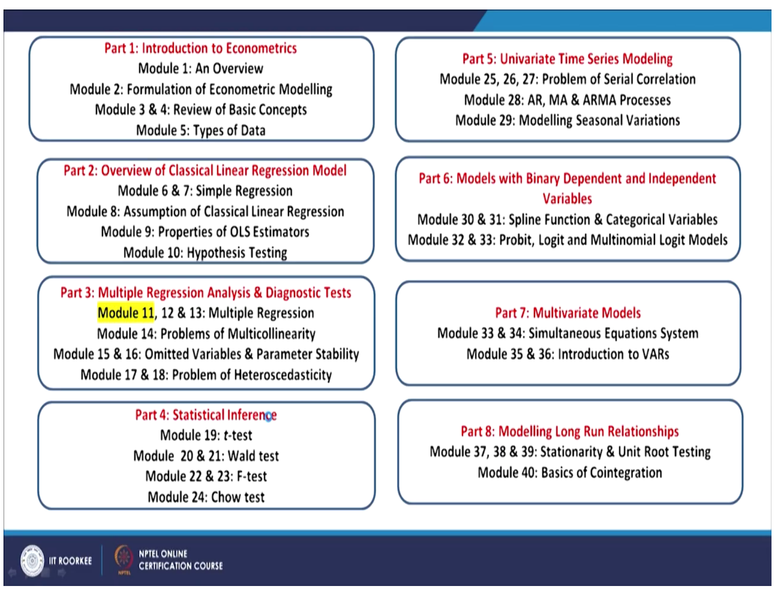
**Econometric Modelling**
**Professor Sujata Kar**
**Department of Management Studies**
**Indian Institute of Technology, Roorkee**
**Lecture – 11**
**Multiple Regression - I**

Hello, and this is the eleventh module of the course on econometric modelling. So, far we have introduced the simple regression analysis where we had only two variables. Now, we are getting into multiple regression analysis where we will be having more than two variables, or there will be more than one independent variable.

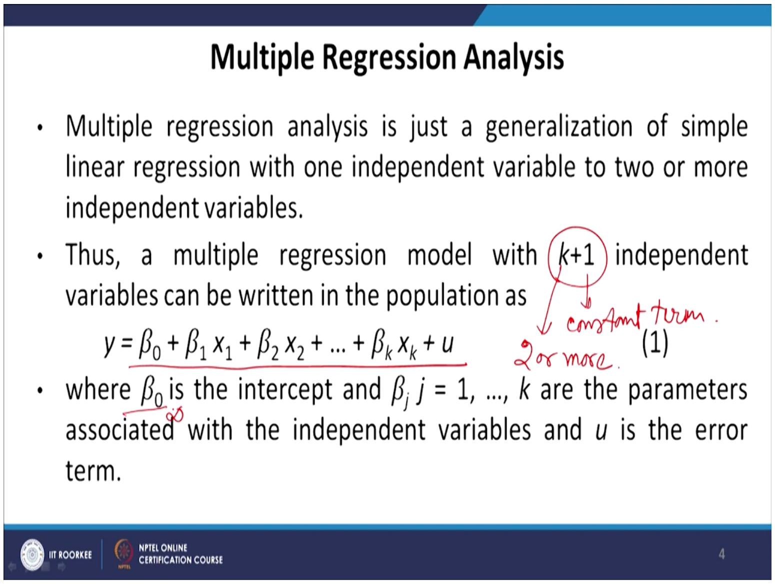
So, when we say that there were only two variables, one is the independent variable and the other was a dependent variable. Along with that, there was a constant term. Constant terms are always there, unless and until there is a special situation or case.

(Refer Slide Time: 01:20)




When we talk about multiple regression analysis, then it basically involves one dependent variable, but several independent variables. So, it can be two or more. We have three modules on the basics of multiple regression analysis. This is the first module on multiple regression analysis.

