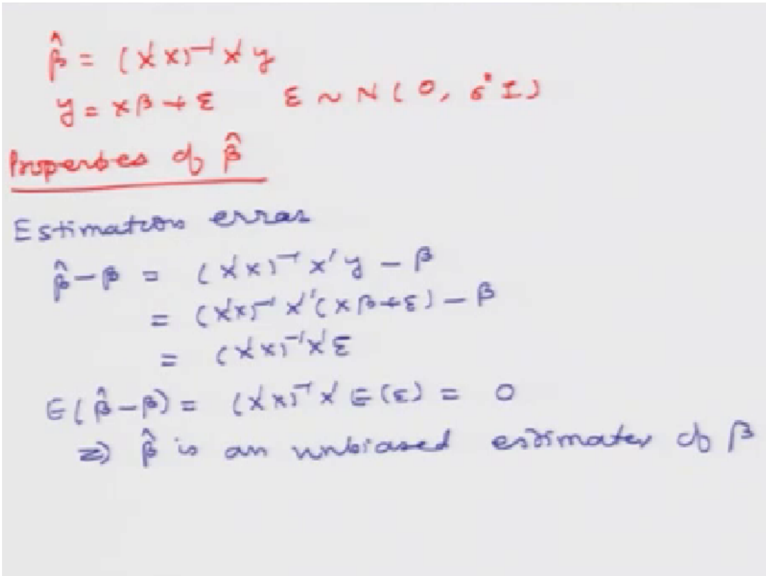
Regression Analysis and Forecasting
Prof. Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology-Kanpur

Lecture -13
Estimation of Model Parameters in Multiple Linear Regression Model (continued)

Welcome to the lecture you may recall that in the earlier lecture we had derived the ordinary least square estimator of regression coefficient vector beta as a beta hat = x transpose x whole inverse x transpose y under the model y = x beta + epsilon assuming that epsilon follows a normal distribution with mean vector0 and covariance matrix sigma square i.

(Refer Slide Time: 00:27)

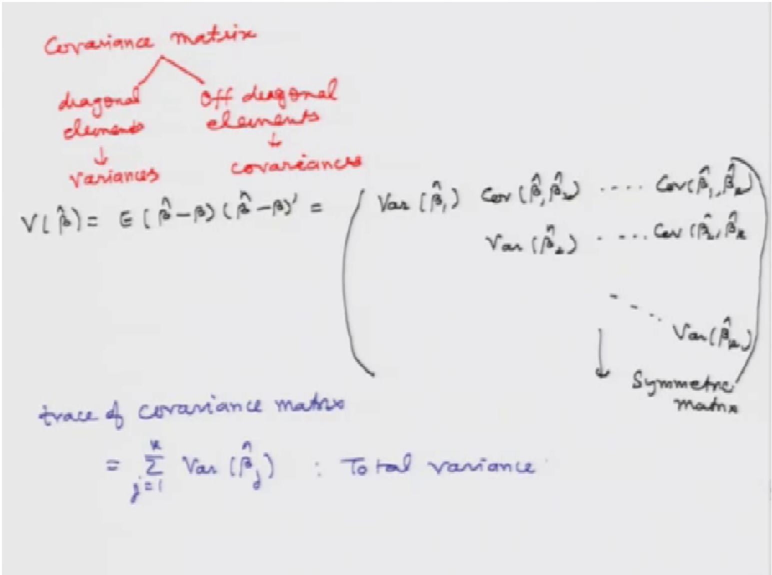


So now we try to investigate the properties of beta hat, first we find out the estimation error which is defined as beta hat - beta so this becomes x transpose x whole inverse x transpose y minus beta now y= x beta + epsilon, so we substitute it here and we obtain like this and this is nothing but x transpose x whole inverse x transpose epsilon because beta cancels out.

So you can see now here that when we try to take the expectations of beta hat - beta then because we have assumed that x is a non-stochastic matrix, so I can write this expectation as follows and since we assumed that expected value epsilon = null vector, so this comes out to be a null vector okay so this clearly implies that beta hat is an unbiased estimator of beta.

And remember in case of simple linear regression model also we had shown that the least square estimator of intercept term beta0 and slope parameter beta1, they were also unbiased, so this is again an extension of that result in the case of multiple linear regression model. Next we try to find out its covariance matrix you may recall that in case of a simple linear regression model we had found the variances of beta0 hat and beta1 hat.

