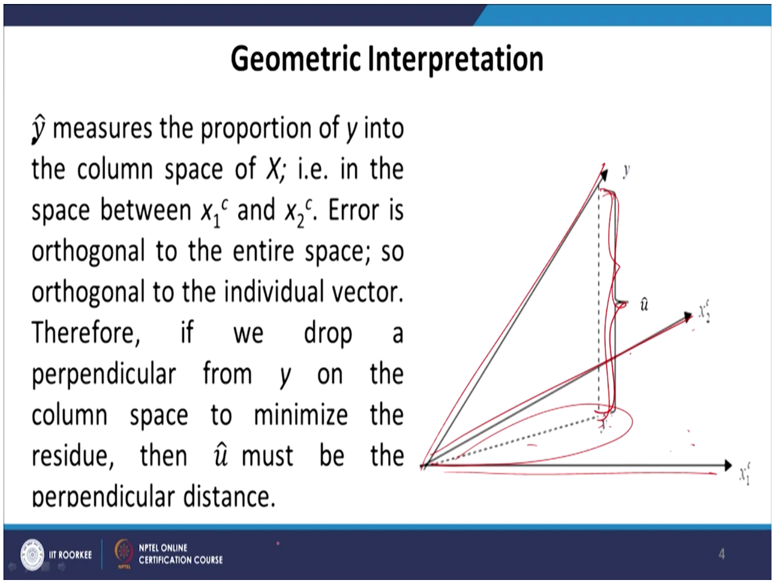
**Econometric Modelling**
**Professor Sujata Kar**
**Department of Management Studies**
**Indian Institute of Technology, Roorkee**
**Lecture – 12**
**Multiple Regression - II**

Hello everyone, this is module twelve of econometric modelling, this also deals with multiple regression analysis. So, in the previous module, I just briefly introduced certain concepts related to multiple regression.

(Refer Slide Time: 00:43)


Specifically, what we talked about was, how do we arrive at or derive the estimated parameters. And then we talked about assumptions and properties, important properties of the estimated parameters. Now, we will give you the geometric interpretation, followed by one important theorem and its application in deriving the residual variance.

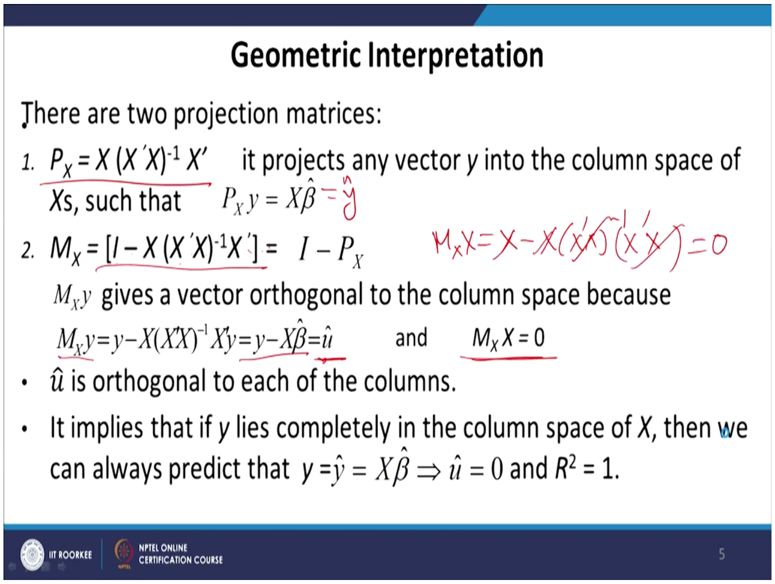
(Refer Slide Time: 01:09)

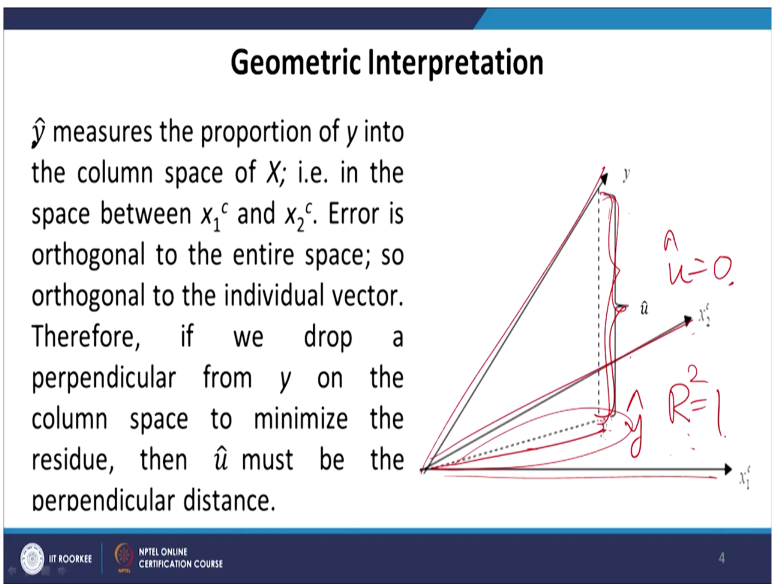



Talking about geometric interpretation, we by now know that y hat measures the proportion of y into the column space of X, that is what percentage of y is explained by the independent variables or maybe, whatever portion the part of y that is being explained by the independent variable. So, these are the column space of the independent variables. Suppose independent variables together explain y. So, this is my y hat, but these are actual observations of y (*see the slide time 1:09).* So, what remains is basically u hat or unobserved or maybe rather the residuals. So, the error or the residuals are orthogonal to the entire space, orthogonal to the individual vectors. Therefore, if we drop a perpendicular from y on the column space to minimize the residue, then u hat must be the perpendicular distance. Because you can see that u hat is perpendicular to the column space. This also implies that u hat is independent of the column space or individual independent variables. So, this is how we interpret the multiple regression analysis. Of course, when you work with n variables, then it is very difficult to explain an n dimension graph or explaining the n variable regression analysis geometrically or graphically.