(Refer Slide Time: 01:39)

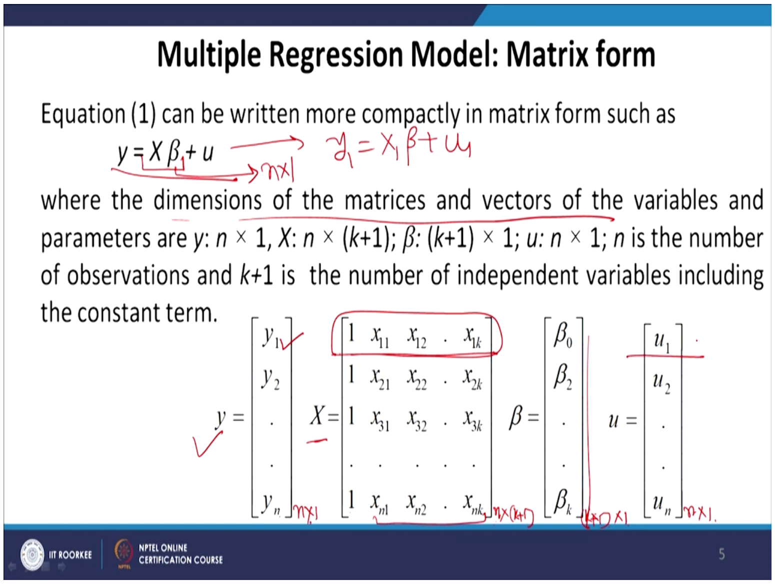



Multiple regression analysis is just a generalization of simple linear regression with one independent variable to two or more independent variables. So, we move from one independent variable in the case of simple linear regression to two or more independent variables. There is a multiple regression model with k plus 1 independent variable can be written in the population as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u \qquad \textit{(equation-1)}$$

Now, we have mentioned here k plus 1, because 1 here refers to the constant term. And this k can take any value, which is two or more. So, in case k is 1, we have the simple regression case, when we have k equals 2 or greater than 2, then we have the multiple regression case. So, when beta naught ($\beta_0$ ) is the intercept, which we were earlier denoting by alpha. So, $\beta_0$ is equivalent to our previous alpha. And $\beta_j$ equals 1 to k are the parameters associated with the independent variables. And u is the error term, the usual error term.

(Refer Slide Time: 03:00)




So, we can write the multiple regression model in matrix form, which actually saves us a lot of computation. And it is much more convenient to deal with multiple regression in matrix form. So, *equation-1* can be rewritten more compactly in matrix forms *(refer to slide above, time 03:00)* such as y equals x beta plus u. So, y is a vector of observations on the dependent variable.

We are now dealing with say, cross-section data. So, we have n observations, and consequently, the dimension of y will be n by 1. So, there are n rows and only one vector, I have also mentioned here the dimensions of the matrices and vectors of the variables and parameters. And so, accordingly, you can see that x, the matrix of independent variables will be having n rows and k plus 1 column *(refer to slide above, time 03:00)*.

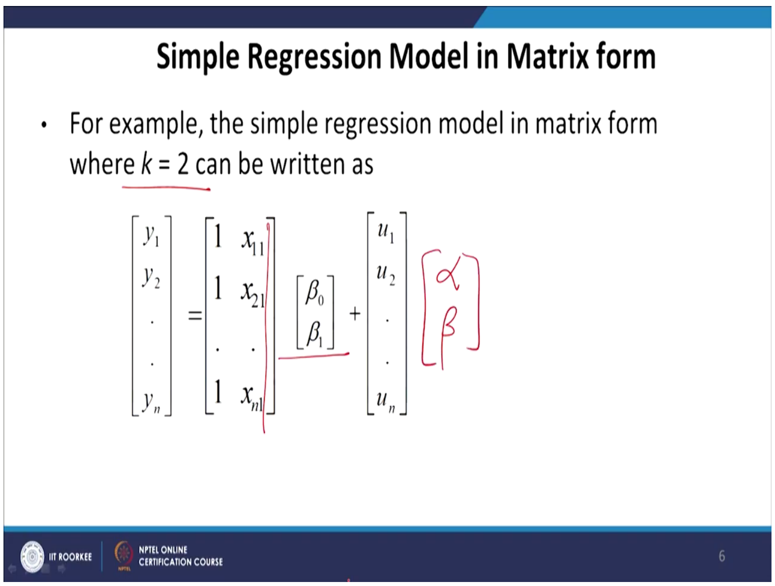
So, in the one column, and there are the k columns for the k independent variables. So, we have a k plus 1 column. And beta is again a vector, which is of the dimension k plus 1 multiplied by 1. So, there are k plus 1 row and only one vector. And u, the error terms has a dimension of n by 1. So, there are n observations, corresponding to that, there is only one column. So, you can see that we multiply x with beta.

