why are we trying the line blue one line which we fitted, you would basically have some query that why did we use that. Why did we place it here, not above not below something? So, I have done it arbitrarily, but there is a reason for that. What we are trying to do is that we are trying to basically minimise the sum of the squares. Now, listen to this statement very carefully, if I am saying that I am trying to minimise the sum of the squares or finding out the average of the sum of the squares and trying to minimise that technically it means that you are trying to do something to do with the variance, because variance basically means the sum of the difference of the squares.

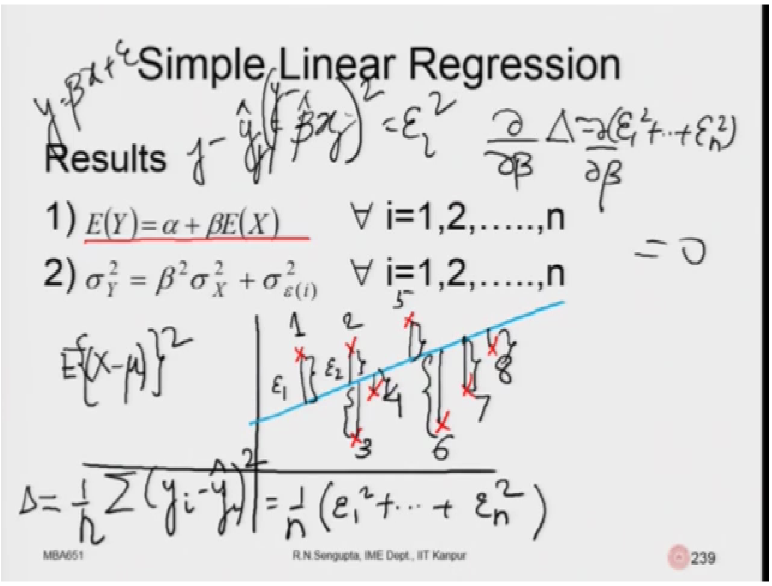
So, variance technically I we take as expected value of X minus mu whole square. So, what is X, X are the actual values. What is mu, mu is the average value. So, this actual value minus average value is the error square it. This actual value minus the mean value is the error, square it. So, this is the error for the third case square it; this is the error for this fourth case square it. This is the error for the fifth case, sixth case, seventh case, eighth case. So, what I am doing is that finding out the errors, summing them up and trying to basically minimise and minimise obliviously you put it to 0, because you want to basically minimize the error.

Now, question is that you will minimise with respect to what. Now, if you remember I did mention that when we are when we were discussing about alpha and beta we said that given X 1 to X n and Y 1 to Y n, X 1 to X n and Y 1 to n, we want to basically predict that what would be the predictive value or the forecasted value of Y in nth plus 1 period compare that with Y actual at the nth plus 1 period, and try to find out that how does this error basically match.

So, technically what we are and we also mentioned that the alpha and beta values are unknown. So, in the case, if the alpha values and beta are unknown, in this case you have only one alpha and one beta. Consider for the first case alpha is 0. So, in this case, this error which you have written down the square values that is actually sum of y i minus y i hat is basically the predicted values square it and you divide by n, because there are n number of readings, this is 2. So, this is basically delta or the. So, this is the error. So, I will put it as delta.

Now what we do is that, square this distance, square this distance, square this distance, square this distance, for all of them add them up and differentiate. Now, this is what you are doing for the differentiation. So, I will remove some portion here.

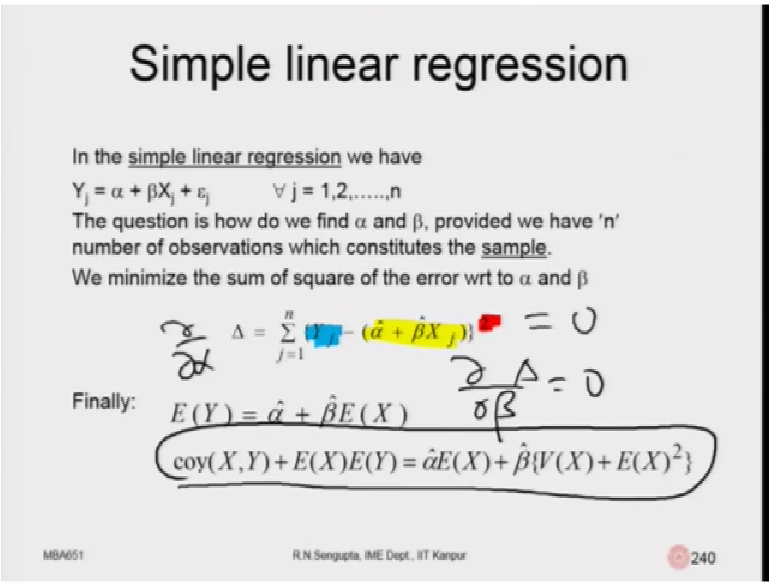
(Refer Slide Time: 21:20)



