Statistics for Experimentalists
Prof. Kannan. A
Department of Chemical Engineering
Indian Institute of Technology - Madras

Lecture - 38
Regression Analysis: Part B

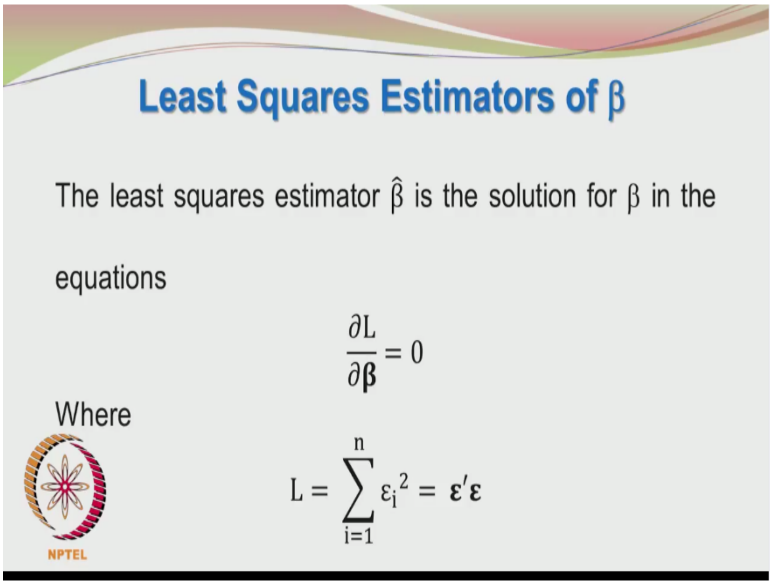
So we wrote down n equations and the moment we have n equations, we think that we can solve all those equations and get n estimates of the parameters but I told you that even though we perform n experiments, the number of regression parameters or regression coefficients we estimate is smaller than n. Essentially, that means we are not solving those n equations and n unknowns right.

It makes no sense also to have a huge number of parameters in our regression model. We are having such a large number that they are=the number of experimental observations. We need only a small set of parameters. So we do not solve those equations simultaneously. We adopt some other method.

You can solve n equations in n unknowns using the matrix method but we are going to use another method also involving matrices to get good estimates of the parameters beta hat, beta 0, beta 1 hat, beta 2 hat so on to beta k hat. So we have totally p parameters to estimate, p=k+1, the k regression coefficients associated with the k regressor variables+the intercept beta 0 hat, which makes it k+1, we call p=k+1.

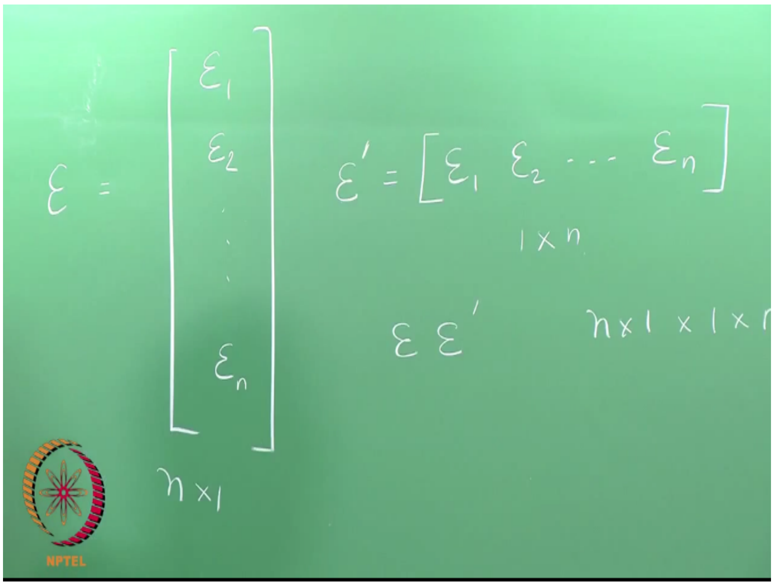
All these things are very clear. The method we are going to adopt is the least squares estimation technique for the parameter set given by the beta column vector.

(Refer Slide Time: 02:26)



So what we do is we define a function L which is defined as sigma i=1 to n epsilon i squared. In matrix notation, epsilon i squared may be written as epsilon prime epsilon. This is very simple. I will explain this.

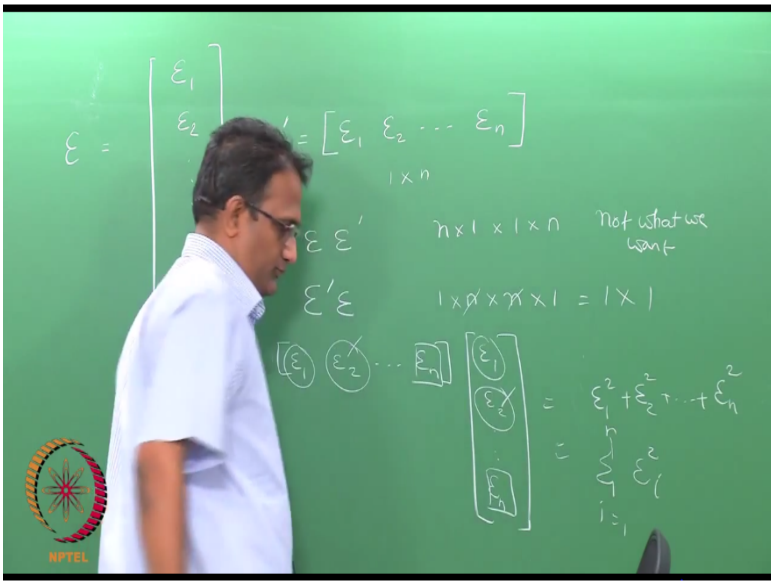
(Refer Slide Time: 02:54)



Epsilon was given by epsilon 1, epsilon 2 so on to epsilon n. When you take transpose that is what is given by epsilon prime you get epsilon 1, epsilon 2 so on to epsilon n. So this is n rows one column, this is one row n column, you cannot do epsilon, epsilon prime because then you will get n/1*1/n n rows one column, one row n column okay and when you do that you will get n/n. How is that possible?

Okay epsilon 1 squared and then you will get epsilon 12, so this is not what we want.

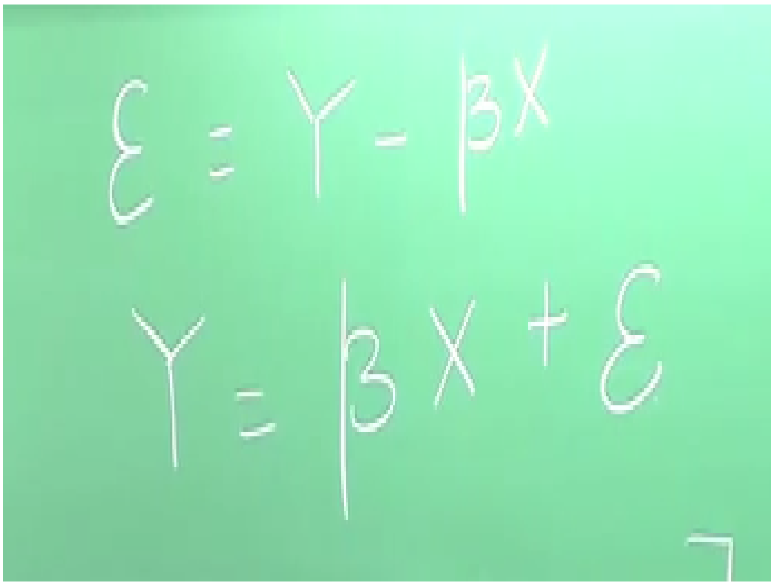
(Refer Slide Time: 04:00)



We want epsilon which is 1 by n*n by 1, then these 2 will get canceled out and then you get 1 by 1. What is epsilon prime epsilon? You have epsilon 1, epsilon 2 so on to epsilon n and then you have epsilon 1, epsilon 2 so on to epsilon n and here you get epsilon 1 squared+epsilon 2 squared+so on to+epsilon n squared. So what is happening is epsilon 1 is multiplied with epsilon 1 here+epsilon 2 is multiplied with epsilon 2 here so on to epsilon n is multiplied with epsilon n here.

So that we get sigma epsilon i squared, i running from 1 to n. So what we want to do is we want to identify the parameters beta such that dou L/dou beta=0. We want to minimize L. what is L? L is the sum of the square of the errors and what is error? That is very simple. What is error? This is very important.

