Essentials of Data Science with R Software - 2
Sampling Theory and Linear Regression Analysis
Prof. Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology Kanpur

Linear Regression Analysis
Lecture - 48
Multiple Linear Regression Analysis
Properties of OLSE and Maximum Likelihood Estimation

Hello friends, welcome to the course Essentials of Data Science with R Software 2, where we are trying to understand the topics of Sampling Theory and Linear Regression Analysis. In this module on linear regression analysis, we will consider the multiple linear regression analysis. So, you will recall that in the earlier lectures what we have done up to now, just a quick review.

We consider the multiple linear regression model and we have obtained the ordinary least square estimators of the regression coefficient vector β, right. And then, we obtain the fitted model and then we obtain the residuals. And we also learnt how to implement and how to obtain those values in the R software.

Now, I will try to continue, and in this lecture I am going to concentrate on two aspects; first is this whatever ordinary least square estimator we have obtained, I will try to state the properties of those estimator. And then I will try to come to the aspect of maximum likelihood estimation. So, the first question comes that, why should I state you these properties and then why only state why not proof?

The first thing is this, when you are trying to obtain a model from on the basis of some small set of data. And this model is going to be valid for a very big population, so the question is this you have to be confident the way you are obtaining the model is good or bad number one. Ultimately, whatever is the definition of statistical modelling, for us now it is a very simple thing we have to estimate the parameters.

So, now the question is this, there are different types of estimation methods for estimating the parameters. Now, I would like to know that whatever values I am estimating for my parameters on the basis of random sample of data, are they giving us a

good value or not? Now, once I come on the aspect of good, then this good is defined with respect to the statistical properties.

Now, the question is this, I expect from you that you have a reasonable background in statistics, but there may be some audience, some students, some faculty who have not done a course on statistical inference like what we do in the undergraduate or say post graduate courses in statistics. So, giving the proof to them will make possibly make their life difficult.

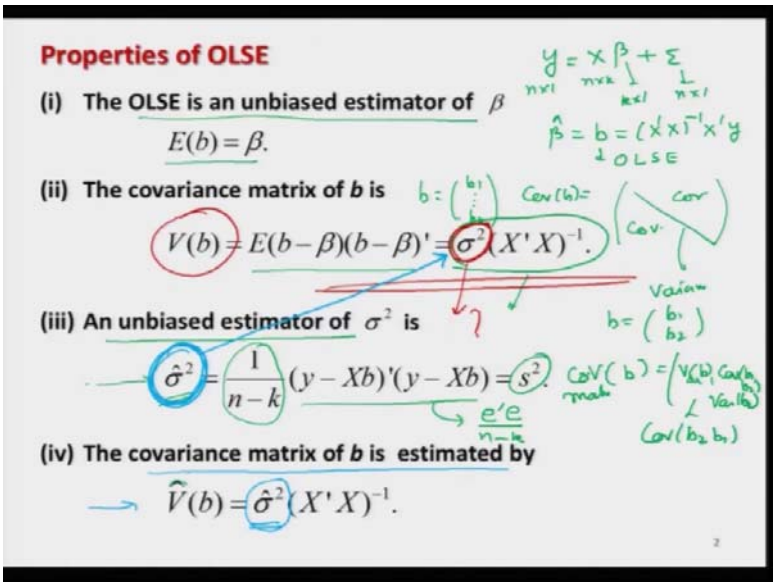
But, as a user I believe that they must know that what are the different properties and you have to believe on me that if you want you can go to the bo, oks and had to and lo, ok at the proofs of those results. But this will give them a confidence that whatever I am doing, that is correct.

And more so ever, when I come to the field of data sciences many people are now venturing into this area. People from computer science, people from different types of subject, they are coming; they are trying to learn this data sciences. It is not only those people who have done a full course on statistical inference they are doing or they want to become only the data scientists.

So, keeping in mind all this factors, what I have done, I have just comprehended all the results, so that you can be confident. Wherever the results are simple to prove, there I am trying to do it and rest I would leave it up to you, that how much you want to learn in the theory.

I will try to give you the sufficient concept with sufficient theory. And after that I will try to show you how to conduct the maximum likelihood estimation in the framework of multiple linear regression model, right, ok. So, let us now begin our lecture, ok.

(Refer Slide Time: 04:38)



So now, you can recall that our model was y equal to X β + ε. This is our multiple linear regression model, where y is a n x 1 vector of observation on study variable, X is a n x k matrix of observations on k independent or explanatory variables, β is a k x 1 vector of regression parameters and ε is an n x 1 vector of random error components.

And based on that we had employed the principle of least square and we had obtained an estimator of $\hat{\beta}$ which was indicated by b and this was $(X'X)^{-1}X'y$ and b is called as ordinary least square estimator OLSE, right. So, now, we are going to state the properties of this OLSE. So, this OLSE is an unbiased estimator of β.

So, the E(b) = β. So, some of these properties I will try to prove after stating it, but at this moment you can just take them as a result. And, if you recall in the case of simple linear regression model also, we had proved this result. If you recall in the case of simple linear regression model, we had obtained the variances of the least square estimators of slope parameter and intercept term estimators, and we had obtained their covariance's also.

So, we had obtained the variances of b_0 and b_1 and we also obtain the covariance between b 0 and b 1. Now in the this case, you have here b is equal to b_1 , b_2 ,…, b_k there are k parameters. So, for that we try to obtain the covariance matrix. So, which is defined here as a E(b - β)(b - β)'and this is actually going to be a matrix here like this, where the

diagonal elements are going to indicate the variances and off diagonal elements are going to indicate the covariances, right.

So, if I try to say suppose, I have here b is equal to b_1, b_2, then the covariance matrix of b, if I write try to write covariance matrix of b this is going to be something like this. On the diagonal elements you will have variance of b_1 and variance of b_2 and off diagonal elements will be covariance between b_1 and b_2 and here also covariance between b_2 and b 1. So, this is how we obtain the covariance matrix of the; of an estimator.