So, we work with, at the max two variables, and that too excluding the constant term. So, what we here try to do is to minimize this orthogonal distance, while trying to explain y using various combinations of the Xs.

(Refer Slide Time: 03:12)




There are two projection matrices. Now, we introduce these two projection matrices in order to explain certain concepts and also in order to prove certain theories. So, first of all, the theory that will be coming up next will be using these matrices extensively. So, the two main matrices, the first one is a projection matrix. *(Refer to slide time 03:12).* So, you can see that this is the matrix that gives us the value of y hat. Because y hat is the projected values of y by the combinations of Xs.

So, the projection matrix gives us the value of y hat. The way we construct the projection matrix denoted by Px, which is X into X prime X inverse X prime, it projects any vector y into the column space of X's such that Px y is equal to X beta hat. And you know, X beta hat is actually equal to y hat. So, when Px is multiplied by y, then we get the projected values of y, that is y hat. And that is why Px is called the projection matrix.

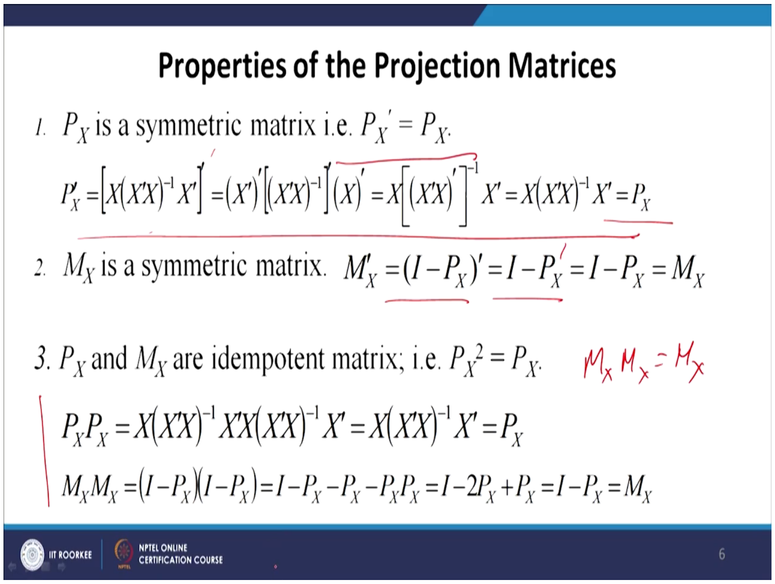
The other matrix is Mx *(Refer to slide time 03:12)*, is actually giving a vector orthogonal to the column space, because when Mx is multiplied by y, then you can see that by putting the values of y or rather by multiplying this expression, which is I minus X into X prime X inverse X prime alternatively Mx is defined as I minus Px. This is a matrix, which gives a vector orthogonal to the column space. So, you can see that this gives us y minus X beta hat, which is equal to E hat.

This is because Mx X equals 0, if you multiply Mx with X, then you will be having X minus X into X prime X inverse X prime X. So, X prime X and X prime X inverse they cancel out, and I am left with X minus X. So, Mx is equaled to 0, Mx X equals to X minus X into X prime X inverse X prime X. These two cancels out, and then these two cancels out. So, that is how 0. So, Mx y gives us u hat *(Refer to slide time 03:12)*.

And since Mx X equals 0, we can say that this gives a vector that is orthogonal to the column space. And u hat are all basically orthogonal to the column space, it implies that if y lies completely in the column space of X, then we can always predict that y equals y hat equals X beta hat, which implies u hat equals to 0 and R square equals 1. That is, if y is completely explained by the column space, then what would happen?

What would happen is that there will be no orthogonal distance. So, if there is no orthogonal distance between y hat and y, this means u hat is 0. There is no residual. It is completely explained by the combination of the independent variables. And as a result of which, the R square as goodness of it measures that we had discussed earlier will be equal to 1. That is, the explained sum of square is exactly equal to the total sum of square, so we have E s s upon T s s equal to 1.