(Refer Slide Time: 02:58)



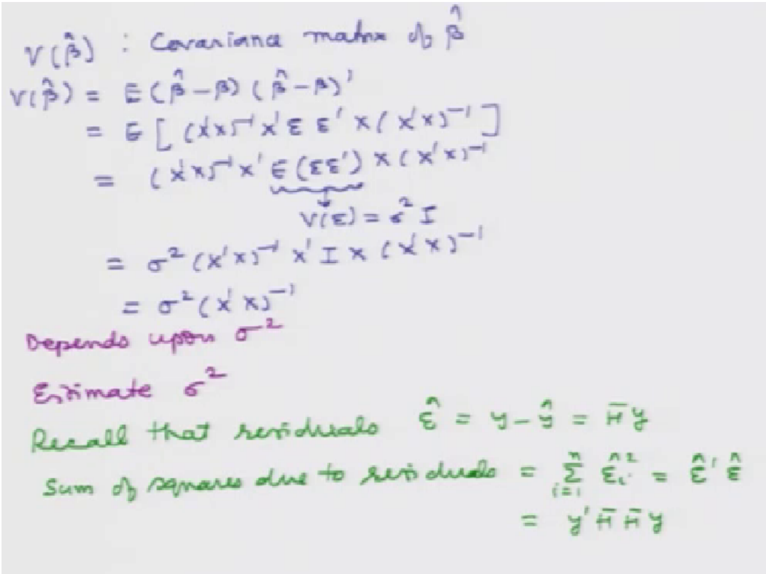
We also found the covariance between beta0 hat and beta1 hat, similarly when we talk of the covariance matrix we are moving one step further and here in this case we want to find the covariance matrix of beta hat, earlier in case of simple linear regression model we had only 2parameters beta0 and beta1 and their estimates as a beta0 hat and beta1 hat but now we have k estimates.

Beta1 hat beta two k hat beta k hat, so when we are trying to talk about the covariance matrix this covariance matrix as two types of elements diagonal elements and off diagonal elements. The diagonal elements give us an idea of the variances and off diagonal elements give us an idea of the covariance. For example when I try to write down the variance covariance matrix of beta hat this is nothing but expected value of beta hat - beta beta hat - beta transpose.

And its structure will be something like this, that on the diagonal elements we will have variance of beta1 hat, variance of beta2 hat and so on, variance of beta k hat and on the off diagonal element for example here this will be a covariance between beta1 hat and beta2 hat up to here say covariance between beta one hat and beta k hat and similarly here covariance between beta2 hat and beta k hat.

So this is a symmetric matrix that we know because covariance between xi and xj is the same as the covariance between xj and xi, and there is another concept here that when we try to find out the trace of covariance matrix, matrix which is equal to something i goes from here one to k, variance of here beta or is should use j beta j. This gives us an idea of the total variance. So this is how we try to interpret the covariance matrix.

(Refer Slide Time: 05:40)



Let us now try to find out the covariance matrix of beta hat, so covariance matrix of beta hat this is expressed as expected value of beta hat - beta, beta hat - beta transpose and you may recall that here in this slide we already had obtain the expression for beta hat - beta. So I simply have to substitute it here and we obtain it like this epsilon, epsilon hat x x transpose x whole inverse.

Now since we have assumed that x is a non-stochastic matrix, so I can take this expectation operator inside and I can write like this. Now we have assumed that expected value of epsilon, epsilon transpose is nothing but the covariance matrix of epsilon which = sigma square i, so I can write down here this thing as sigma square x transpose x whole inverse x transpose ix, x transpose x whole inverse.

So this is nothing but now sigma square x transpose x whole inverse, so this is the covariance matrix of beta hat, the diagonal elements of this matrix will be indicating the variances of beta1 hat, beta2 beta k hat and the off diagonal elements will be indicating the covariance between beta i hat and beta j hat. Now we have another issue here now, this variance of beta hat is based on the population values.

For example it depends on sigma square, and sigma square is a population value, so incase if I need to know this covariance matrix on the bases of a given sample of data we cannot obtain it here and in order to know it we need to estimate sigma square, otherwise this value is an unknown to us. So now in order to estimate the sigma square, we will try to follow the same philosophy that we have developed in the case of simple linear regression model.