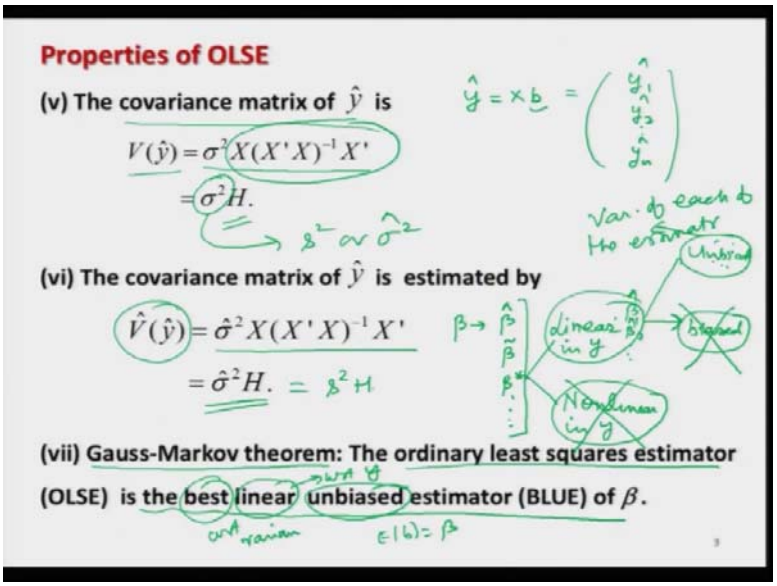
So, the covariance matrix of ordinary least square estimator is given by $\sigma^2 (X'X)^{-1}$. So, here also you can see that unless and until you make an assumption that rank of X is a full column rank, you cannot obtain this unique inverse and hence, you cannot obtain this variance covariance matrix, right.

So, after this we try to find an unbiased estimator of σ^2 . This is obtained by this expression and let $\hat{\sigma}^2$ hat indicates the estimator of σ which is given by 1 upon n - k times y - X b'y - X b. And if you try to see, this term is simply your here y - X b is e. So, this become e', e divided by n - k. Very simple to remember, right.

So this will give you the estimator of σ^2 and we are indicating this estimator of σ^2 by small s^2 . And now, if you try to lo, ok at this expression, try to see my pen in red colour, this expression for covariance matrix of b, right. Here we have got the term σ^2 . So, now, if you want to find out the value of covariance matrix of b on the basis of given sample of data, this σ^2 is unknown, so we cannot find it out. So, this creates a trouble.

So, now, we have here a solution. The solution is this, now you have estimated σ^2 by $\hat{\sigma}^2$ here, you try to put it here, right. So, once I try to do it, I can obtain the an estimate of the covariance matrix of b like this, just by replacing σ^2 by $\hat{\sigma}^2$ here times X'X whole inverse and this expression will give me an estimate of the covariance matrix of ordinary least square estimator. So, this is $\widehat{Var}(b)$. So, hat is indicating that this is an estimator, ok.

(Refer Slide Time: 09:38)



Now, I come to another aspect. Up to now we were considering the properties of ordinary least square estimator b. Now, you have obtained the fitted values, $\hat{y}$ = Xb. So, you can now think, $\hat{y}$ is the value which is of very, is very important for us, because if this value is close to y, then I can say my model is good and if this value is very far away from the model then I will certainly say that my model is not good.

So, definitely this also depends on b, b is a statistics a random variable. So, definitely you would like to study the properties of $\hat{y}$ also. For example, if this is if this has got a very high variance, possibly you can say that the model is not well fitted model is not good. So, that is why I am trying to write down these two expressions.

So, first I try to find out the covariance matrix of $\hat{y}$. $\hat{y}$ because if you remember, this will also be a vector like $\hat{y}_1$, . So, this will also be a so it is variance it could be in the form of a covariance matrix. So, this $V(\hat{y}) = \sigma^2 X(X'X)^{-1}X'$.

And you can recall that this matrix here $X(X'X)^{-1}X'$ is your H hat matrix which we have discussed earlier, right. Now again, we have the same issue that this variance of $\hat{y}$ depends on the covariance matrix of $\hat{y}$ depends on σ^2 which is unknown to us. So, we try to estimate it by $\hat{\sigma}^2$ and we try to replace it by here say s^2 or $\hat{\sigma}^2$.

So, then I can estimate the covariance matrix of $\hat{y}$ like this, which is $\hat{\sigma}^2 X(X'X)^{-1}X'$ and which is same as $\hat{\sigma}^2 H$. So, briefly I can write it down here say $s^2 H$, right. So, this will also help us in getting the further details of the model.

Now, next we have a theorem which is very important for us, which is called as Gauss Markov theorem. So, this theorem help us in taking a final call that the ordinary least square estimator which we have obtained to get the values of the unknown parameters on the basis of given sample of data are they good or bad. So, this theorem states that the ordinary least square estimator of β, what you have obtained here is the best linear and biased estimator of β. What does this mean?

Suppose, if I say suppose there is an unknown parameter β. And this can be suppose estimated by say some estimator $\hat{\beta}$. Now, somebody suggest a different estimation procedure and the value of β or the estimator of β is obtained here as a $\tilde{\beta}$. Similarly, somebody suggest another estimator and this estimator is a indicated by β* and so on.

Suppose, there are so many estimator for estimating the same parameter, that can happen. Somebody may like to use arithmetic mean to estimate the population means, somebody may use median or mode or harmonic mean, geometric mean to estimate the population mean, right. So, now from these possible estimator we can classify them into two groups those estimators which are linear and those estimator which are non-linear.

Remember one thing; I am not talking of the linear model or non-linear model. I am talking of the linearity of estimator. So, this means; linear in y and non-linear in y. If you remember, we had expressed the estimator b_1 in case of simple linear regression model as a $\sum k_i y_i$. So, it was a linear function of y.

Now, some of the estimators out of this possible estimator will come here and some of the estimator will come in this group. Now, try to ignore this group. Now, there will be suppose some estimator $\hat{\beta}_1, \hat{\beta}_2, \ldots$ and so on suppose, there are couple of estimator which are all are linear in y.

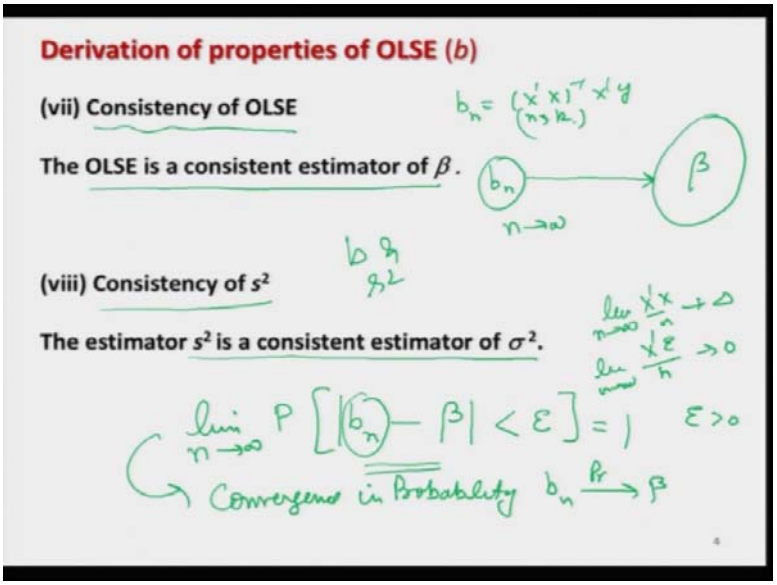
Now, I try to impose here one more condition. I try to divide them into two groups; one group here is something like which is consisting of those estimator which are unbiased,

and then means another groups those estimator which are bias for β. Now, ignore those bias estimator and try to consider all those estimator which are here unbiased.