(Refer Slide Time: 05:58)

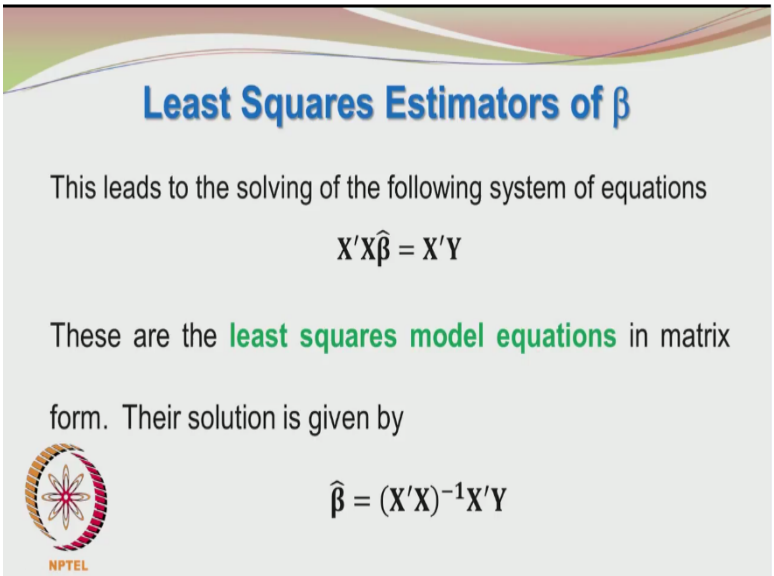


Error is Y-beta X. We wrote the model Y=beta X+epsilon. So the error term=Y-beta X. Sometimes you may have positive errors, sometimes you may have negative errors and to account for the error in an impartial manner, we do not want to add the positive errors and the negative error and then show a net small error or 0 error. We square the errors, so that irrespective of whether the error is positive or negative you are squaring all the errors.

So that all of them become positive and we get a complete total error, so we have sigma epsilon i squared and that may be written in matrix notation as epsilon prime epsilon. Now we have the sum of the square of the deviations and we want to minimize that. This is the good old least squares principle. You might have done that in may be higher secondary or in the second year of your engineering program.

But the main idea is the same. In matrix notation, we have dou L/dou beta=0, epsilon prime epsilon is also scalar.

(Refer Slide Time: 07:35)

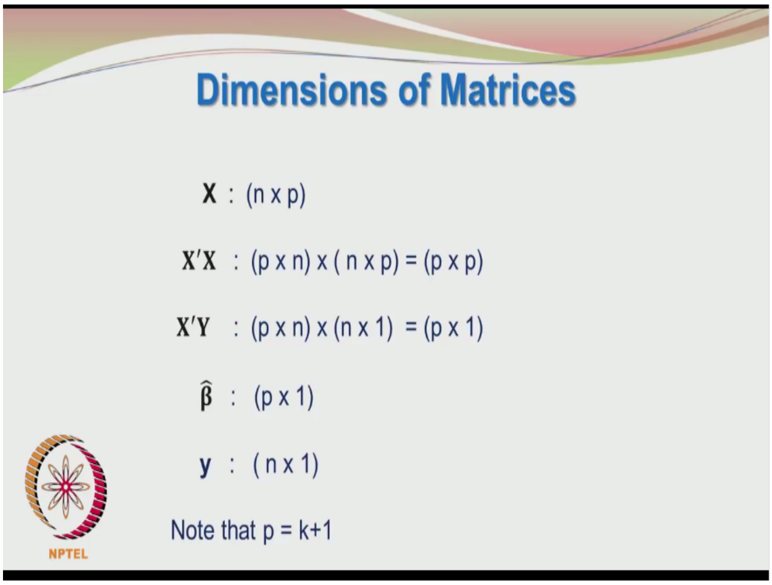


So you have after the dust has settled down X prime X beta hat=X prime Y or beta hat=X prime X inverse X prime Y. This is a very famous equation. You can predict the parameters for your regression model by multiplying X prime Y matrix with X prime X inverse. To find beta hat we pre-multiply X prime Y with X prime X inverse. So this is going to directly give us the set of parameters.

I am not giving you the proof and this follows from this relation. So to find beta hat, we pre-multiply X prime Y with the inverse of X prime X inverse. So if I put X prime X inverse on both sides, X prime X inverse*X prime X will become the identity matrix and then you have this matrix inverse multiplying with X prime Y. That will give you directly beta hat. You do not have to do these calculations by hand.

Sometimes for large matrices finding the inverses may become cumbersome and error prone, you can use the aid of mathematical software like MATLAB for instance to do the matrix manipulations.

(Refer Slide Time: 09:28)



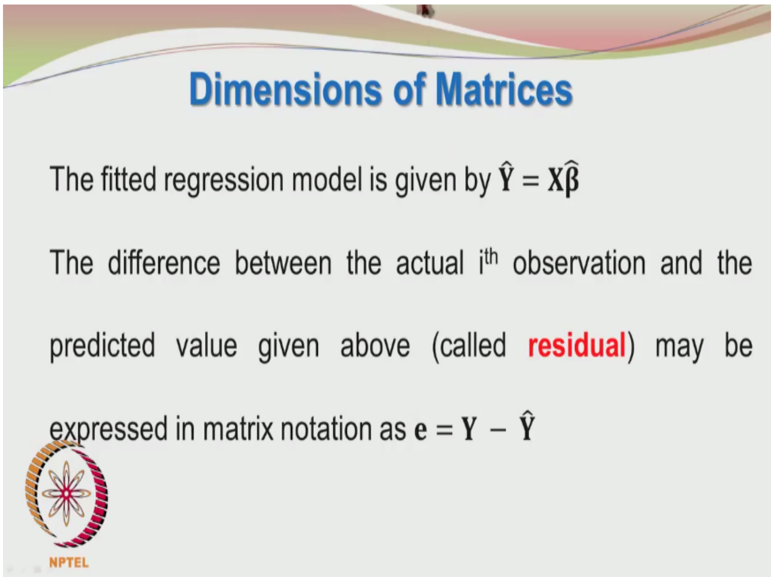
So let us look at the dimensions of these matrices. We are having X prime X beta hat=X prime Y beta hat=X prime X inverse*X prime Y. So are the dimensions consistent, it is a good time to summarize the dimensions of the different matrices involved. X matrix is having n rows and p columns, n experimental observations and p regression coefficients or the parameters beta 0 so on to beta k which would be k+1 parameters.

X prime X transpose of X multiplying with X, so if X is n/p transpose of X would be p by n. When I am saying transpose, I am interchanging the rows and columns. So suppose I have a matrix with a certain number of rows and certain number of columns. When I take the transpose of that matrix, I am converting rows into columns and columns into rows. So if I am having n by p for X matrix the transpose of the X matrix would have dimensions of p by n.

You had n rows and p columns originally in the X matrix. In the X prime matrix, you have p rows and n columns because the rows and columns have interchanged. So when I multiply these two, the n/n cancel out and then we have p cross p. X prime Y would be p cross n as we saw here and Y is a column vector of n observations so it will be n by 1 so X prime Y would be p cross 1.

And beta hat would be p cross 1, it is a column vector comprising of p parameters and one column. The p parameters are arranged row wise and this should be capital Y, which is the vector of responses, n rows and one column. I will just make it as capital Y.

(Refer Slide Time: 11:41)

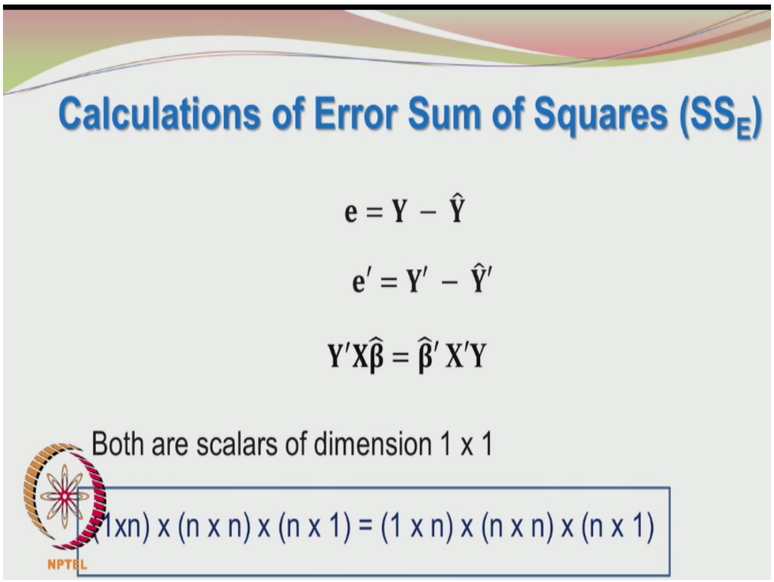


So the fitted regression model is given by Y predicted=X beta hat. The difference between the actual ith observation and the predicted value given above is called as a residual and may be expressed in matrix notation as e=Y-Y hat. So we have been using epsilon and now I am using e. There is a reason for this change.