That we will start with some of the square due to residuals and from there we will try to construct an estimator for sigma square. So let us try to follow the same philosophy and we recall that the residuals epsilon hat which was defined as y - y hat we had defined as h bar y and based on that if I try to define the sum of squares due to residuals this is nothing but i goes from 1 to n epsilon i hat whole square which is nothing but epsilon hat, transpose epsilon hat.

And we obtain epsilon hat as h bar y, so I can write down here y bar, h bar, h bar y and since h bar is an idempotent matrix, so I can write down here as y bar h bar y.

(Refer Slide Time: 09:40)

Other forms of SS_{res}

$$SS_{res} = (y - X\hat{\beta})'(y - X\hat{\beta})$$
$$= y'y - 2\hat{\beta}'X'y + \hat{\beta}'X'X\hat{\beta}$$

(normal equation $X'X\hat{\beta} = X'y$)

$$SS_{res} = y'y - \hat{\beta}X'y$$

$$SS_{res} = y'\bar{H}y$$
$$= (X\beta + \varepsilon)'\bar{H}(X\beta + \varepsilon)$$
$$= \varepsilon'\bar{H}\varepsilon$$

Using
$$\bar{H}X = [I - X(X'X)^{-1}X']X = X - X(X'X)^{-1}X'X = 0$$

Result: $Z = (Z_1, Z_2, \ldots Z_n) \sim N(0, I)$

$$Z'AZ \sim \chi^2(p)$$

if and only if A is an idempotent matrix of rank p

There are some other forms also of sum of square due residual. For example I can also express sum of square due to residual as a y - x beta hat transpose y - x beta hat and this i if I try to open it this is will become y transpose y - twice beta hat transpose x transpose y plus beta hat transpose x transpose x beta hat. Now if you recall we had obtained the normal equation as x transpose x beta hat = x transpose y.

So if I try to use it then this sum of a square due to residual can be written as y transpose y - beta hat x transpose y. Similarly there is another form SS residual this was obtained earlier as y transpose h bar y and y is in nothing but our x beta + epsilon whole transpose h bar x beta + epsilon, and if I try to open it this will come out to be epsilon transpose h bar epsilon.

And here actually we have used a result that h bar x this is nothing but your i - x x transpose x whole inverse x transpose times x and this is nothing x - x transpose x whole inverse x transpose x, so this will come out to be a null matrix. So we have used this result and using this thing we obtaind here an alternative form of the some of the square due to residuals. Now I am going to write down here result and using this result we are going to find out and estimator of sigma square.

See here z which is something like z1 z2 zn so this is n cross 1vector and suppose this follows a multivariate normal distribution with mean vector0 and covariance matrix here i identity matrix then z transpose Az follows a chi-square distribution with degrees of freedom p if and only if A is an idempotent matrix of rank p. So now using this result we try to obtain the estimator of sigma square.

(Refer Slide Time: 12:47)

$\varepsilon \sim N(0, \sigma^2 I)$

$y = X\beta + \varepsilon$

$E(y) = X\beta$

$V(y) = V(X\beta + \varepsilon) = V(\varepsilon) = \sigma^2 I$

$y \sim N(X\beta, \sigma^2 I)$

$y'\bar{H}y \sim \chi^2(\operatorname{tr} \bar{H}) \qquad (\operatorname{tr} \bar{H} = n-k)$

$= \chi^2(n-k)$

$E(y'\bar{H}y) = (n-k)\sigma^2$

$E\left(\frac{y'\bar{H}y}{n-k}\right) = \sigma^2$

$\Rightarrow \hat{\sigma}^2 = \frac{y'\bar{H}y}{n-k} = \frac{SS_{res}}{n-k} = MS_{res}$ (Mean sum of squares due to residuals)

$\downarrow$

unbiased estimator of σ^2

So we just note down that we have assumed that epsilon is following a normal distribution with mean vector0 and covariance matrix sigma square and y = x beta + epsilon, so expected value of y is same as x beta and covariance matrix of y, this is covariance matrix of x beta + epsilon and this is same as the covariance matrix of epsilon which is sigma square i.