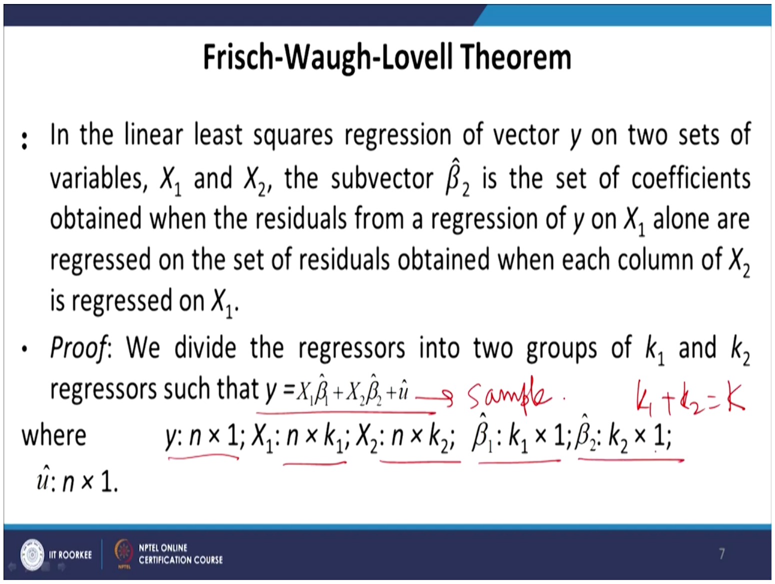
(Refer Slide Time: 06:50)




Now, we talk about certain properties of the projection matrices, because again, that will be useful *(Refer to slide time 06:50)*. Px is a symmetric matrix, which implies that the Px prime is equal to Px. Now, the proofs are also given here, you can see that Px prime is actually a Px the n prime that is the transpose is taken. And if I break the transpose, then we actually arrive at Px again. So, that is why we call Px a symmetric matrix. Similarly, Mx is also a symmetric matrix.

Because simply Mx is I minus Px prime. So, I minus Px prime will also be I minus Px and that is equal to Px. Px and Mx are idempotent matrices, that is if I multiply Px with Px or Px square, then again we arrive at Px. Similarly, Mx into Mx is actually equal to Mx. And the proofs are also given here *(Refer to slide time 06:50)*. So, these are the two very important properties of the projection matrices. First, they are symmetric matrices. Second, they are idempotent matrices.

(Refer Slide Time: 08:01)

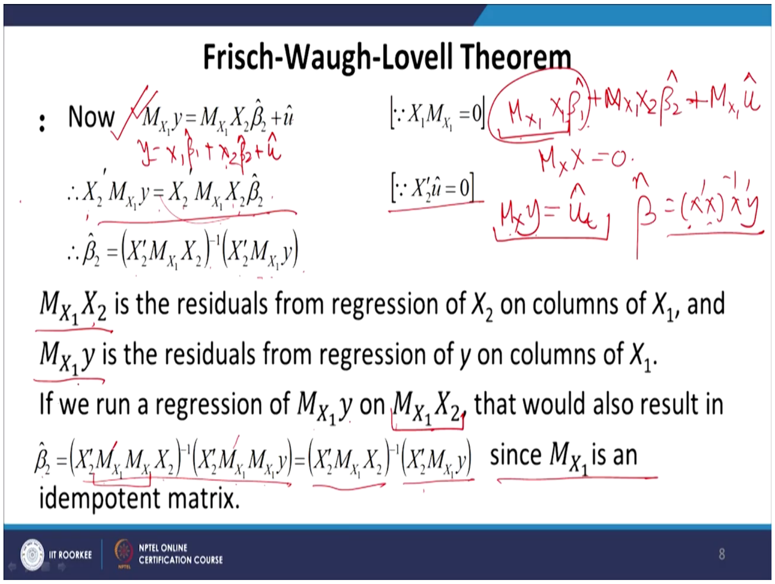



Now, we talk about a theorem which is Frisch Waugh Lovell theorem. In the linear least squares regression of vector y on two sets of variables $X_1$ and $X_2$, the theorem states that the sum vector beta $_2$ hat is the set of coefficients obtained when the residuals from a regression of y on $X_1$ alone are regressed on the set of residuals obtained when each column of $X_2$ is regressed on $X_1$.

So, the theorem states that by regressing y on a set of independent variables and obtaining the residuals again, we regress the rest of the independent variables on the first set of independent variables and retain the residuals. And now, if I regress the first residual on the second set of residuals, then we can derive the parameter estimates of the second set of independent variables. So, the matter appears a little complex, that is actually explained.

And this theorem is also proved in the subsequent slides. So, we divide the regressors into two groups, k1 and k 2. Such that, k1 plus k2 is equal to the total number of independent variables, which is k. So, we write it as y equals $X_1$ beta 1 hat plus $X_2$ beta 2 hat plus u hat. If we had regressed y on $X_1$ and then $X_2$, then we would have arrived at this expression in the sample. Now, they are, here the dimensions of y and $X_1$ $X_2$ and beta 1 beta 2 hat are mentioned. So, y is the n by 1 vector, as usual, $X_1$ is n by k1, $X_2$ is n by k2 beta 1 hat k 1 by 1, beta 2 hat, k 2 by 1 and u hat is again, n by 1.

(Refer Slide Time: 10:15)




Now, the Frisch Waugh Lovell theorem uses, of course, these matrices. So, first of all, let me show you that;

$$M_{X_1} y = M_{X_1} X_2 \hat{\beta}_2 + \hat{u}$$

Why this is so? Because, I have just mentioned that y is actually X1 beta 1 hat plus X2 beta 2 hat plus, u hat *(Refer to slide time 10:15)*. So, given that, y is X1 beta 1 hat plus X2 beta 2 hat plus u hat, Mx 1 y and then replacing the values of y into this expression, I will be having Mx 1 X1 beta 1 hat plus Mx 1 X2 beta 2 hat plus Mx 1 u hat.