Epsilon represents the true error, the random component of the experiments reflected in the form of epsilon, the random experimental error but I am using e. I am saying that e is a residual and that residual may be only due to the random error or it may also be including the unexplained effects in the experiment because of inadequate modeling okay.

If my model is not fully able to explain the variations in the experimental response, then that discrepancy cannot be dismissed as random error. So my residual contains possibly unexplained variability and also the experimental error, so I am using e here. If all the possible variability has been accounted for in the model, then the residuals would be a true reflection on the random error component.

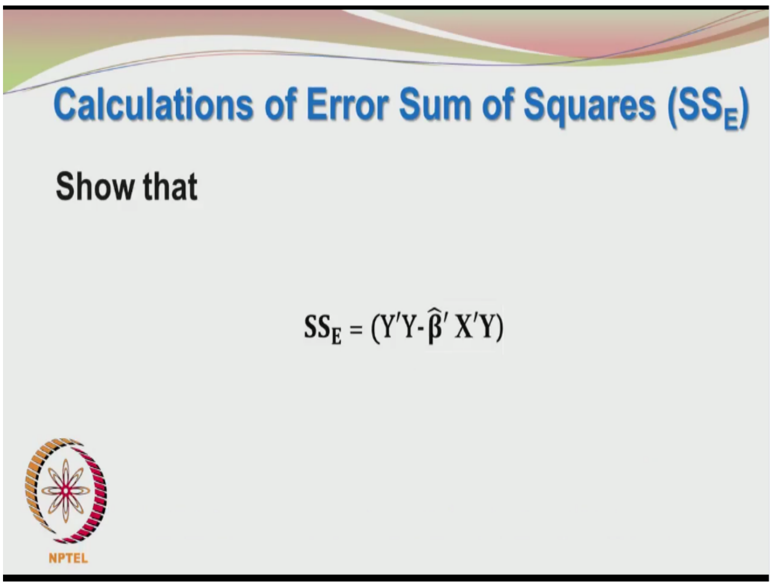
(Refer Slide Time: 13:30)



So first we will calculate the error sum of squares. The moment we see the term sum of squares, we can guess that some analysis of variance is involved. So you have Y-Y hat=e which is the residual, e prime will become Y prime-Y hat prime and then we have Y prime X beta hat=beta prime X prime Y but both are the same. Both are the same because both of them are scalars.

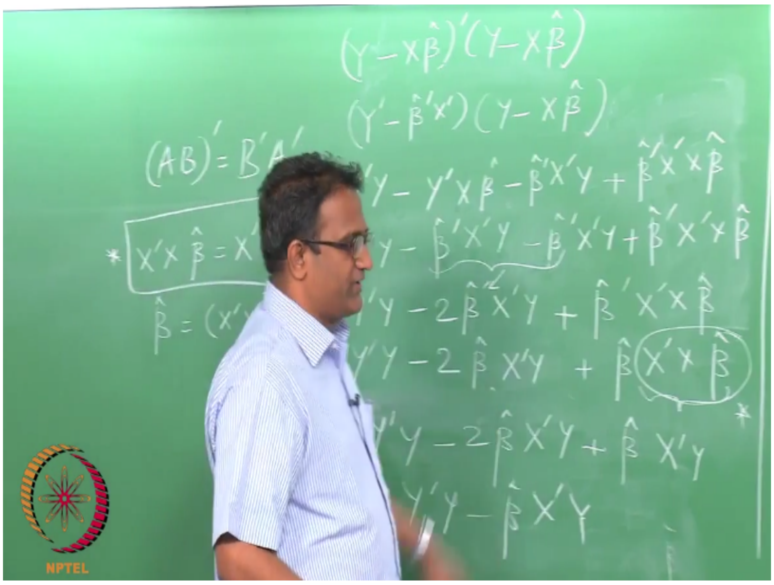
It can be easily shown that Y prime X beta hat is a scalar, it has dimensions of 1 by 1 that calculation is shown here and then you also have beta prime, X prime Y which is again a scalar and that dimensions are also shown in this calculation. The n's will cancel out nicely leaving 1 by 1 right.

(Refer Slide Time: 14:32)



Now we can show elegantly that the sum of squares of the error is given by Y prime Y-beta hat prime X prime Y. This is very interesting. Let us see the proof.

(Refer Slide Time: 14:52)



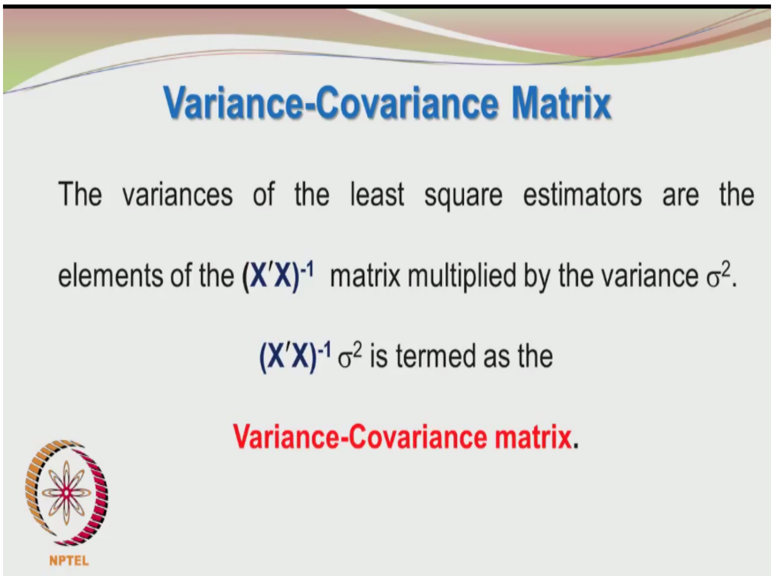
Turn to board and then do the derivation directly so that people who are not familiar with these can follow the steps. So we have Y-X beta hat prime*Y-X beta hat, so I take the prime inside and we know that when you take A of B prime this becomes B prime A prime so using that, there is no problem with Y, it becomes Y prime. X beta hat whole prime becomes beta hat prime*X prime*Y-X beta prime.

So this becomes Y prime Y-Y prime X beta hat-beta prime X prime Y+beta prime X prime X beta hat okay. So I am just multiplying the last two terms here. So this will be beta hat prime X prime X beta hat. So we saw that from the previous slide that Y prime X beta hat=beta prime X prime Y. So you also have –beta prime X prime Y+you have this term. So these 2 can be written as -2 beta hat prime X prime Y.

So we get Y prime Y-2 beta hat prime X prime Y+beta hat prime X prime X beta hat and we also know by definition that X prime X beta hat=X prime Y. You may recorrect that, we found the parameters beta hat by taking X prime X inverse X prime Y, so this is the equation, which I can use here and that becomes beta hat X prime Y, you have -2 beta hat X prime Y, so you get Y prime Y-beta hat X prime Y. So this completes the derivation.

This is always good to go to the board and do the derivations. PowerPoint's also have their own charm and make teaching more convenient. In other hand, writing on the board you make mistakes, you correct the mistakes and learn in the process.