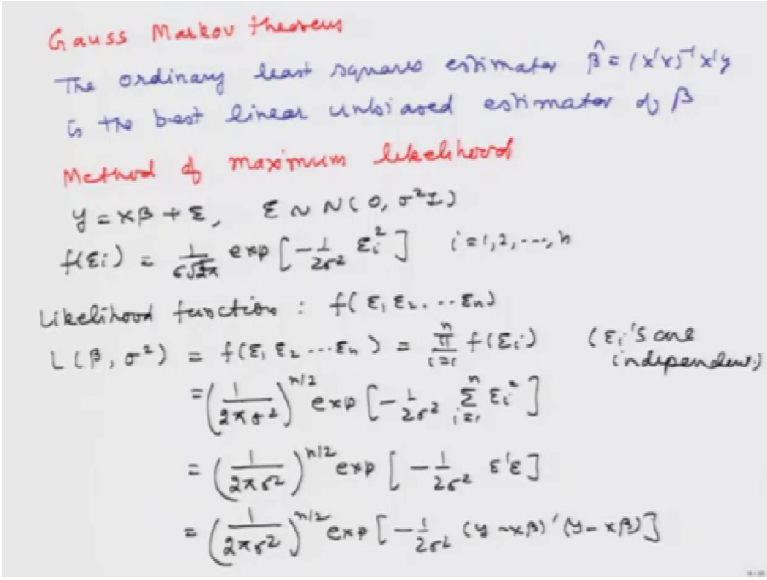
And y is a linear function of epsilon, so I can write down that y is also following a multivariate normal distribution with mean vector x beta and covariance matrix sigma square i. Based on that I can write down that y transpose h bar y will follow a chi-square distribution with degrees of freedom face of h bar, because h bar is an idempotent matrix, so I can use the result and based on that I can write down that the quadratic function y transpose h bar y will follow a chi-square distribution with degrees of freedom which are = trace of h bar.

And we may recall that we had found the trace of h bar= n - k, so this is nothing but your chi-square distribution with n - k degrees of freedom and we also know that if there is some random variable here which is following a chi-square distribution then the expectation of that random variable is same as a degrees of freedom and its variance is same as that twice of degrees of freedom.

So using this result I can write down expected value of y transpose h bar y = n - k sigma square and hence y transpose h bar y over n = k = sigma square and this implies that sigma square hat = y transpose h bar y over n - k this is actually nothing but sum of square due to residual divided by n - k and we define it here as MS res which means mean sum of squares due to residual.

So this turns out to be an unbiased estimator of sigma square, so we have now obtain the least square estimates of the regression coefficient vector beta1 hat beta2 hat beta k hat and we also have obtain now an unbiased estimator of sigma square, so now we have obtained the estimates of all the model parameters

(Refer Slide Time: 16:37)



Now there is another question that how do we ensure that the estimate of regression coefficient vector which we obtain to the principle of least square as beta hat= x transpose x whole inverse x transpose y is it a good estimator, so for that we have a theorem what we call as Gauss Markov Theorem and this theorem states that the ordinary least squares estimator beta hat = x transpose x whole inverse x transpose y is the best linear unbiased estimator of beta.

Well, we are skipping the proof of this theorem, but I would like to explain what is this theorem is trying to say, this theorem is trying to say that if we have a parameter beta, and suppose there are more than stimator available for estimating beta, out of those estimator first we try to identify that which one of them are linear estimator and from the group of or from the class of those linear estimator then we try to indentify that which of the estimators are unbiased estimator of beta.

Now those estimator which are linear and which are unbiased estimator of beta if we try to find out the variances of those estimator then the estimator based on ordinary least square estimator will have the minimum variance, so the ordinary least square estimator are having the minimum variance in the class of linear and unbiased estimator and that is the massage what id given by the Gauss Markov Theorem.

Now after this our next objective is that we demonstrate how to estimate the parameters using the method of maximum likelihood, so method of maximum likelihood, we have used in estimating beta0 and beta1 in case of simple linear regression model all most on the same lines we can demonstrate here that how to estimate the regression coefficient vector and sigma square.