So, basically I will have the error which is basically sum of epsilon 1 plus epsilon n whole square. I will basically differentiate with respect to beta if there is only beta and put it to 0. Now, you will say that from where does the beta come. Now remember the y hat value is actually considering there is no alpha would be beta hat actually you want to find out multiplied by x that means, y i or j whatever would be this. So, when we put it and obviously, here see the error is not there.

So, technically the y actual value is equal to beta into x plus epsilon. So, when you put the expected and the predicted value error is not there. So, the difference between y minus this is equal to y minus this one whole square is equal to exactly equal to y i whole square. So, differentiate with respect to beta put to 0 and then basically utilize that beta for the prediction of the nth plus 1 value nth plus 2 value so on and so forth.

(Refer Slide Time: 22:43)

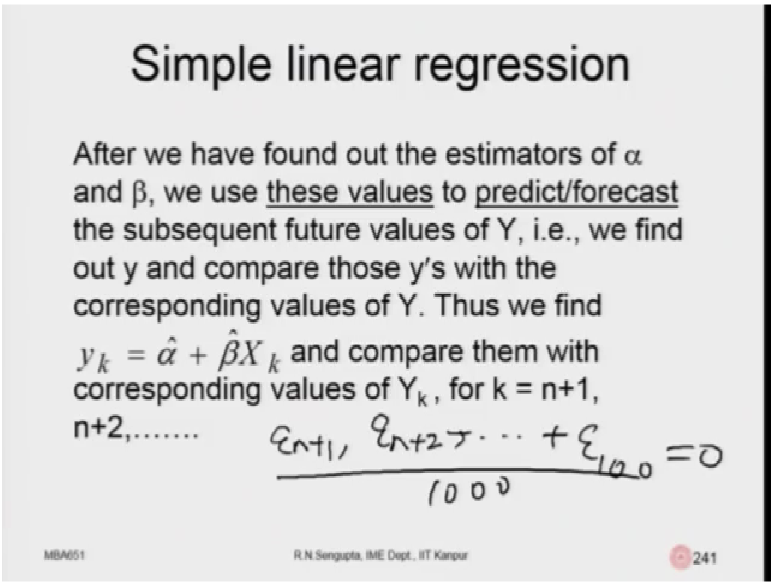


So, in the linear regression system, we have Y is equal to Y j or Y i is equal to alpha plus beta X j plus epsilon; and the j is equal to 1 to n number of readings. The question is how we find out alpha and beta as I mentioned. Provided we have n number observation which constitutes the sample we minimise the sum of the square is a error with respect alpha and beta and this is the error which I told. So, this is the this is the actual value; this is the predicted value, difference one that square it up. So, I am squaring it up. And this error is differentiated with respect to alpha or beta or gamma whatever it is.

So, this will be with alpha put to 0, this will be with respect to beta is equal to 0, solve the equations which you have and your work is done. So, finally, the expected value of Y would be equal to the predicted value or the estimated value of alpha which is alpha hat plus beta hat which we have already found out, beta hat into expected value of X because in that case the error is not there because you are predicting.

And now the covariances of obviously, there would be a covariance of X and Y we did mention the covariances of the error with X is not there. But obviously, there would be covariances between X and Y and that covariance of X and Y plus the expected value of X n into expected value Y would be given by the equation. We will be utilizing it later on accordingly.