And then since Mx 1 X1 equals 0, if you remember the way we define Mx, we prove that Mx X is equal to 0. So, in a similar fashion, we are defining Mx 1. So, Mx 1 X1 will also be 0. So, since Mx 1 X1 is actually equal to 0, I am left with Mx 1 X2 beta 2 hat plus u hat. Now, X2 now I am multiplying this expression with X2 prime. So, multiplying this with X2 prime, I have X2 prime Mx 1 X2 beta 2 hat plus X2 prime u hat. But X2 prime u hat is equaled to 0 *(Refer to slide time 10:15)*. Because, u hat the residuals are independent of all functional forms of X, that is all the independent variables. So, X2 prime u hat will be equal to 0. As a result of which, I have only this expression, If you remember, Mx y was the, was equal to ut hat or u hat, which implies that when we are regressing y on X, and then multiplying Mx with y, then we are, arriving at the residuals obtained from the regression of y on X. So, Mx 1 X2 must be giving us the regression or the residuals from the regression of X2 on X1. So, that is what is mentioned here, that Mx 1 X2 is the residuals from the regression of X2 on the columns of X1 and Mx 1 y in a similar fashion is the residuals from the regression of y on the columns of X1.

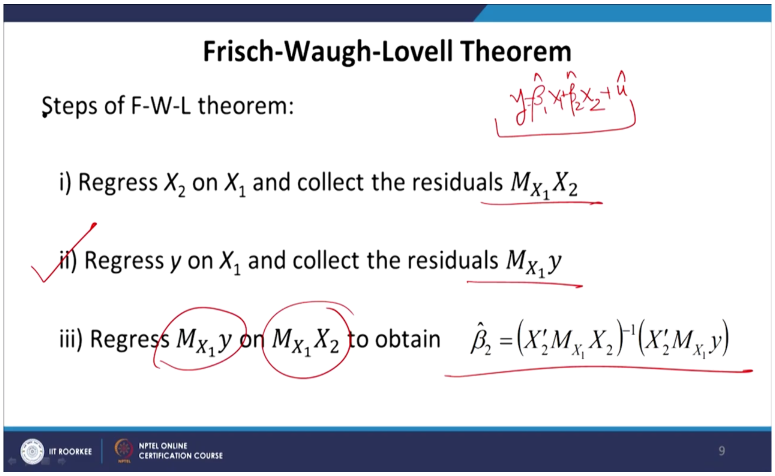
So, if we run a regression of Mx 1 y on Mx 1 X2, that would also result in beta 2 hat equals to

$$\hat{\beta}_2 = \left(X_2' M_{X_1} M_{X_1} X_2\right)^{-1}\left(X_2' M_{X_1} M_{X_1} y\right) = \left(X_2' M_{X_1} X_2\right)^{-1}\left(X_2' M_{X_1} y\right)$$

See now, I draw an analogy with beta hat, which is X prime X inverse X prime y. Now, here my X is Mx 1 X2, that is we are running a regression of Mx 1 y on Mx 1 X2. So, Mx 1 X2 is my X here, so, I will be having Mx 1 X2 prime which is Mx 1 X2 prime. Since Mx 1 prime and Mx are the same, they are symmetric matrices.

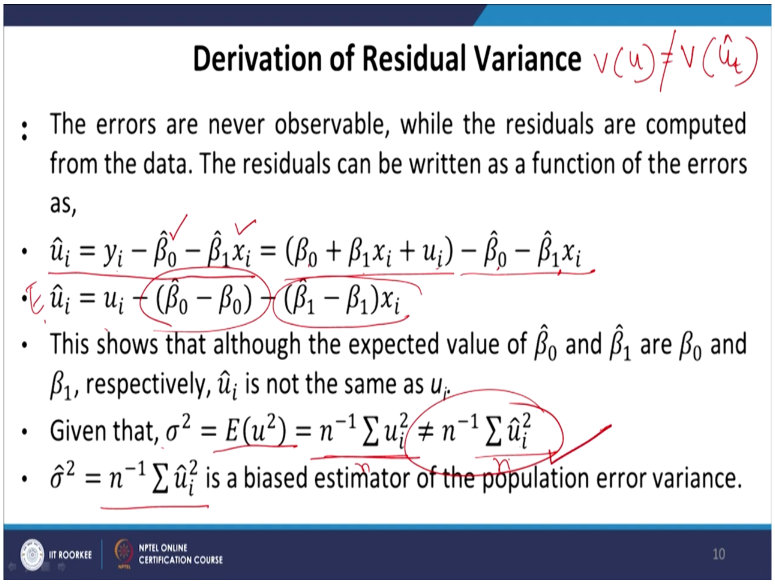
So, that is why mentioning a prime or not mentioning it are the same thing. So, I have X prime which is X2 prime Mx 1 prime X Mx 1 X2 inverse X2 prime Mx 1 prime Mx 1 y. So, that is equivalent to our X prime X inverse X prime y. Since Mx 1 prime Mx 1 equals to Mx 1 Mx 1 is equal to Mx 1. Because they are idempotent matrixes. So, I have X2 prime Mx 1 X2 inverse X2 prime Mx 1 y. Since, Mx 1 is an idempotent matrix. So, this proves that beta 2 hat can be obtained from the regression of the residuals of a regression of X1 on y, on the residuals from a regression of X2 on X1.