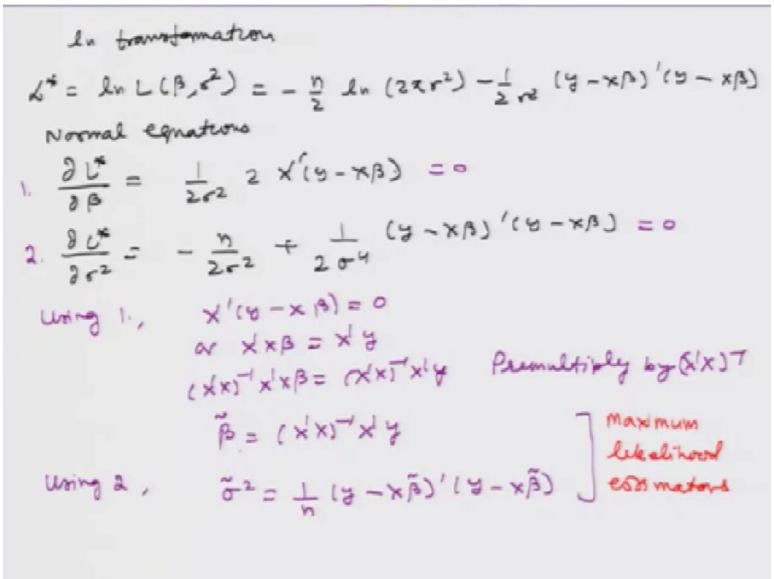
So in the model y = x beta + epsilon we have assume that epsilons are following a multivariate normal distribution with mean vector0 and covariance matrix sigma square i and we know that the probability density function of epsilon i is 1 over sigma root 2 pie exponential of - 1 over 2 sigma square epsilon i square where i goes on here 1, 2 here and n, now I can write down the likelihood function.

We had defined the likelihood function in the case of simple linear regression model as the joint density function of all the random variable, so in our case the likelihood function is nothing but the joint probability density function of epsilon1, epsilon2, epsilon n. So in our case we denote this likelihood function as L, which the function of beta and sigma square which is joint density function x1, epsilon1, epsilon two, epsilon n.

And since we have assumed the independence, so I can write down that this is nothing but product of f of epsilon i's because epsilon i’s are independent. So this becomes here nothing but 1 over 2 pie sigma square rest to the power of here n by 2 exponential of - 1 over 2 sigma square summation i goes on 1, 2 n epsilon i square and the same thing can be written 2 pie, 1 upon 2 pie sigma square power of n by 2 exponential of - 1 over two sigma square epsilon transpose epsilon.

And this can further be expressed as 1 upon 2 pie sigma square is power of n by 2 - 1 over 2 sigma square y - x beta transpose y - x beta. So now since we know the log transformation is monotonic and it is easier to handle the log of L beta square rather than the likelihood function, so we try to take the log transformation and we try to defined here

(Refer Slide Time: 22:04)



say L star which is natural log of L beta sigma square and this is nothing but - 1 - n by 2 log of 2 pie sigma square - 1 over 2 sigma square y - x beta hat y - x beta hat y - x beta. Now I will simply use the principle of maxima and minima to obtain the values of beta and sigma square, such that this likelihood function is maximized.

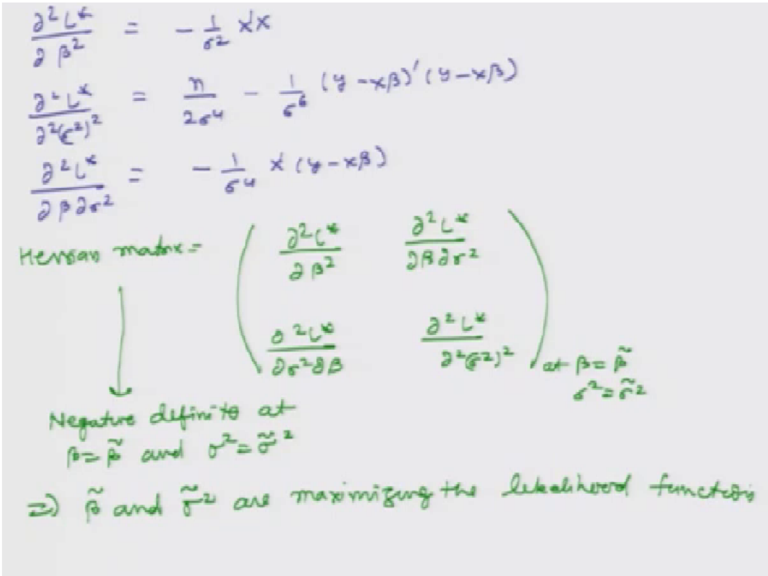
We obtain the normal equations by partially differentiating L star with respective beta and with respective sigma square. We obtained here one over two sigma square twice of x transpose y - x beta and here we obtain - n over sigma square + 1 over 2 sigma the power of here four y - x beta transpose y - x beta and we put them = 0.