(Refer Slide Time: 18:12)



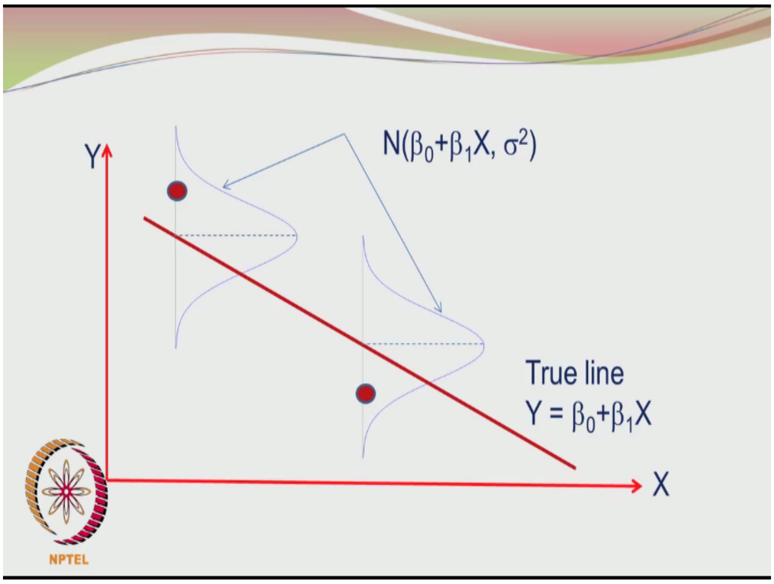
I hope that you would also feel interested to do the derivations independently on a paper and see whether your results are matching with the final expected results. So now we are going to talk about a very important property of linear regression analysis. It is the variance-covariance matrix. So obviously we are going to talk about a matrix and we have found the parameters beta 0, beta 1 so on to beta k.

We want to know how precise these parameter estimates are. So to have an idea about that we can use the variance-covariance matrix. So let me just introduce the matrix to you and then we will talk about how to apply it in real regression problems. So the variance of the least square estimators are the elements of X prime X inverse matrix multiplied by the error variance sigma squared.

Where did we see sigma squared previously? I showed a figure where we had the true relationship line and we showed the experimental data scattered around this line. We said that the scatter was described by a probability distribution, which was a normal distribution. The mean of that distribution was given by the equation but the data was not exactly aligning with that value given by the equation or the data was not present exactly at the mean value.

But it was present somewhere else because of random effect. This probability distribution was having a variance sigma squared. All the data points which were scattered around the true line had the same variance sigma squared. Let me go to that figure again.

(Refer Slide Time: 20:33)

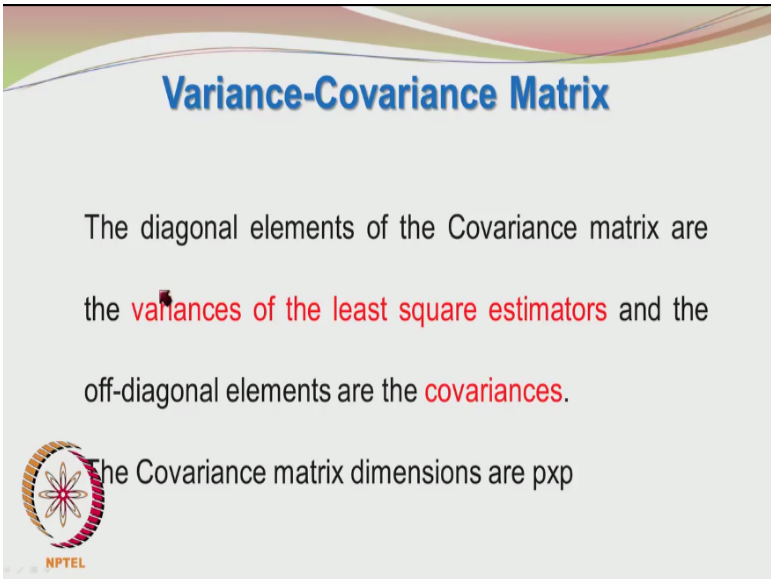


So here we have this particular figure. You have the true line and then you have the data scattered around this line and the scatter is because of experimental error. The scatter is described in terms of probability distributions and these are normal distributions most conveniently and the mean of these distributions are given by the equation. This mean would be different from this mean.

Because this X value is different from this X value. Here you have one X value, here you have another X value. Then you have the data points, which are scattered around this and the probability distribution is having a mean given by beta 0+beta 1X and changes with X but it has a constant variance sigma squared. This is the sigma squared we are going to use in the variance-covariance matrix.

So now coming back to the variance-covariance matrix. We have X prime X inverse multiplied by sigma squared and that sigma squared was the constant error variance but unfortunately we do not know sigma squared. We need to estimate the parameters beta 0, beta 1, so on to beta k. We also need to estimate the error variance sigma squared.

(Refer Slide Time: 22:14)



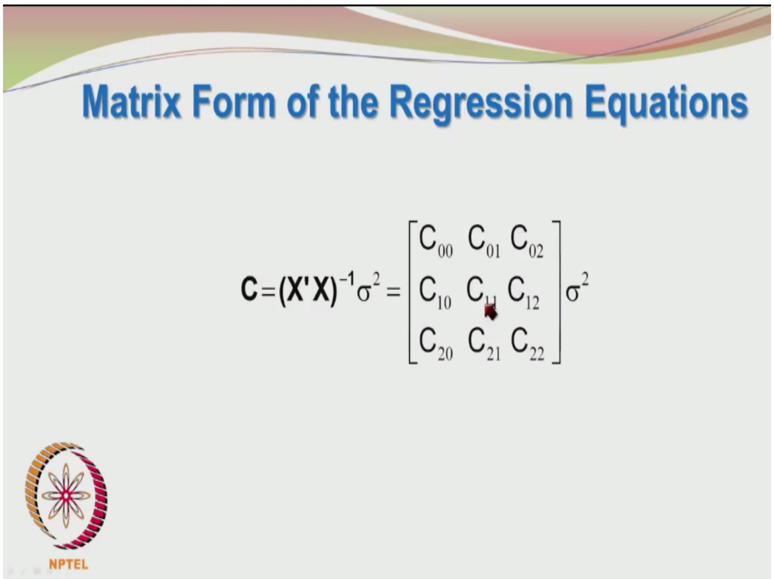
We are having a variance-covariance matrix. That term implies that matrix contains both variances as well as covariances. Which are the variances and which are the covariances? The diagonal terms in this matrix X prime X inverse sigma squared matrix represents the variances. The diagonal terms means first row, first column and the second row, second column, the element corresponding to third row, third column and so on.

So if you look at the matrix, the main diagonal will comprise of the variances of the estimated parameters. The off-diagonal elements of the variance-covariance matrix will represent the covariances between the parameters. So just now we are introducing the variance-covariance matrix. It is important and enough at this point if you understand that the variances are given by the diagonal terms and the covariances are given by the off-diagonal terms.

And it is also important to note that the variance-covariance matrix is symmetric. What is a symmetric matrix? A symmetric matrix is one whose appearance is unchanged when you change rows into columns and columns into rows. So you can have a symmetric matrix and when you interchange the columns and rows it appears to be just as the same. It is also important to note that the variance-covariance matrix are simply called as the covariance matrix.

Sometimes people call it as a variance matrix also. So the covariance matrix dimensions are p cross p. Do not get confused if sometimes you see variance-covariance matrix or on other times you see covariance matrix, both are the same.

(Refer Slide Time: 24:17)

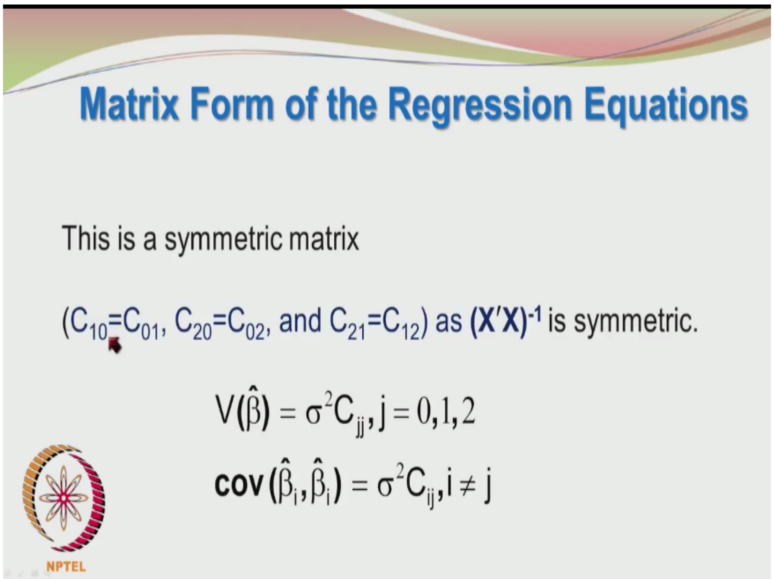


Finally, we have come to the form of the variance-covariance matrix or simply the covariance matrix that is represented by C. It is bold because it indicates a matrix and not a scalar and here we have X prime X inverse sigma squared. Let us represent the columns of X prime X inverse as C00 C01 C02 C10 C11 C12 C20 C21 C22. So this represents the diagonal of the matrix. So C00 C11 and C22 are the diagonal elements of this matrix.