And then, try to find out the variance of each of the estimator, here in this group. Now you will see, out of all possible estimator whatever you have considered here, this ordinary least square estimator will have the minimum variance among all. So, that is why this is called as best linear unbiased estimator. Best is with respect to variance and linear is linear with respect to y and unbiased is that expected value of b is equal to β.

So, that is why this is a very good estimator. And, this Gauss Markov theorem certifies it that the values of β which you are trying to obtain from b, they will they give you a good value, right.

(Refer Slide Time: 16:00)



Now, after this, I try to address on another property which is consistency. Consistency, I am not considering here the details, but in general I can say here, that you can see here b $= (X'X)^{-1}X'y$.

And, this depends on here n so, you can see here, this is a function of n, k so, b is a function of here n, that is the sample size. And what are you trying to do? There is some unknown population here which has got the parameter vector β. And you are trying to estimate this parameter by here b n.

So, what I am trying to say here that, as n is going to infinity this b n will be converging to a towards β. So, the in simple language I can say, if you try to increase the sample size, then this b n will be converging to β. And, what we try to say here, if I try to translate it into a statistics part means, I can write you know in a much simpler way, that let me try to write down here, try to see the steps of my pen.

So, I have here a parameter β which I am trying to estimate by estimator b_n so, I want to see what is the difference between b_n and β. So, we expect that this difference should be as small as possible if my model is good. So, I try to say here, this is suppose less than ε, where ε is some positive quantity, small quantity.

Now, I am trying to say this difference b_n - β can be positive or negative, but it does not make any difference for us, so we try to take here its absolute value. Now, this is my event, that I want that b_n and β should be close enough. That means, I have to put some condition that what happens such that b_n is converging to β.

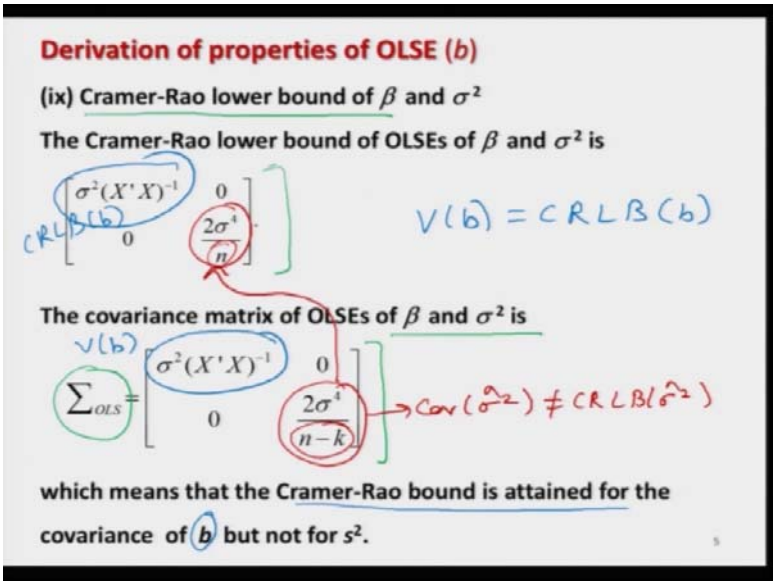
So what I say here, that let me try to compute the probability of such an event, that the difference between b_n and β is as small as possible. And I am trying to say here, if you try to increase the sample size n and I try to compute this probability as limit n tends to infinity; that means, sample size is becoming larger and larger, then what do you expect? This probability should be equal to 1.

That means the difference between b_n and β becomes 0, right. If this is happening then you would say that b_n is a consistent estimator of β. So, in this case I can just show you here, means I will not prove it, but I can inform you that both the ordinary least square estimator of β as well as ordinary least square estimator of σ^2 both means, b and s^2 both are consistent estimator. b is a consistent estimator of β and s^2 is a consistent estimator of σ^2, right.

So, this is actually called as convergence in probability. And this is indicated as b_n tending to is tending to β in probability. And if you remember, when we had discussed the basic assumptions of multiple linear regression model, then the last assumptions that limit (X'X)/n goes to some positive definite matrix delta as n goes to infinity and this vector X'ε upon n as limit n goes to infinity goes to 0.

These were the two assumptions we had made, and these two assumptions are made when I try to establish the consistency property of b and s^2, ok.

(Refer Slide Time: 20:34)



So this is how I am I want to go. Similarly, there is another property in this statistics which is called as Cramer-Rao lower bound. So, Cramer-Rao lower bound gives us a the value of the minimum variance which an estimator can achieve. That means, the variance cannot go beyond that limit.

Well, there is a whole theorem. I am not explaining here the entire theorem, but if you try to use that theorem and if you try to find out the Cramer-Rao lower bound for the ordinary least square estimators of β and σ^2, this will come out to be here like this. And, means if you try to find out the covariance matrix of OLSEs of β and σ^2 this will come out to be here like this which is indicated as σ OLS, right.

So, you can see here that this value, I am trying to write down here in blue colour this and here this they are the same. So, that means, b is attaining the Cramer-Rao lower bound this is the covariance matrix of b. So, this is and this is the Cramer-Rao lower bound of b, both are actually same. So, you can see here variance covariance matrix of b and Cramer-Rao lower bound CRLB of b they are the same.