(Refer Slide Time: 15:10)




So, what are the steps involved in Frisch Waugh Lovell theorem? First of all, I regress X2 on X 1 and collect the residual Mx 1 X2, then I regress y on X1 and collect the residual Mx 1 y. And after that, we regress Mx 1 y that is, this residual on Mx 1 X2 that is these residuals to obtain beta 2 hat. We could have obtained beta 2 hat directly by regressing y on X1 and X2 separately. But Frisch Waugh Lovell theorem says that, if I follow these procedures, then we are going to exactly arrive at the same set of estimates that would have been given by this kind of regression. So, this is the theorem. Now, it has certainly its applications and usefulness, that is why the theorem has been developed. So, the theoretical application will be discussed very soon. While we actually try to derive the residual variance. Now, first of all, why residual variance would be different from the population error variance.

(Refer Slide Time: 16:25)




So, what I am trying to claim here is that the population error variance is actually not equal to the variance of the residuals. The errors are never observable, while the residuals are computed from data. The residuals can be written as a function of the errors as u i hat equals y i minus beta 0 hat minus beta 1 hat xi *(Refer slide time 16:25*). So, now look at here, that I am actually working with a simple regression model that is, we, I have a constant term and I have only one variable here.

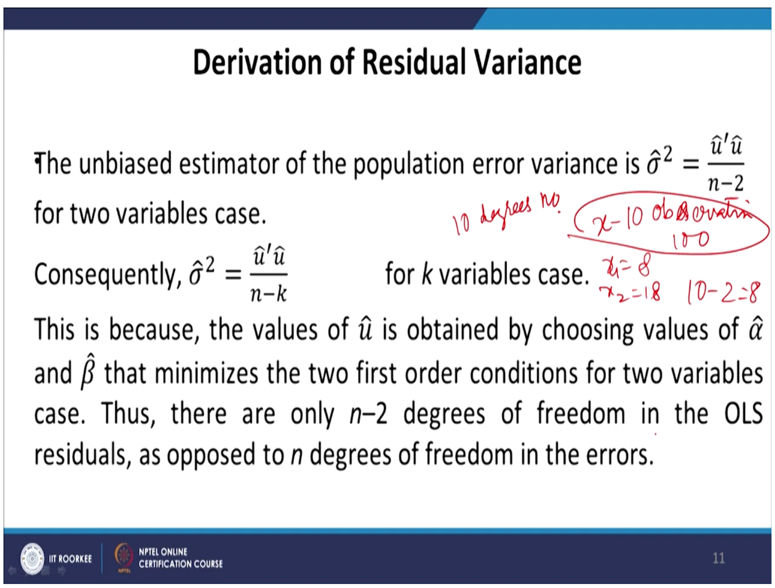
Now, yi can be further replaced with its population expression or model. So, beta naught plus beta 1 xi plus ui minus beta naught hat minus beta 1 hat xi. So, ui hat is actually ui coming here, and then I am collecting these terms beta naught and beta naught hat. So, this is collected here and beta 1 xy xi and beta 1 xi hat, are collected here *(Refer slide time 16:25*).

So, this shows that although the expected value of beta naught hat and beta 1 hat are supposed to be beta naught and beta 1 respectively, ut hat is not the same as ut. So, ui hat is not the same as ui *(Refer slide time 16:25*). See, I do not take an expected value here. So, I am actually not, saying that these are the same thing. So, individually they can be different, the population error can be very different from the sample residual.

So, given that sigma square is the population variance, that is this is equal to the expected value of u square. Alternatively, it can be written as summation ui square divided by n or n raised to the power minus 1 summation ui square *(Refer slide time 16:25*). But this is not equal to summation ui hat square divided by n. Since, ui hat and ui, they are not the same thing. So, sigma square equals n raised to the power minus 1 summation ui hat square is a biased estimator of the population error variance.

Now, I further discussed that why it is a biased estimator, first of all, we have shown that they are not the same thing. So, by dividing them by n we certainly do not get the same thing. That itself indicates that this is not the right estimator of this.

(Refer Slide Time: 19:16)




The unbiased estimator of the population error variance is actually, u hat prime u hat divided by n minus 2 for 2 variable cases and consequently, this is u hat prime minus u hat divided by n minus k for k variable case, or in case I have k plus 1 variable then this will be n minus k minus 1. This is because, the value of u hat is obtained by choosing values of the alpha hat and beta hat that minimizes the first, the two first-order conditions for two variables case.