And the off-diagonal elements are given by those entities, which are not along the diagonal. So all elements other than C00, C11 and C22 are off-diagonal elements. So you have sigma squared outside. You may as well take sigma squared inside and multiply all these terms with sigma squared.

So this is a symmetric matrix. What is a symmetric matrix? If I change rows into columns and columns into rows, the matrix appearance is unchanged that means C01 will be=C10, C02 will be=C20, if C01=C10 then it would appear as if there is no change. Similarly, C02 will be=C20 and here also C21 will be=C12 and the C20 will be=C02. So then you have a symmetric matrix.

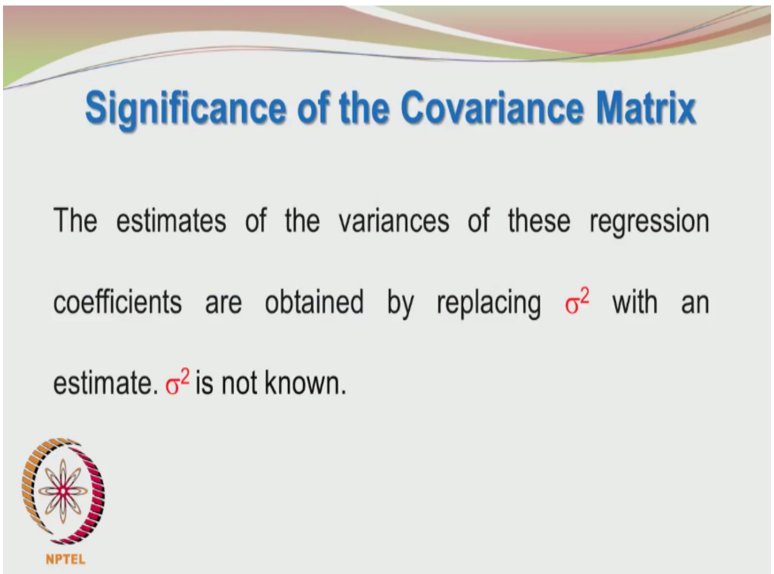
(Refer Slide Time: 26:14)



And that is what I have represented here. C10=C01, C20=C02, and C21=C12 and this is X prime X inverse matrix as a symmetric matrix and the variance of the estimated parameters beta hat are given by the diagonal term sigma squared Cjj where j is=1 or 0 or 2 and the covariance between two different parameters, beta i and beta j are given by the off-diagonal terms.

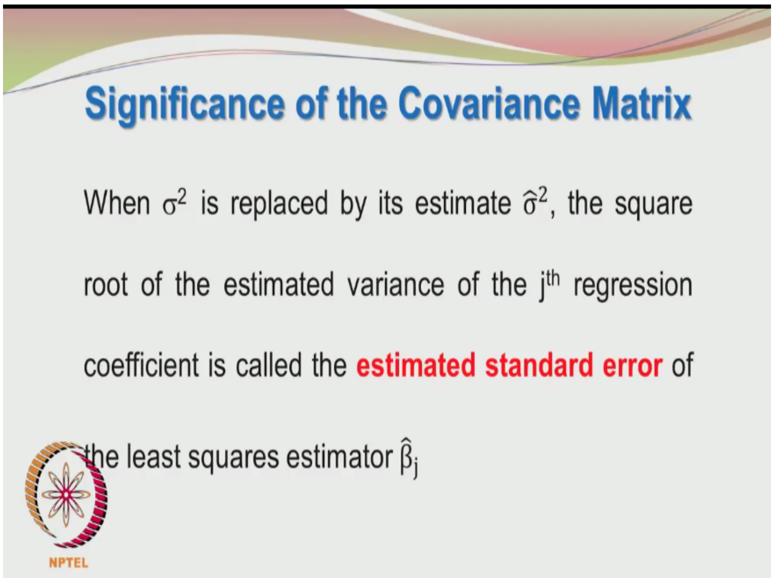
So there is obviously a typo here, which I will correct. So that is what you have here, the covariance between two different parameters beta i hat and beta j hat will be sigma squared Cij where i is!=j.

(Refer Slide Time: 27:09)



As I said earlier, we do not know the value of sigma squared and hence we need to have an estimate for the error variance and we also have to make sure that this estimate is the true error variance and not have any systematic influences.

(Refer Slide Time: 27:34)



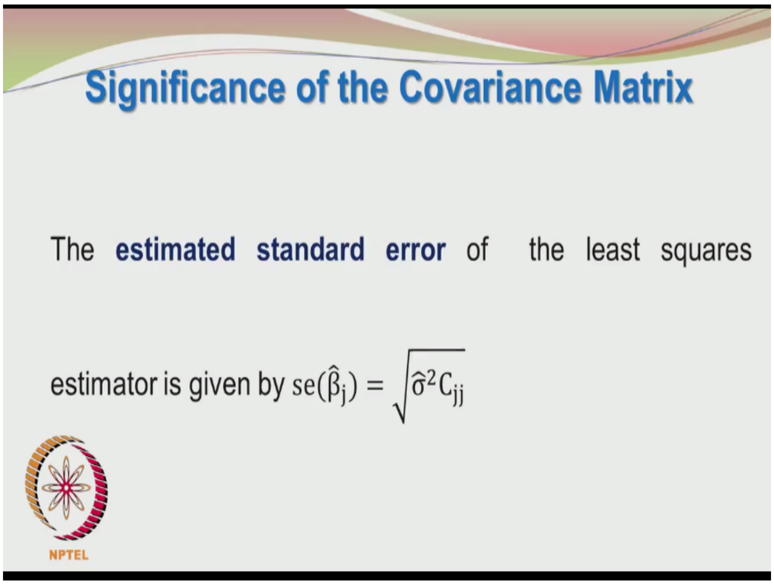
So we want to replace sigma squared by an estimate of sigma squared, we call it as sigma hat squared. In the true regression model, we had Y=X beta+error, that beta is the column vector of true parameters describing the experimental phenomena or rather the phenomena investigated by the experiments. Since we do the parameter estimation based on the available experimental data, the influence of errors are also there.

And hence our model is not accounting for the random phenomena. It is only accounting for the systematic phenomena and hence we are only getting estimates of beta and that is given by beta hat. Even though we would like beta hat to be as close as possible to beta, we may not be able to achieve the aim because the data is subject to experimental uncertainty, which is not included in our regression model okay.

It is accounted for separately as the error term. Similarly, when you want to use sigma squared, since the true value of sigma squared is not known, we need an estimate of sigma squared, which we use in our calculations and we call that sigma squared as sigma hat squared.

How to find it out? We will see shortly and then once you are able to find sigma hat squared, the square root of estimated variance of the jth regression coefficient is called as the estimated standard error of the least squares estimator beta hat j. Now we do not use the term standard deviation here. We are again taking the square root of the estimated variance of the jth regression coefficient and we call it as estimated standard error. It is not called as standard deviation.

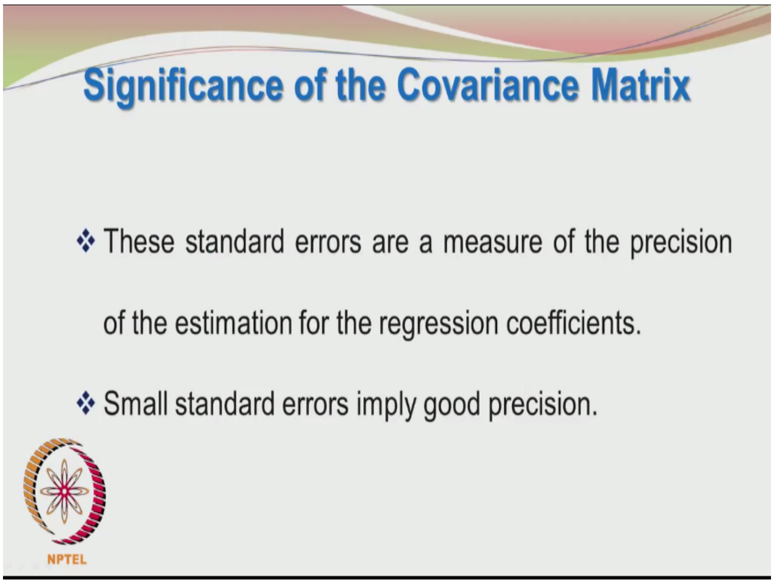
(Refer Slide Time: 29:57)



So the estimated standard error of the least squares estimator is given by se, se stands for standard error for beta hat j and that is given by square root of sigma hat squared Cjj. So this j here is matching with the j's given here and this is an estimated value and we are using the variance-covariance diagonal element multiplying it with the estimated error variance and when we take the square root, we get the standard error of the estimator beta hat j.