So, multiplying x with the beta, we have n by k plus 1 multiplied by k plus 1 by 1. So, x beta gives us a vector of n by 1, y is n by 1 and u is anyway n by 1. So, equality has been established.

And accordingly, we can always write this model as $y_1 = X_1 \beta + u_1$. $X_1$ actually corresponds to the observations on k independent variables related to the first unit, that is, this is the first observation or first entity from which we are collecting data.

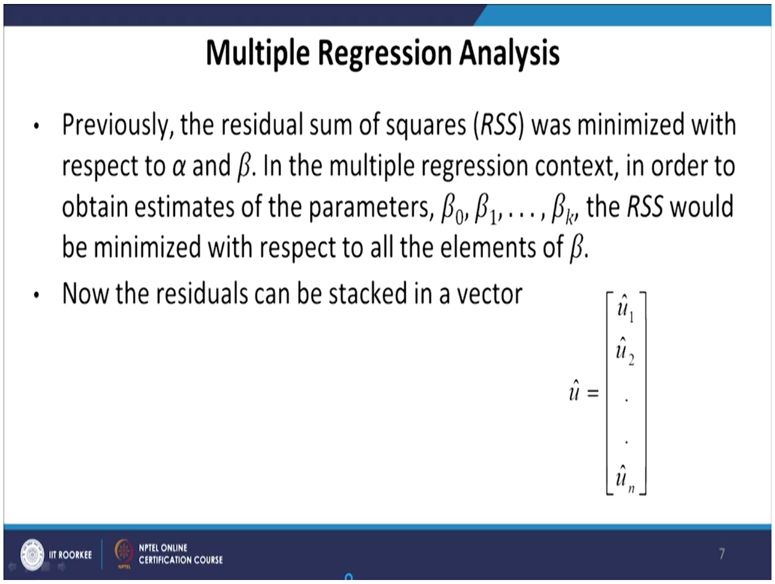
And this actually corresponds to the observations from the first entity on k variables. And then we have, of course, the k parameter estimates or k parameters in the population, and there will be one error term, u 1. And similarly, we can have yi equals xi beta plus ui and so, on.

(Refer Slide Time: 06:01)




So, if we try to draw an analogy between this k plus one variable case, and two variables case, that is when k is equal to 2 or rather here k is equal to 1, but in totality, we have only two variables. So, in that case, I have one independent variable, the inner vectors of parameters are $\beta_0$, $\beta_1$ which are actually our original alpha and beta. And then we have a vector of the error terms.

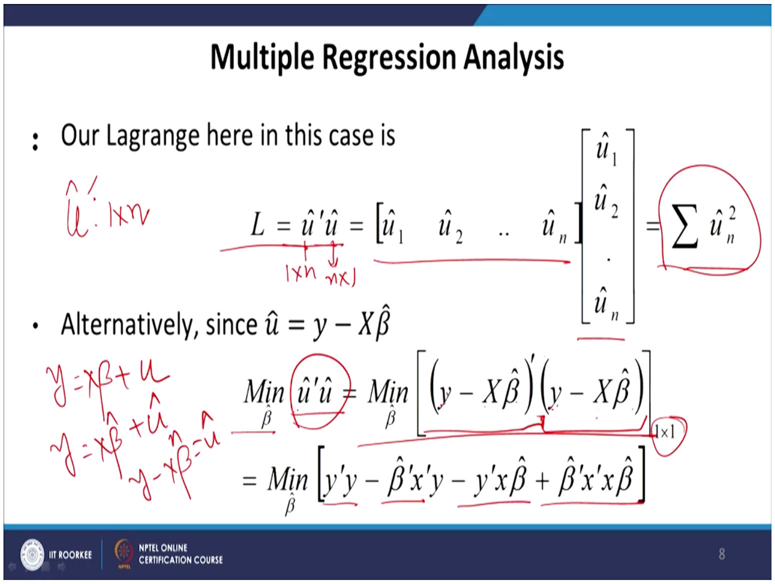
(Refer Slide Time: 06:33)