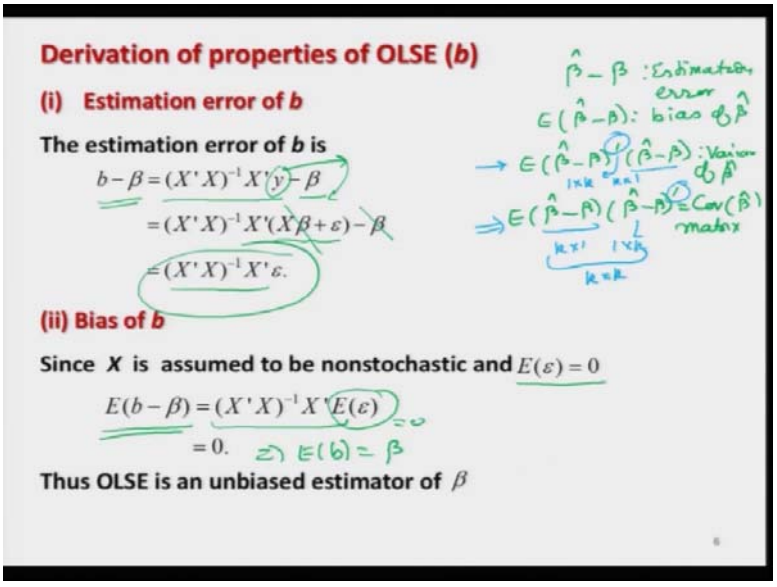
So I can say that that, b attains the Cramer-Rao lower bound or the Cramer-Rao lower bound is attained for the covariance of ordinary least square estimator b. So, you have a

reason that b is going to give you a good value. But now, if you try to see, the covariance matrix of $\hat{\sigma}^2$, this is something like $\hat{\sigma}^2$, right. This is here $\frac{2\sigma^4}{n-k}$ and the Cramer-Rao lower bound for this σ s^2 hat is 2 σ raise power of 4 divided by n.

You can see that this covariance matrix of $\hat{\sigma}^2$ is not equal to the Cramer-Rao lower bound of say $\hat{\sigma}^2$. So, s^2 does not attain the Cramer-Rao lower bound. So, that means, but you can see here that if n is becoming large then there is not much difference. If you try to see, this is here n and this is here n - k. But yeah, theoretically the it is not attaining, that may reached reach as close as possible. So, that is why we always say that it is a better option if you have a larger data set.

Now, if you ask me that what is large and what is the small? That is a very difficult question. And that is that depends on the experiment that, what should be the optimum sample size so that you get a good outcome, right, ok.

(Refer Slide Time: 23:36)



After this, after stating all these properties, let me try to give you a quick proof means the basic steps how one can obtain the proof. So, this will help you if you want to study further, then you can use a book and you will very easily understand it. So, whenever we are trying to prove that either that any estimator is biased or unbiased, what we try to do?

Suppose, we have a parameter β which is estimated by $\hat{\beta}$, so we try to find out the difference.

This is called as estimation error. And then what we try to do? We try to take its expectation. So, this will give us an idea about the bias of $\hat{\beta}$. And, in case if you try to take the expectation of $\hat{\beta}$ - $\hat{\beta}$' $\hat{\beta}$ - $\hat{\beta}$ then this will give you an idea about the variance of β hat.

And if you try to consider E($\hat{\beta}$ - β)E($\hat{\beta}$ - β) then this gives you an idea about the covariance of $\hat{\beta}$. This is the covariance matrix. Yeah, sometimes students get confused whether this prime should be on the second bracket or on the left bracket, right. So, I can give you a very simple trick.

You can just follow this rule, variance is a scalar quantity and covariance matrix is a matrix quantity. So, once you are once you are trying to take here the expectation sign here, you can see here $\hat{\beta}$ is a vector of order k by 1 and here is the transpose so this vector will become 1 by k and the second vector is k by 1. So, the entire product will be of order 1 by 1, so that is a scalar.

So, when you want to find out the variance then you try to put the transpose in the first bracket on the first bracket. And, once you are trying to consider this second quantity, so you can see here this is of order k by 1 and this is of order 1 by k, so this is going to be of order k by k so, this is going to be going to give you a matrix. So, when you want to find out the covariance matrix, then you try to put the transpose on the second bracket.

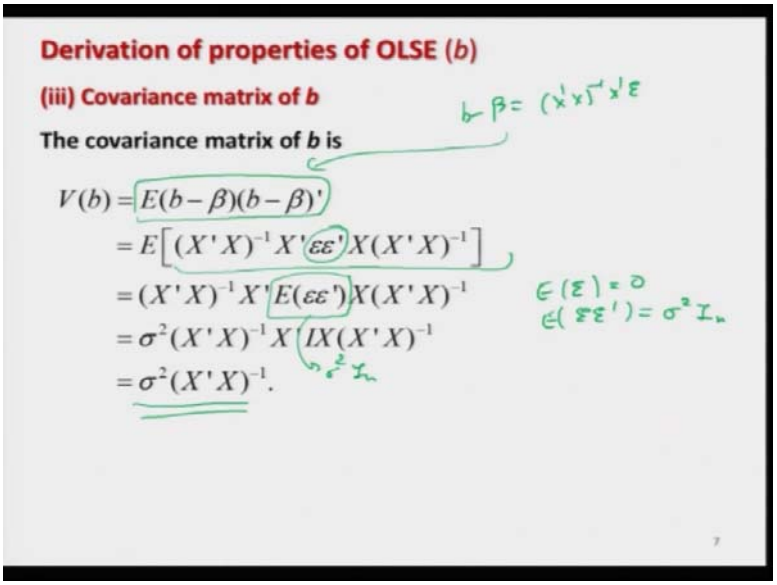
I think there cannot be a better way to remember all these things, right, ok. So, first I try to find out here the estimation error of b so, this is b - β. So, this is the expression for b - β. And here, in place of here y, I try to substitute y = X β + ε. And, if you try to open it, the first term will get cancel out and you will get here $(X'X)^{-1}X'\varepsilon$. So, this is your estimation error.

Now, since you have assumed that X is a non stochastic, this is constant and you also have assumed that expected value of ε is 0, so if you try to take the expectation of this

quantity, you get here expected value of b - β which can be written as this factor into expected value of ε which is 0.

So, that means, expected value of b - β is 0, this implies E(b)=β and hence, ordinary least square estimator is an unbiased estimator of unknown parameter vector β.

(Refer Slide Time: 26:57)

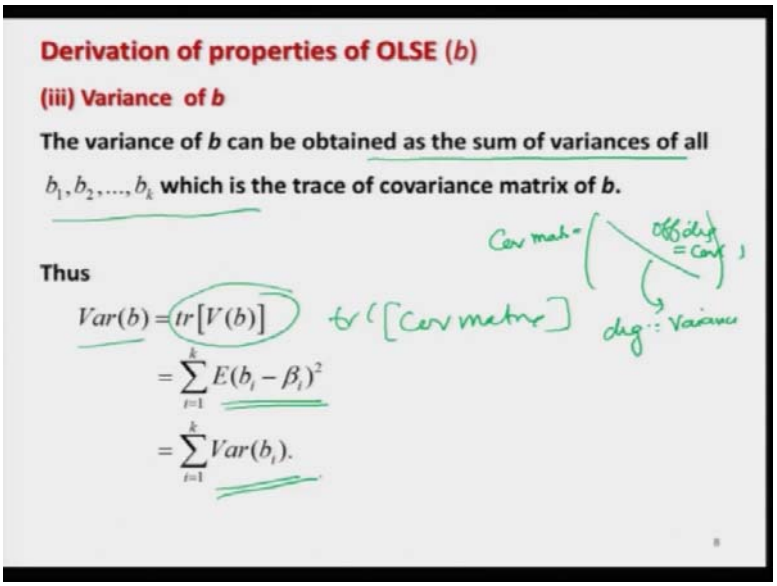


Now, in case if you want to find out the covariance matrix, so as I told you, you simply have to find out the expected of the E(b - β)(b - β)'. So, now, you already have obtained this b - β = $(X'X)^{-1}X'\varepsilon$ so, you simply have to substitute these values over here. So, this will come out to be here like this, εε'in the middle part. Simply try to substitute these values over here, right.