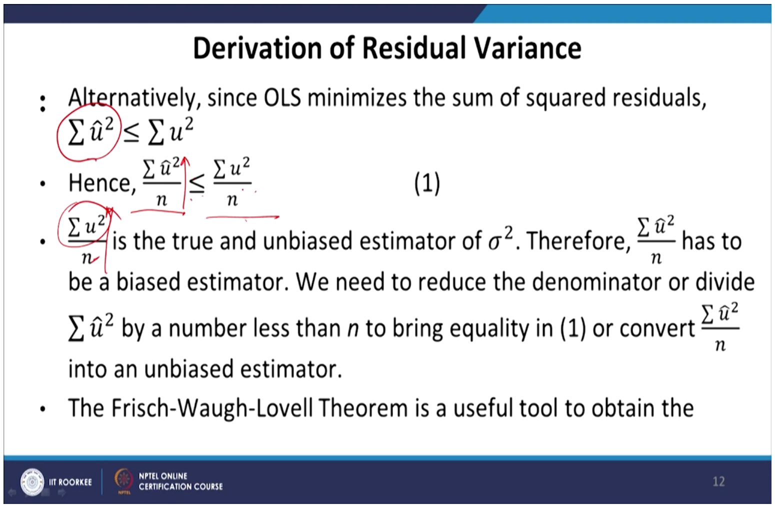
And, in case we have k plus 1 variables and there will be k plus 1 first-order conditions. Thus, they are only n minus 2 degrees of freedom in the OLS residuals as opposed to n degrees of freedom in the errors. So, when we talk about degrees of freedom, what it actually implies? In case, the population error actually can if I consider n observations then the population error can be any n observations. But, what happens is that, when it comes to sampling residuals, we are choosing alpha and beta in order to minimize the values of the, sample, the squared sum of the sample residuals. So, as a result of which, it says that since there are two first-order conditions, the sample residuals can actually take n minus 2 observations or 2, n minus 2 free observations. So, I mean I give a simplistic example of the concept of degrees of freedom.

Suppose, I say that there are ten observations of the variable X. So, the variable X takes ten observations and I do not specify what are the observations. But I only specify that these ten observations add up to a number say 100. So, I can take up any numbers below 100, so that these numbers add up to 100. But if I specify two observations, so this is a case where there are 10 degrees of freedom.

I can take any ten numbers, those will add up to 100. But, in case I specify two numbers. So, suppose I specify that X1 is equal to 8 and X2 is equal to 18. Then, the rest of the 8 numbers have to adjust themselves accordingly, so that the total sum is actually 8 plus 18, that is 100 minus 26. So, now, you can see that I have imposed two conditions. And that is why two degrees of freedom are reduced.

Now, the degree of freedom is actually equal to 10 minus 2, that is 8. There are only 8 observations who are free to take any number, which adds up to 100 minus 26. So, in a similar fashion, because there are two constraints imposed that is while choosing alpha hat and beta hat, that reduces the degrees of freedom for the sample residuals by n minus 2. Otherwise, there had been n observations or n degrees of freedom for the sample residuals as well, unless or if we had not followed the procedure of OLS.

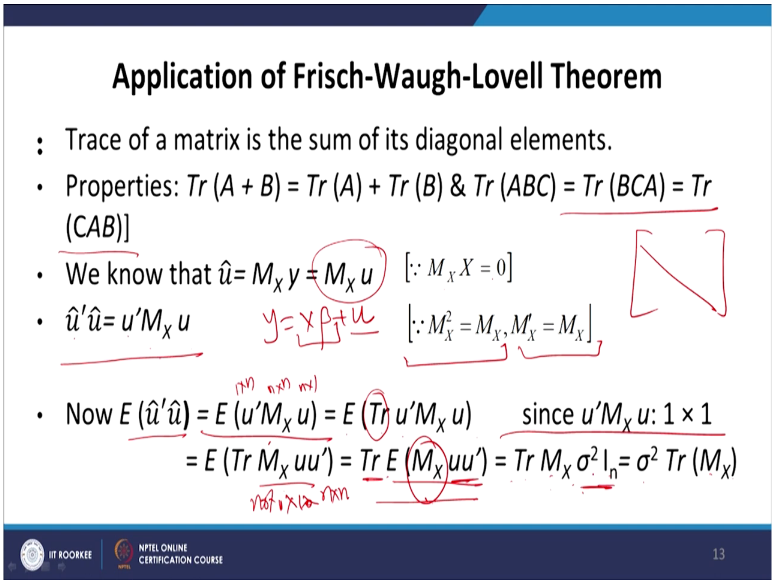
(Refer Slide Time: 23:03)




So, alternatively, OLS minimizes the sum of squared residuals. Of course, this is a deliberate attempt to make it the smallest possible. As a result of this, summation ut had square so, u hat square must be less than equal to summation u square, that is the population counterpart. Hence, summation u hat squared divided by n should be less than equal to summation u square by n. Summation u square by n is the true and unbiased estimator of sigma.

Therefore, summation u hat square by n has to be a biased estimator. We need to reduce the denominator because this is something that is fixed. So, we can actually only change the denominator. So, we can reduce the denominator in order to make the entire expression going up. It goes up, it becomes equal to summation u square by n. So, you need to reduce the denominator or divide summation u hat square by a number less than n to bring equality in 1, or convert summation you had square by n into an unbiased estimator.

(Refer Slide Time: 24:14)




So, the Frisch Waugh Lovell theorem is a useful tool to obtain this. By now, I have explained the application of Frisch Waugh Lovell theorem. So, while discussing it, I first use another concept, which is the trace of a matrix. So, the trace of a matrix is the sum of its diagonal elements. So, this is a matrix, the sum of its diagonal elements is called the trace. Properties of trace which are going to be used here are also discussed briefly.