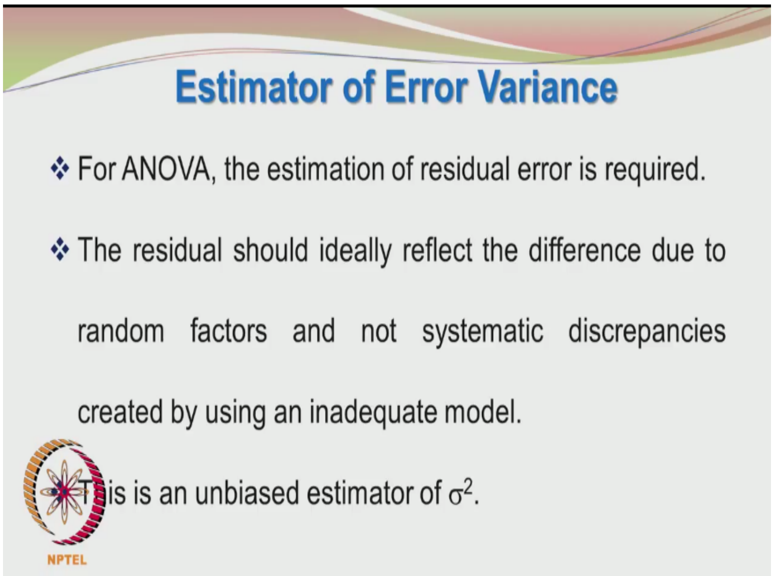
How did we get the beta j hat or beta hat j? It is the least square method we adopted to find this parameter and hence it is called as the least square estimator of beta j and it is represented by beta hat j.

(Refer Slide Time: 31:00)



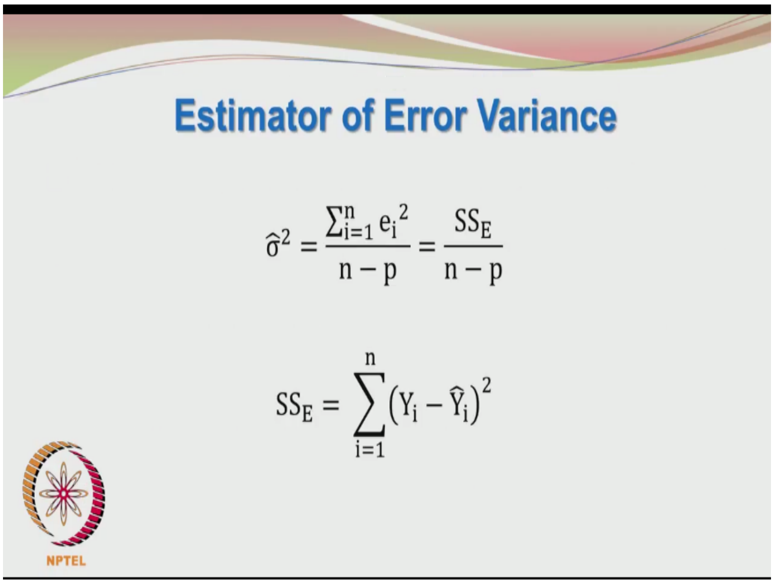
The standard errors are a measure of the precision of the estimation for the regression coefficients. Small standard errors imply good precision. So it is a measure of the fuzziness associated with the estimated regression coefficients. If the fuzziness is too much, then there is a big spread around the estimated beta hat but if the spread given by the standard error is quite narrow, then we have estimated beta's or the beta hats with good precision or reasonable precision.

(Refer Slide Time: 31:38)



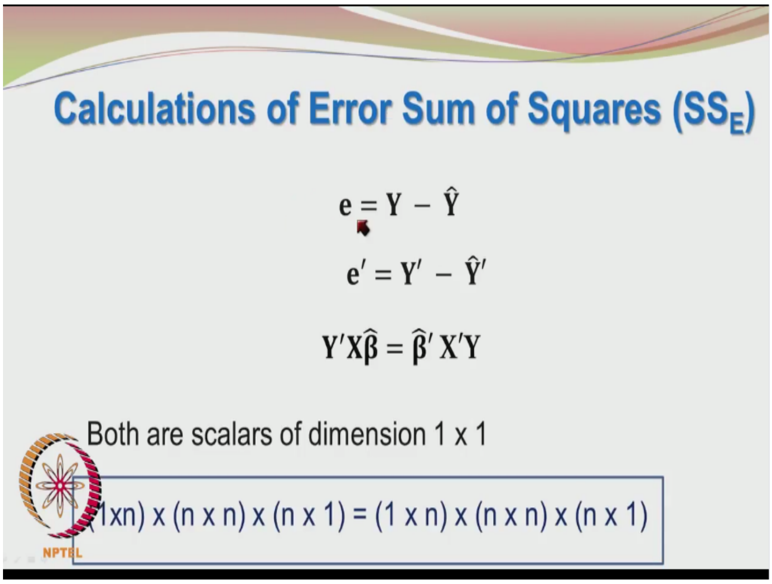
For analysis of variance purposes, the estimation of the residual error is required. In order to also get an estimate of sigma squared, we need the estimation of residual error. The residual should ideally reflect the difference due to random factors and not systematic discrepancies created by using an inadequate model and the residual error is also an unbiased estimator of sigma squared.

(Refer Slide Time: 32:12)



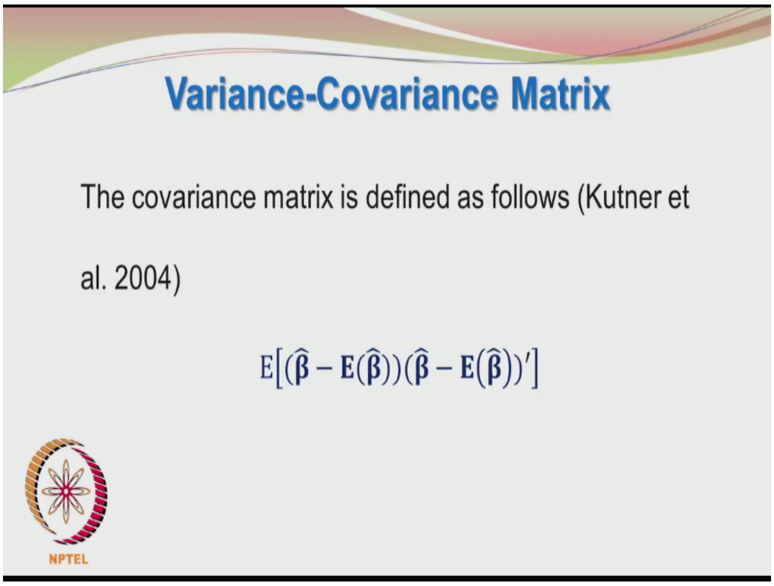
So we define the error variance in the following manner. We have sigma hat squared=sigma i=1 to n ei squared/n-p and that is given by sum of square of the error/n-p. It is the difference between the actual experimental value and the predicated value.