Now, I can take this expectation sign inside the bracket. So, X is a non stochastic matrix. So, the expectation sign inside will be on the random part which is εε'and you already have assumed actually that E(ε) = 0 and E(ε ε')= $\sigma^2 I_n$, right.

So, I can just substitute here this quantity to be here as σ^2 here I_n and you can see here, that once you do it, you will get here $\sigma^2 (X'X)^{-1}$. So, this is the covariance matrix of ordinary least square estimator of β, which is not difficult to obtain. You just need to know the basic matrix theory.

(Refer Slide Time: 28:14)

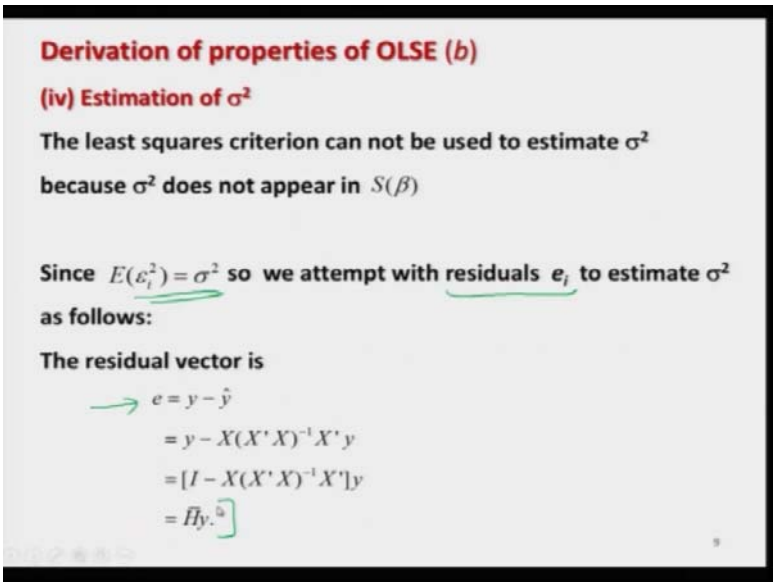


Now, there is a concept that when you are trying to find out the covariance matrix, there are sometimes people are interested in finding out the variance. So, finding out the variance in case of a vector has a little bit different definition. So, suppose you have got here a covariance matrix, then you know that the diagonal elements they are the variances and of diagonal elements, they are the covariances, that we know.

So, now, if you want to find out the variance of b then this is defined as sum of all the variances, sum of all the variances of b_1, b_2, …, b_k,; that means variance of b_1, + variance of b_2 up to variance of b_k,. So, that is equivalent, that if you have obtained the covariance matrix then you can simply take the trace of the covariance matrix, right.

So, the variance of b can be defined as trace of covariance matrix, simply trace of covariance matrix. And this quantity is something like this, $\sum_{i=1}^{k} E(b_i - \beta_i)^2$, and this is nothing but the $\sum_{i=1}^{k} Var(b_i)$, as simple as that, right.

(Refer Slide Time: 29:40)

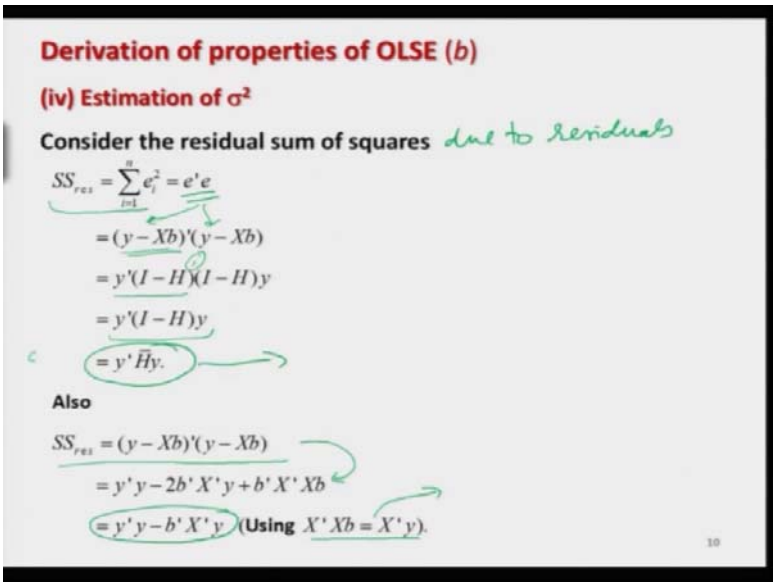


Now, after this we come to the aspect the that, how are you going to estimate the variance σ^2 ? This estimation process is similar to the process what we have learnt in the case of simple linear regression model, right. Because you remember that we started with the estimator by guess and then we converted it converted it into an unbiased estimator.

So, what we try to do here, that we try to consider the sum of square due to residual and then we try to take its expectation and try to obtain an unbiased estimator of σ^2 , right. So, as such it is difficult to it is not possible to obtain the value of σ^2 as we have found the value of b from the by the least square method by minimizing the function as β. Whereas, when we try to do the maximum likelihood estimation you will see it is possible, right, ok.

So, now we try to attempt it. Because, we know that $E(\varepsilon_i^2) = \sigma^2$ for all i. So, what I can do that I can use the residuals to estimate the variant σ^2 , how? If you remember, that we had obtained the residual say (y - $\hat{y}$) and which was expressed that $\hat{y} = \bar{H}$ y.