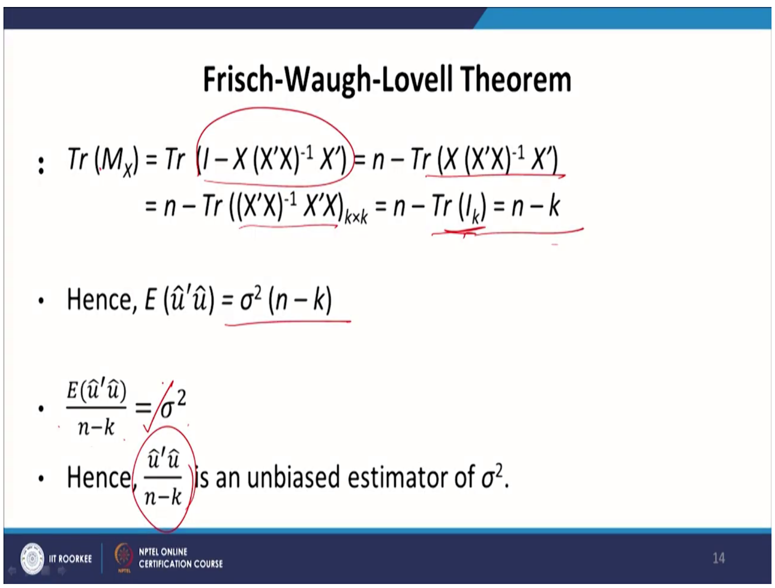
First of all, trace A plus B equals trace A plus trace B, and trace ABC are all different matrices is equal to trace BCA equals trace CAB. Now, we know that u hat is equal to Mx y. Now, Mx y is equal to Mx u also, because y is equal to x beta plus u. And Mx X equals to 0. So, X beta multiplied by Mx becomes 0, I am left with Mx u. Now, we see that u hat prime, u hat is actually equal to u prime Mx u. Because Mx square is equal to Mx, and Mx prime equals to Mx.

Because of this, as Mx is a symmetric matrix as well as the idempotent matrix, we can write it like this *(Refer to slide above).* So, u hat prime multiplied by u is Mx u prime multiplied by Mx u. So, u prime Mx, Mx u is u prime Mx u. Now, the expected value of u hat prime u hat is the expected value of this expression. Now, you can see that this is a 1 by n matrix, Mn is an n by n matrix. And u is an n by 1 matrix. Since this is a number, I can always write trace of u prime Mx u and before the expected value.

So, a number's trace is basically the number itself. But now, since applying this property of the trace of a matrix, I can always write it as Mxu u prime trace u prime Mx u is equal to trace Mx uu prime. Now, this is not a 1 by 1 matrix, this is actually a n by n matrix now. Now, I take the trace operator outside and expectation goes inside. And this is actually non-random given that this is depends only on the values of X and conditional upon x.

So, Mx is non-random, Mx comes out the expected value of uu hat is sigma square In that has already been proved. So, sigma squared being a constant it actually comes out. So, I am left with a trace of Mx, Mx I is actually Mx. Because I is an identity matrix. So, sigma square trace Mx.

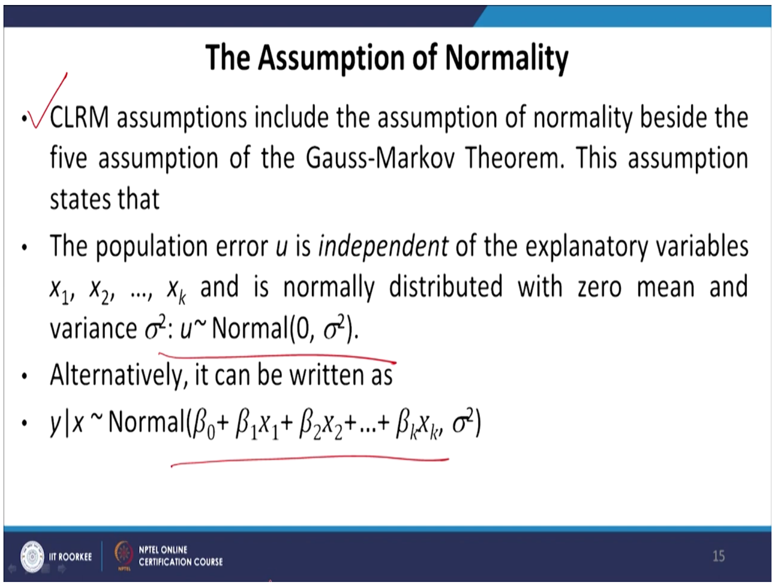
(Refer Slide Time: 27:30)




Now trace Mx is I expand or input the value of Mx, the trace of the matrix I will be n. And then trace of this matrix rearranging terms I arrive at a trace of this matrix as k. So, the trace of Mx is actually n minus k. Now, you remember, I had the expected value of u hat prime, u hat is equal to sigma square trace of Mx, a trace of Mx is n minus k. So, I replace it with n minus k, which implies that the expected value of u hat prime u hat divided by n minus k is equal to sigma square.