(Refer Slide Time: 32:37)



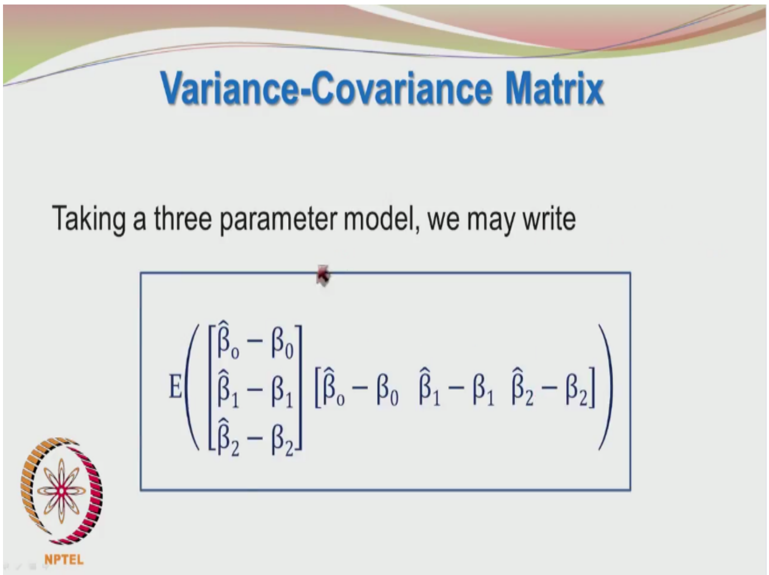
The error in the column vector form has Y-Y hat experimental vector and this is the model predicted vector, each entity we may represent it as ei, i running from 1 to n, so it will be Yi-Y hat i for the residual i or the ith residual. Sum of square of the error is actually given here itself, i=1 to n Yi-Y hat i whole squared and that is summed to give the sum of square of the error.

(Refer Slide Time: 33:16)



So if you look at Kutner et al 2004 reference, we can define the covariance matrix or the variance-covariance matrix as the expected value of beta hat-expected value of beta hat that is multiplied by again transpose of beta hat-expected value of beta hat. So you have this and we have to expand it and express it in matrix form.

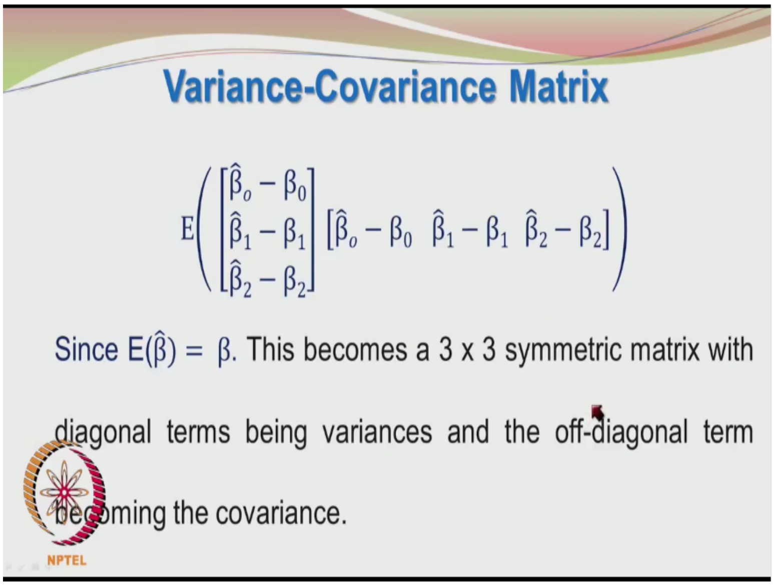
(Refer Slide Time: 33:45)



So we have expected value of beta 0-beta 0 beta 1 hat–beta 1 beta 2 hat-beta 2 so we have that matrix multiplying with this matrix comprising of only one row and 3 columns. So we have transpose of this written here. Transpose I am converting rows into columns so I am having one column here with 3 rows and converting it into a matrix with 1 row and 3 columns. So this becomes beta 0 hat-beta 0 and then this goes to the row element beta 1 hat-beta 1.

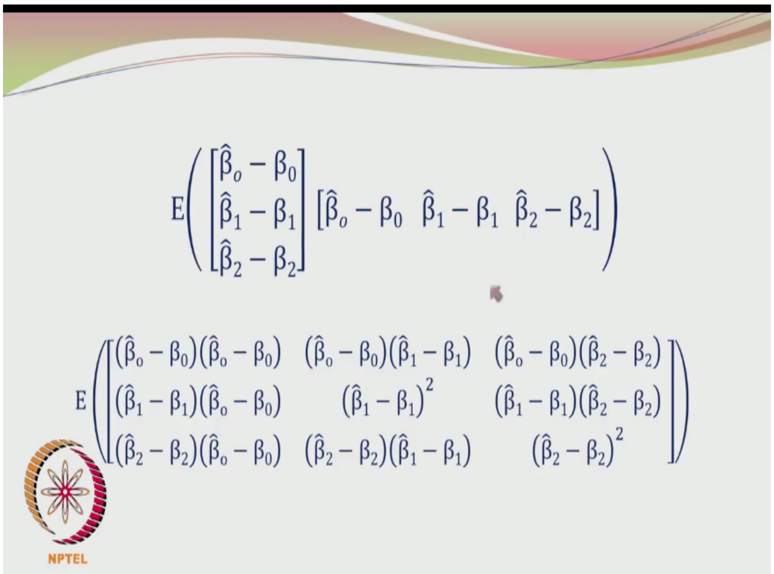
Then it goes to the next row element beta 2 hat-beta 2. So this is what we have. I can first multiply this and then find the expected value as given by E.

(Refer Slide Time: 34:42)



And also an important thing to realize is the expected value of beta hat is the parameter beta itself. So this is an unbiased estimator beta hat. Now when we first multiply this, it becomes a symmetric matrix and the expected value of the diagonal terms will become the variances and the expected value of the off-diagonal terms will become the covariances.

(Refer Slide Time: 35:21)

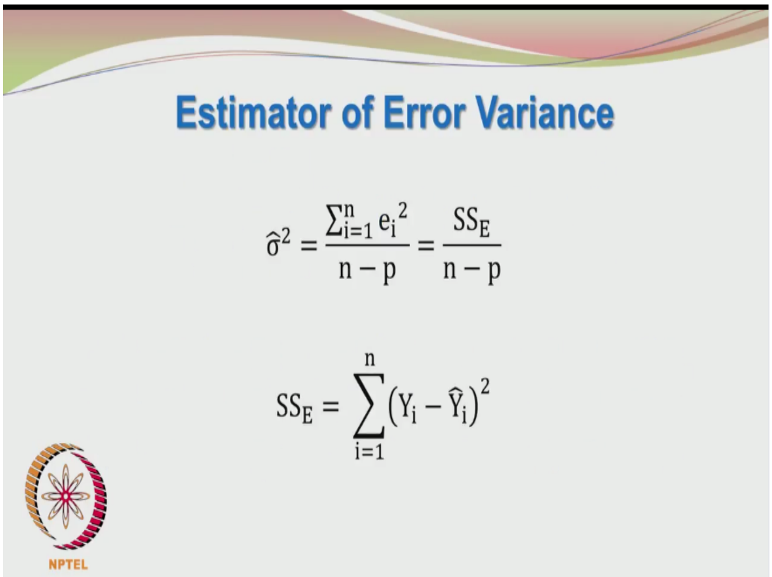


So let us do the multiplication, beta 0-beta 0*beta 0 hat-beta 0 that is what you have here. Beta 0 hat-beta 0*beta 1 hat-beta 1 and that is what you have here. You may want to do the calculations on your own to see whether you get this particular form. So this term here is=this term and this term here is=this particular term so the matrix is symmetric, we can show it for other terms also.

What are the terms we can show? So you have 32 will be=23 third row and second column element is this that should be=the second row and third column element and that is what these two are matching. So we can conclude that the matrix present inside the argument is symmetric. Then we apply the expectation to all these elements in the matrix and we know that the expected value of beta 0 hat-beta 0 squared is nothing but the variance of beta 0 hat okay.

So that becomes quite straight forward and the off-diagonal terms will become expected value of beta 0 hat-beta 0*beta 1 hat-beta 1 and that would relate to the covariance beta 0 hat and the beta 1 hat.