(Refer Slide Time: 31:05)



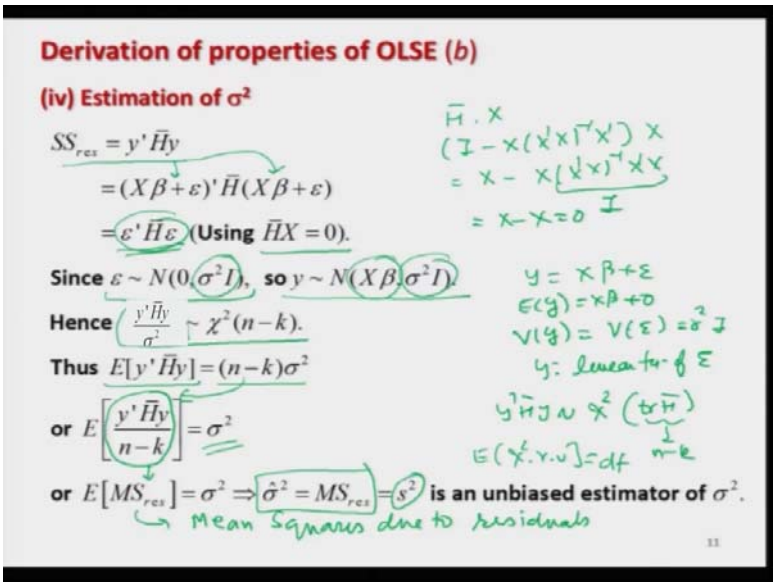
So, I tried to use it here and I tried to use the same concept that I had developed in case of simple linear regression model. I tried to consider here a term SS $_{res}$ that is sum of square due to residual, sum of squares due to I can complete this due to residuals, ok. And this was defined as $\sum_{i=1}^{n} e_i^2$, which can be written in terms of vectors as e'e and e here you have obtained by here (y - X β)'(y - X b).

And which can be using the earlier result what we have just obtained, y - Xb can be written as $y'(I-H)(I-H)y$, but since I - H is a symmetric matrix, so this'will not make any difference. So, I can write down here is and this i - H is an idempotent matrix. So, I can write down finally, as y'(i – H)y and I - H is your $\bar{H}$. So, this sum of square due to residual can be written as y' $\bar{H}$ y.

And, just for your information there are some other forms also, for example, if you try to write down SS res like this, and if you try to open it and when you will get this expression also y'y - b'X'y. And this expression has been obtained using the normal equation X'X b = X'y.

So, this is just for your information that sometime if you use it or sometime in the book it is given, so you must know that this is an alternative form of sum of square due to residuals.

(Refer Slide Time: 32:41)



Now, I can write down sum of square due to residual as a y' $\bar{H}$ y. Now, this y = X β + ε, you can say, right. And if you see here, I can show you here this is the result which I have used here that SS $_{res}$ equal to this, right. So, once I write down here y = X β + ε, then we know one thing here more, that if you try to write down here $\bar{H}$ into X you can see here (I - X (X'X)$^{-1}$ X') X.

So, if you try to open it this will be (I - X (X'X)$^{-1}$ X'), so this becomes here I and this becomes here X - X which becomes here 0. So, this $\bar{H}$ X is equal to 0 and using this property I can write down that this that sum of square due to residual can be also written as ε' $\bar{H}$ ε.

And now we need to use some distributional properties. So, now we are trying to associate the probability density function. So, we assume that εs are following a multivariate normal distribution with mean vector 0 and covariance matrix σ^2 I.

So, this y will also follow the multivariate normal distribution with mean vector X β and covariance matrix σ^2 I. Why? Because, because this b because y =X β + ε so, you know that expected value of y =X β + 0 and covariance matrix of y will be the covariance matrix of X β which is 0 + covariance matrix of ε which is σ^2 I.

So, and y is a linear function of ε which is normally distributed. So, y will also be a normal with mean vector Xβ and covariance matrix $\sigma^2 I$. And we also know from the properties of quadratic form that y' $\bar{H}$ y upon σ^2 will follow a chi square distribution with n - k. Actually, the rule here is this, y' $\bar{H}$ y will follow a chi square with the degrees of freedom which are trace of $\bar{H}$.

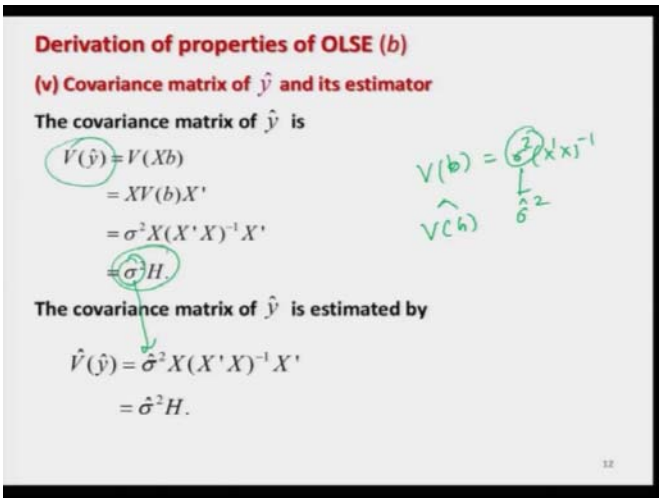
The trace of $\bar{H}$ you already have found to be n - k so, this will follow a chi square distribution with n - k degrees of freedom. And we know that once you are trying to take the expectation of a chi square distributed random variable, then the expectation of this random variable comes out to be same as its degrees of freedom.

So, now, since y' $\bar{H}$ y upon σ^2 is following a chi squared distribution and the degree of freedom is n - k, so expected value of y' $\bar{H}$ y will become n - k upon σ^2 and, if you try to bring this n - k factor here then you get here y' $\bar{H}$ y upon n - k, its expectation is equal to σ^2. And this quantity here, y' $\bar{H}$ y upon n - k this is indicated by MS_{res}, right.

That you will see when we try to conduct the test of hypothesis, then I will introduce this quantity actually this MS_{res} means, mean squares. Because I am going to use it, so I am just introducing it here, so, right. So, this is mean squares due to residuals, right. So, this comes out to be σ^2 and this $\hat{\sigma}^2$ now becomes same as MS res, mean square due to residual which we have denoted by s^2.

And you will see that, when you are trying to implement it in the R software, you will get the value of MS res. So, that is why I am trying to introduce it here, right.