(Refer Slide Time: 24:30)

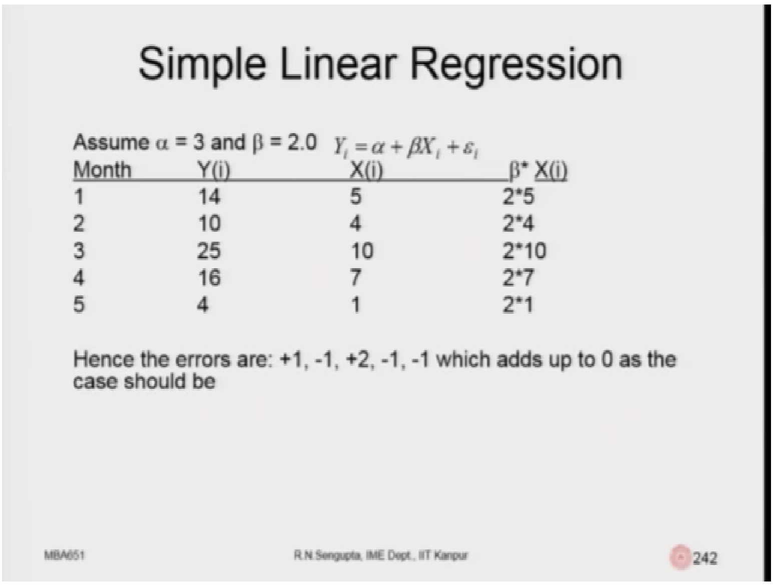


So, after we have found on the estimators as alpha and beta, we use these values to predict or forecast the subsequent future values of Y that is we find out small y and compare this small y's with corresponding values of Y capital Y which is basically the actual value. So, the hat which you put is basically the small values which are the predicted or the estimated and the capital values of the actual one difference between them is the error which you want to minimise the sum of the squares.

Thus we find out the equation y k is equal to alpha hat plus beta hat into x k and compare them with the corresponding values of capital Y k for k is equal to n plus 1 n plus 2 n plus 3 and so on and so forth. So, say for example, if the prediction is absolutely perfect, then the difference in the actual value and the and the predicted value for nth plus 1 reading, for the nth plus 2 reading, nth plus 3 reading and so on and so forth. So, these errors would be there for the nth plus 1 time. So, this is error for nth plus 1 time, they would be an error for the nth plus 2 time. So, you will continue see for example, you would take such 1000 values consider.

So, if you add them up and divide by 1000, this is the expected value. So, we know from the formula the expected value is 0 which basically corroborates. Actually if it is 0 it corroborates the fact that the assumptions which we have taken. Corresponding to the case that the expected value of the error is 0 and the variance of the errors is sigma square can be as done as a back calculation to prove that the methodology and the way we are proceeding is right.

(Refer Slide Time: 26:16)



So, assume alpha is equal to 3 for our case, beta is equal to 2, and the equation is equal to Y i is equal to alpha plus beta into X i plus epsilon i. So, i is equal to basically the reading number. Consider there are five reading numbers month 1, 2, 3, 4, 5. The actual values of Y for that time period 1 to 5 are given as 14, 10, 24, 16, 4. And the corresponding X i's are given in the third column starting from 5 to 1. So, obviously, when you multiply the beta, beta is given as 2 beta multiplied the X i's gives you the values of 2 into 5, which is 10 till the last values which is equal to 2 into 1 which is 2. Now, here the errors the errors is basically the actual value minus the alpha plus beta into X i.

So, if you find out in the first case the value which you have is basically the you want to basically find out the error. So, in this case, the error would be given by alpha is 3, beta into X i is equal to 10 which is basically 13, and the actual value of Y is fourteen. So, 14 minus 13 is 1. So, here the error comes out as one. In the second case the actual value of Y is 10, 2 into 4 is 8 which is beta into X i plus then the value of alpha is equal to 3. So, in this case, the actual value of alpha plus beta into X i is equal to 11. So, 10 minus 11 is minus 1. So, all the values which we have found out for the first, second, third, fourth, fifth reading.

So, technically if you remember I said that the sum of the errors in the (Refer Time: 27:55) should be 0. So, this is just a hypothetical example where you add up the values of the difference between the Y actual value and the predicted value which is the error they turn out to be 0. And obviously, similarly you can find out the variances of the errors technically it would be constant and that would basically be a certain sigma square value which you should have before hands such that you can basically corroborate the overall calculation which we have done. So, with this I will end this lecture and continue more discussion about the multiple linear regression in the next class.

Thank you very much and have a nice day.