So now we have here 2 likelihood equations 1 and 2, so first we try to use here the equation number 1, we obtain here that x transpose y - x beta = a null vector.or this can be written as x transpose x beta = x transpose y and when I try to pre multiply by x transpose x whole inverse we obtain here x transpose x whole inverse x transpose x beta is equal to x transpose x whole inverse x transpose y.

So we obtain here beta = x transpose x whole inverse x transpose y and this we denote as beta delta, and similarly when I try to use the second equation we obtain the value of sigma square = 1 over n y - x beta transpose y minus x beta, So obviously this cannot be obtain because beta is unknown so in this case what we do that we replace beta by beta delta and this gives us a maximum likelihood estimator of sigma square like this.

Well these are the maximum likelihood estimators. We still need to show that the value of beta = beta delta and sigma square equal to sigma square delta they really maximize the likelihood so in order to show whether beta = beta delta and sigma square = sigma square delta are maximizing the likelihood function we try to obtain the second order derivative, second order partial derivative with respect to beta with respect to sigma square and with respect to beta and with respect to sigma square and we obtain this expansion as a - 1 over sigma square x transpose x.

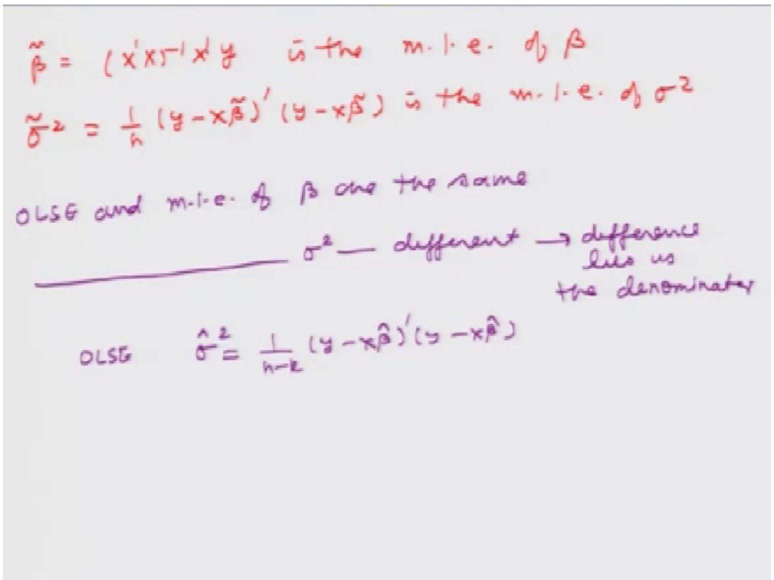
(Refer Slide Time: 24:44)



And this expression is n over 2 sigma is power four over one upon sigma is power of six y minus x beta transpose y - x beta and this is one over sigma the power of here four, x transpose y - x beta and then, I can write down the Hessian matrix, partial derivative of L star with respect to beta square, partial derivative of L star with respect to beta and sigma square and partial derivative of L star with respective sigma square and so on.

Then I try to obtain it's a value at beta = beta delta and sigma square equal to sigma square delta and this Hessian matrix comes out to be negative definite at beta = beta delta and sigma square = sigma square delta, so this implies that beta delta and sigma square delta are maximizing the likelihood function.

(Refer Slide Time: 27:52)



So now I can say here that beta = x transpose x whole inverse x transpose y is the maximum likelihood estimator of beta and sigma square delta which is one over n y - x beta delta transpose y - x beta delta is the MLE maximum likelihood estimator of sigma square. One thing what you have to notice here is that the ordinary least square and maximum likelihood estimator of beta are the same.

Whereas ordinary least square estimated MLE maximum likelihood estimator of sigma square is different, the difference is, difference lies in the denominator. In case of ordinary least square estimator, we had seen that the denominator was one over n - k, y - x beta hat and in this case, and in the case of maximum likelihood estimator, the denominator here is n and that is obvious means the maximum likelihood estimates have establish property that they have the minimum variance.

So whenever we are trying to divide the quantity y - x beta delta transpose y - x beta delta by n we are going to get a lower variance. So now we stop here, and in the next turn we will try to explore some more aspect, till then good bye.