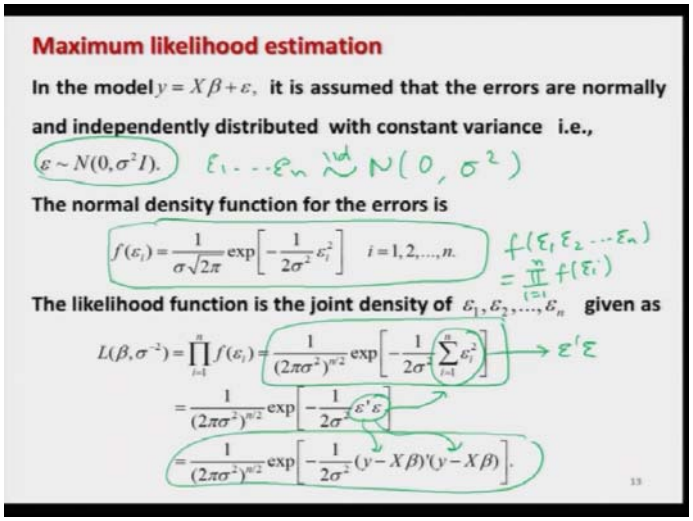
(Refer Slide Time: 36:52)



So, hence, now if you want to estimate the covariance matrix of $\hat{y}$, you can simply replace the quantity σ^2 by $\hat{\sigma}^2$. So, if you remember, we had obtained this covariance matrix of $\hat{y}$ fitted values which was σ^2 say H, now I can replace here σ^2 by; σ^2 by $\hat{\sigma}^2$ and I can obtain the estimate of covariance matrix of $\hat{y}$.

And similarly, I can also find out that variance, if you try to see covariance matrix of b is obtained at $\sigma^2 (X'X)^{-1}$, so I can replace the σ^2 by $\hat{\sigma}^2$ and I can obtain the estimate of covariance matrix of b, right. So, I am just taking this example to show you that what is the use, ok.

(Refer Slide Time: 37:52)



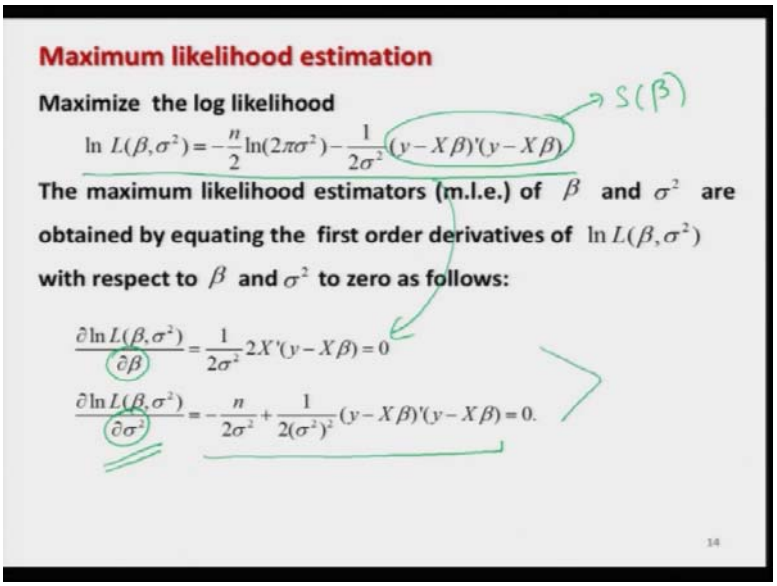
So, now, after this let me give you a quick idea about the maximum likelihood estimation, the theory and philosophy behind the maximum likelihood estimation I already had explained you in case of simple linear regression model. So, I am trying to use the same thing, same steps here also.

So, now we assume that εs are following a multivariate normal distribution with mean vector 0 and covariance matrix σ^2 I. This means, all $\varepsilon_1, \varepsilon_2, \varepsilon_n$, they are IID identically and independently distributed following normal distribution with mean 0 and variance σ^2 . And hence, means I can also find out the distribution of y which I already have done in the earlier case, right.

So, the probability density function of this ε_i is given by this quantity, $f(\varepsilon_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}\varepsilon_i^2\right] \quad i = 1, 2, ..., n$. And, the likelihood function that is obtained as joint density function of $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ and since we are assuming that they are independent, so I can write down them the product of individual $f(\varepsilon_i)$'s so, this is here like this.

So, and if you try to write down here, you will get the likelihood function as here like this, summation ε i square. Now, if you try to see try to lo, ok at this expression summation ε i square, $\sum_{i=1}^{n} \varepsilon_i^2$ can be written as ε'ε, right. So, I try to write down it in the form of ε'ε and now this ε can be replaced by y - X β. So, now this is here the likelihood function. So, basically what we have to do as I discussed earlier, this likelihood function has to be maximized and then we have to find out the value of β and σ^2 .

(Refer Slide Time: 40:01)

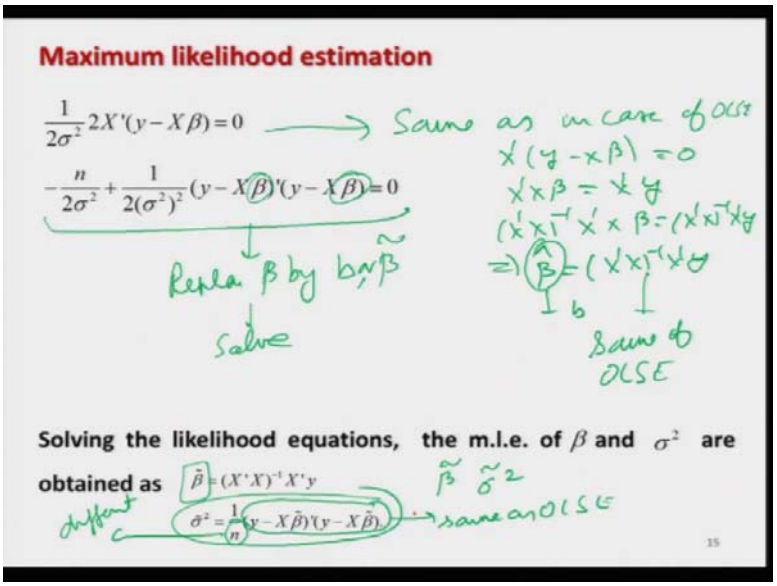


So, instead of maximizing this likelihood function, we maximize the log likelihood, why do we do it? We had explained it in the earlier lecture that both likelihood function and log of likelihood function, they are the monotonic function. So, whatever the solution you are going to obtain by maximizing or minimizing the likelihood function, the same function can be same results can be obtained by log likelihood.

So, I try to take the log on both the side and I try to write down the log likelihood and then after this I simply have to different partially differentiate this log likelihood with respect to β and σ^2 . And, once you try to differentiate it which is now very simple because you can see here, you already had differentiated this function with respect to β in case of $S(\beta)$, when you would found the least square estimator of β you had differentiated the same quantity.

So, this is the same thing. And now, if you try to differentiate it with respect to σ^2 , you get here this equation. Now, if you try to put it then; put these two equation equal to 0 then you get the normal equation. And using those normal equation we can obtain the results.