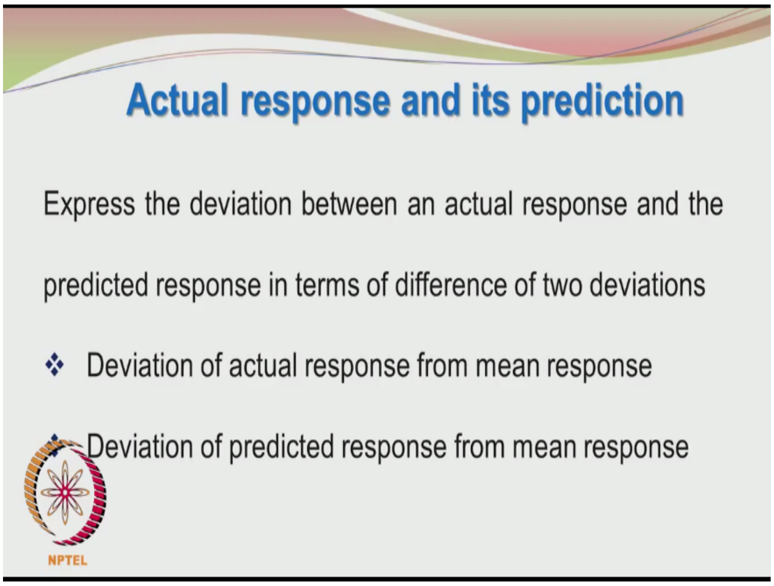
(Refer Slide Time: 37:00)



So to find the sigma hat squared, we have sigma i=1 to n ei squared/n-p that is sum of square of error/n-p. So this is how we are finding an estimate of sigma squared, the error variance and since we are finding an estimate, we denoted as sigma hat squared and sigma i=1 to n ei squared/n-p=sum of square of error/n-p, so that is accounted for and then what we do is we have that value sigma hat squared and then we can multiply all these elements with that sigma hat squared.

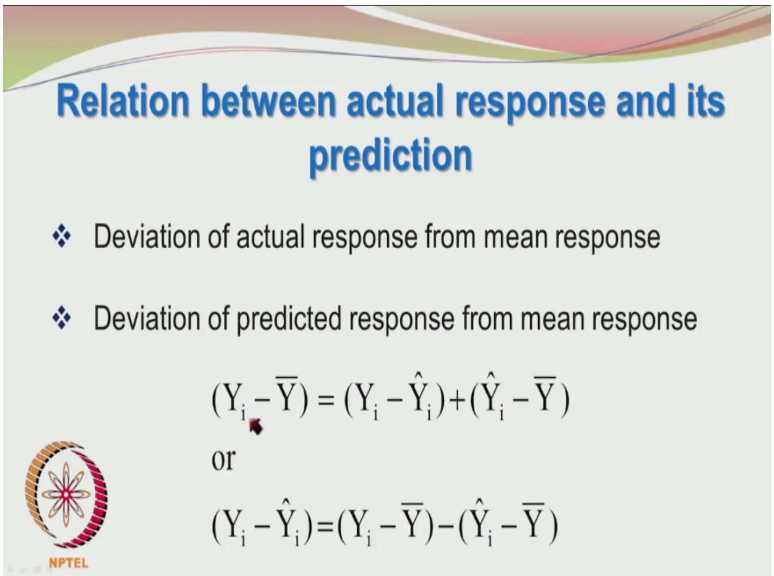
And then we can find the different variances and covariances of the parameters and their combinations. So when I multiply this, this would become the variance of beta 0 hat and this would be the variance of beta 1 hat and so on.

(Refer Slide Time: 38:15)



Now we are going to talk about regression sum of squares and error sum of squares. So what we do here is we express the deviation between an actual response and the predicted response in terms of two deviations. Deviation of actual response from mean response and deviation of predicted response from mean response. So rather than putting it into words let us see in symbols.

(Refer Slide Time: 38:44)

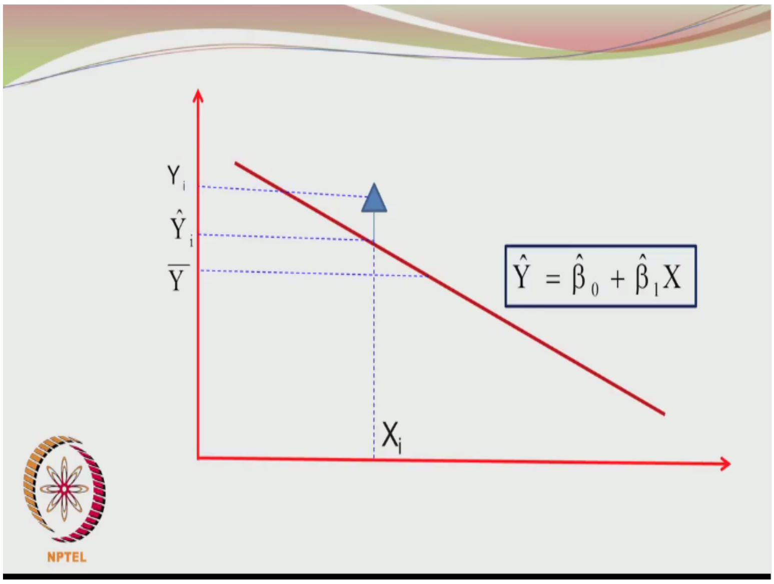


So first we are expressing the deviation of a particular response from the average value of the observations. You are conducting all the n experiments and you take the average value of the response, this is Yi-Y bar and that may be written as Yi-Y hat i+Y hat i-Y bar. This will cancel out with this, so you will be finally getting Yi-Y bar.

Or we can write this equation involving the deviation of the actual experimental data point from the predicted value Y hat i as Yi-Y bar-Yi hat-Y bar, so we are defining the residual as the difference of two entities. Deviation of the experimental observation from the mean value, deviation of the predicted value from the mean value okay. This is very interesting. Instead of directly writing Yi-Y hat i, you are writing it as Yi-Y bar-Yi hat or Y hat i-Y bar.

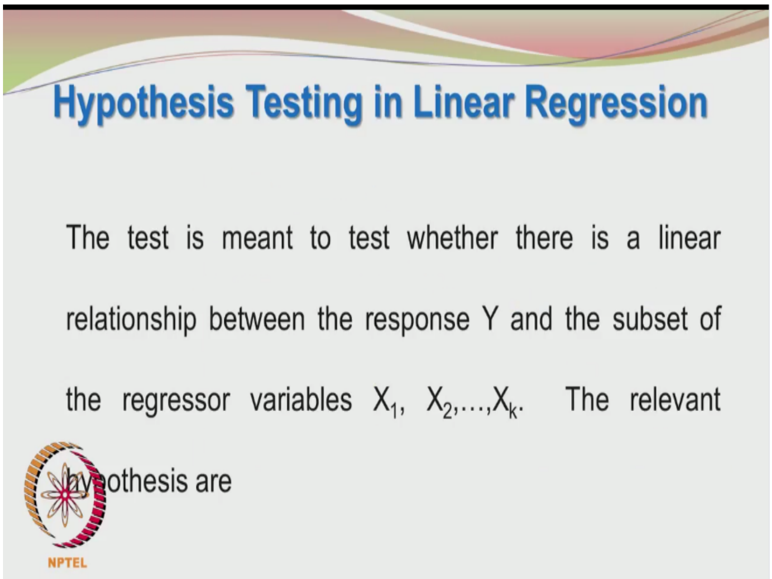
So you are subtracting and adding Y bar in this expression. So the residual which is a discrepancy between the actual experimental value and the predicted value is expressed as the difference between Yi-Y bar and Yi hat–Y bar. So what is the discrepancy of Yi with respect to the mean value that I will subtract with the discrepancy between the predicted value and the average value Y bar.

(Refer Slide Time: 40:57)



So that can be shown graphically in a nice fashion. This is the actual experimental data point and that is slightly off from the prediction value. This is Yi and then you have Y hat i and this is the average value based on all the experimental data points. So we want to find the deviation between these two the residual that may be expressed as the deviation of Yi with respect to Y bar-the deviation of Y hat i with respect to Y bar.

(Refer Slide Time: 41:40)



Next, we go to hypothesis testing in linear regression. So we are going to start a new phase in the regression analysis and it is also interesting to note that whatever we studied in the first part of statistics for experimentalists are coming into play in the second part as well or in the second phase as well. So now you will be able to appreciate with your background in inferential statistics on what we are going to do with linear regression.

When we apply those to linear regression, things will become very clear and you also will understand why we are doing these kinds of tests. So the hypothesis test is what we are going to study in detail in the next lecture. I request you to not only brush up your fundamentals in linear algebra but also look at the concepts we covered in hypothesis testing.

Find out what is meant by level of significance, the p value, the region of acceptance, region of rejection, the confidence intervals. Please refresh your concepts on these topics and if and once you have done so, whatever we are going to discuss next will become very simple. Thank you for your attention.