So, this shows that this is an unbiased estimator of sigma square and not n. So, I am not going to divide u hat prime by u hat by only n, in order to arrive at an unbiased estimator of the population variance. I need to divide it by n minus k, if there are k variables in the regression. If there are two variables then it should be n minus 2 and so on. And then finally, we talk about the assumption of normality. That has already been actually discussed in the case of or in the context of simple regression analysis.

(Refer Slide Time: 28:56)



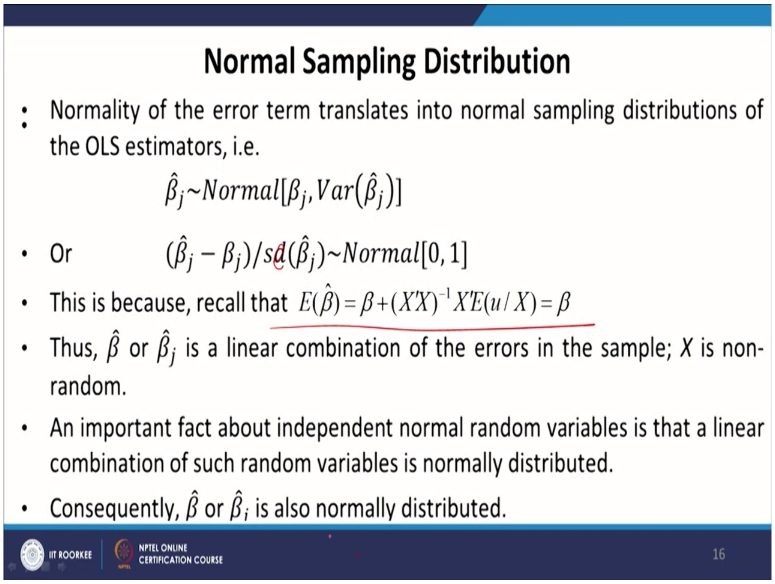

That CLRM that is classical linear regression model assumptions include the assumption of normality, beside the five assumptions of the Gauss Markov theorem, this assumption states that the population error u is independent of the explanatory variables X1 to Xk and is normally distributed with mean 0, and variance sigma square. So, u is normally distributed with 0 mean and sigma square as its variants.

Alternatively, it can also be written as $y|x \sim \text{Normal}(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k, \sigma^2)$

Now, if you remember that, I just need to mention here that in the case of Gauss Markov theorem, we had or we have an additional assumption, that is the assumption of no perfect colinearity between or among the independent variables.

This assumption was not relevant in the case of a simple regression model, because we had only one independent variable. Since, in the case of multiple regression analysis, we have multiple independent variables, so this assumption has been brought into that is an additional assumption we have for multiple regression analysis under the Gauss Markov theorem. But this assumption was actually very much a part of the simple regression analysis as well.

(Refer Slide Time: 30:16)

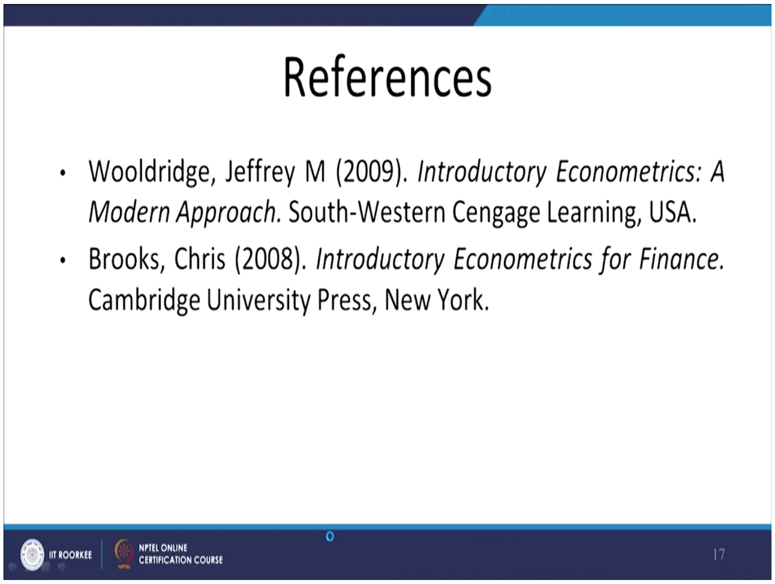



So, from here we also derive the sampling distribution, that is normality of the error terms translate into normal sampling distributions of the OLS estimators, that is beta j hat is also normally distributed with beta j as the mean and variance of beta j hat as the variance, or alternatively, we can write that, beta j hat minus beta j divided by the standard deviation. Ideally, it should be the standard error of beta j hat normally distributed with mean 0 and variance 1.

This is, this comes straight away from the fact that the expected value of the beta hat is equal to beta, that is the unbiasedness. Thus beta hat or beta j hat is a linear combination of the errors in the sample, X is non-random. That is given that X is non-random, beta hat or beta j hat is a linear combination of the errors in the sample. An important fact about independent normal random variables is that a linear combination of such random variables is normally distributed.

Consequently, beta hat or beta j hat is also normally distributed. So, that brings me to the end of some of the discussions on multiple regression analysis. We will further continue with some other topics on multiple regression analysis in the next module.

(Refer Slide Time: 31:51)




You can follow these books for the discussion that I have had so far. Thank you.