(Refer Slide Time: 41:21)



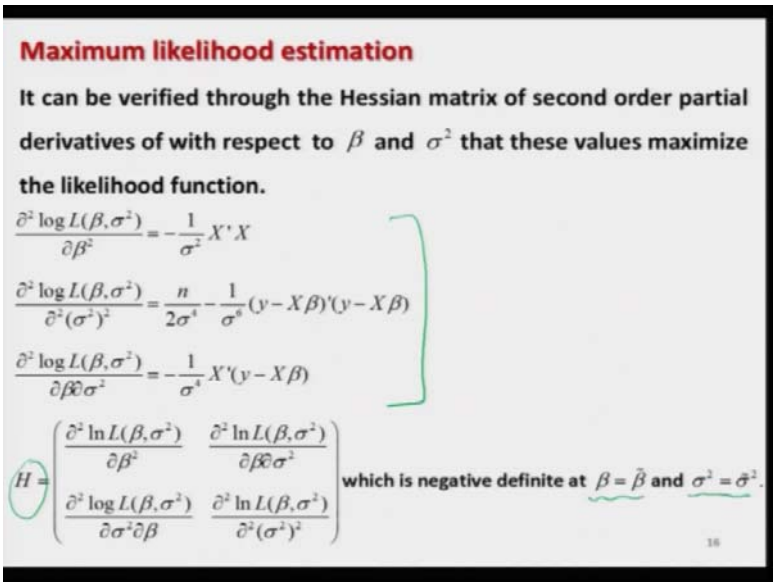
So, this is here the same as in the case of ordinary least square estimator. So, you can see here this is X'y - X β = 0. So, this becomes here say X'X β = X'y. So, on both the sides you try to pre multiply by $(X'X)^{-1}$. So, this comes out to be like this, and this implies that β is equal to $(X'X)^{-1}X'y$ and this is an estimator and we are trying to denote it by b, right.

So, you can see here this is coming out to be same as ordinarily least square estimator, right. Now, you try to take the second normal equation, replace β by say here b or this m l e, let me denote by here say $\tilde{\beta}$, let me try to introduce here the symbol $\tilde{\beta}$ just to indicate the maximum likelihood estimator, although ordinary least square estimator of β and m l e of β that is the maximum likelihood estimators of β they are the same.

So, replace this here b by b or $\tilde{\beta}$ and then solve it. You will get σ^2 is equal to $\frac{1}{n}(y - X\beta)'(y - X\beta)$. The maximum likelihood estimator we are trying to denote by $\tilde{\beta}$ and $\tilde{\sigma}^2$. So, you can see here this is here like this, right. And you can see here, this part in the numerator this is the same as OLSE. The only difference is that this the term in denominator is different.

In the case of m l e the denominator is n, in the case of ordinary least square estimator of σ^2 it was n - k, ok.

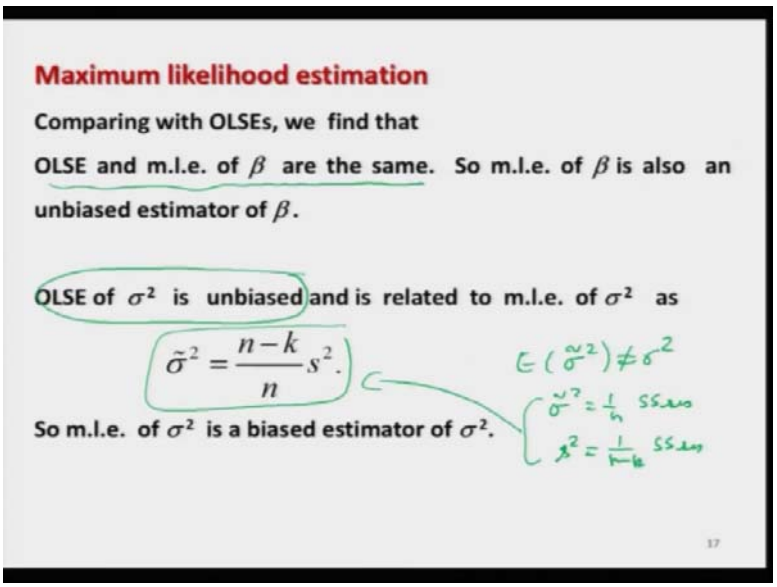
(Refer Slide Time: 43:25)



So, now after this you have obtained these values and you need to verify whether they are going to give you the maximum or minimum values. So, I can obtain the second order derivatives with respect to β σ^2 as well as β and σ^2 and we can obtain the Hessian matrix of second order partial derivatives which is obtained here.

And one can prove that which is the negative definite matrix which indicates that the likelihood function is getting maximized at β equal to $\tilde{\beta}$ and σ^2 equal to $\tilde{\sigma}^2$, right.

(Refer Slide Time: 44:02)



So, now, if you just try to make a quick comparison, you can see here between the ordinary least square estimation and maximum likelihood estimation, you can see that the ordinary least square estimator and maximum likelihood estimator of β are the same, right.

So, all the properties whatever you have obtained for ordinary least square estimator, they will hold for the case of maximum likelihood estimator or β also. Like as this is going to be unbiased estimator is covariance matrix, estimate or covariance matrix they will remain the same, right.

In case of the estimate of σ^2 , the OLSE of σ^2 can be found to be an unbiased estimator, because this is the way it has been obtained, right. But, the m l e of σ^2 which is $\tilde{\sigma}^2$ is not equal to σ^2 . And, and if you try to find out the relationship between the two, it is coming out to be here like this. This is very simple means, right, you can just obtain it here, $\tilde{\sigma}^2$ is 1 over n say, SS res and $s^2 = (1/(n - k))\ SS_{res}$, SS_{res} here is your.

So, and then you can just simply see here SS_{res} is your (y - X b)'(y - X b) and you can obtain this relationship from here, right. So, the m l e of σ^2 is a biased estimator of σ^2 . So, right now we come to an end to this lecture. And yeah, I agree this lecture was little bit longer, but my idea was very simple, I wanted to give you the complete theoretical structure in a single shot you have an option that you can watch this video into two parts or three parts.

But, my idea was that I wanted to make you confirm or confident that whatever values you have obtained they are going to give you a good value. And, you should be confident that these values have been obtained on the basis of a small sample, but they will be valid for entire population, right.

So, my request to you is that you please try to lo, ok into this video and try to find and derive the steps what I have given here, what whatever I have done here, yourself. Unless and until you do this mathematics and algebra yourself with your own hand on your own paper, you will not be feeling confident. So, you do it and I will see you in the next lecture and I will continue with the topics on multiple linear regression model.

So, see you in the next lecture, till then good bye.