Previously, the residual sum of squares or RSS was minimized with respect to alpha and beta, in the multiple regression context in order to obtain estimates of the parameters that is beta 0 beta 1 to beta k, the residual sum of squared would be minimized with respect to all the elements of beta. Earlier it was only two variable cases, now, we would be arriving or trying to derive the fitted regression line by choosing, all the values of the parameters beta naught to beta k.

We minimize the residual sum of squares by choosing the parameter values of beta naught to beta k. Now, the residuals can be stacked in a vector, which is very similar to the population error. The way population error has been stacked, the residuals can also be stacked. Because we minimize the residuals, not the population variance or the population error. So, we need to bring in the concept of residuals because we work with a sample, we do not work with the population most often. And that is why we minimize the residuals and not the population error actually.

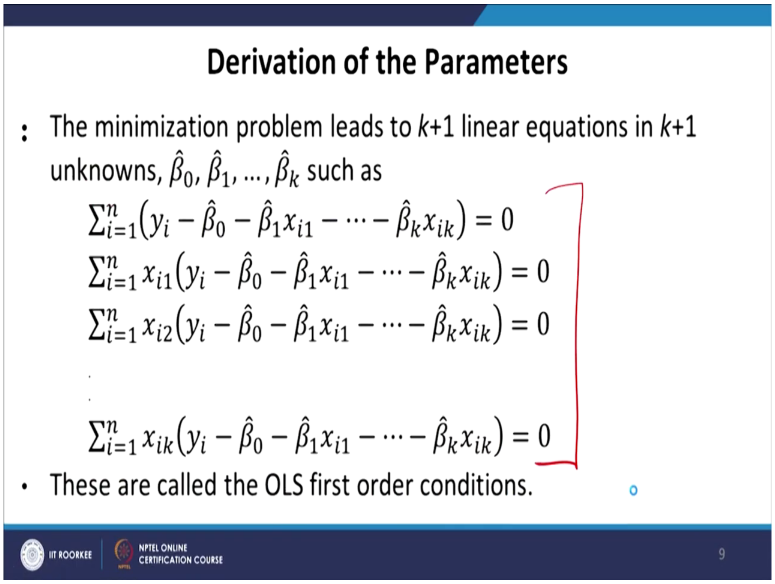
(Refer Slide Time: 07:55)




So, here, in this case, our Lagrange is set as so. Lagrange multiplier or basically the expression by minimizing which we are obtaining the values of the parameter estimates, here, in this case, is u hat prime u hat. You can see that u hat prime, u, u hat was an n by 1 matrix. So, u hat prime, which is the transpose of u hat is actually a 1 by n vector. *(Refer to slide above, time 07:55)*

So, this is actually a number which is u hat n squared. And summed over all n observations. Alternatively, since u hat is equaled to y minus x beta hat, what we are minimizing is u hat prime, u hat with respect to the beta hat. And this is the expression because u hat is y minus x beta hat. You know that y equals x beta plus u. Its sample counterpart would be y equals x beta hat plus u hat. So, y minus x beta hat is equaled to u hat *(Refer to slide above, time 07:55)*. So, it is y minus x beta hat prime corresponding to u hat prime and this is y minus x beta hat. And when I expand it, then I will be having y prime y. So, y prime y, then X prime beta hat transposes beta hat transpose x transpose. So, beta hat transpose x transpose multiplied by y. Similarly, I have y prime x beta hat, and then again beta hat prime X prime x beta hat. Note that, this is of 1 by 1 dimension which implies that this is also a number. Since u hat prime u hat is a number, the left-hand side is a number. So, the right-hand side also has to be a number. *(Refer to slide above, time 07:55)*

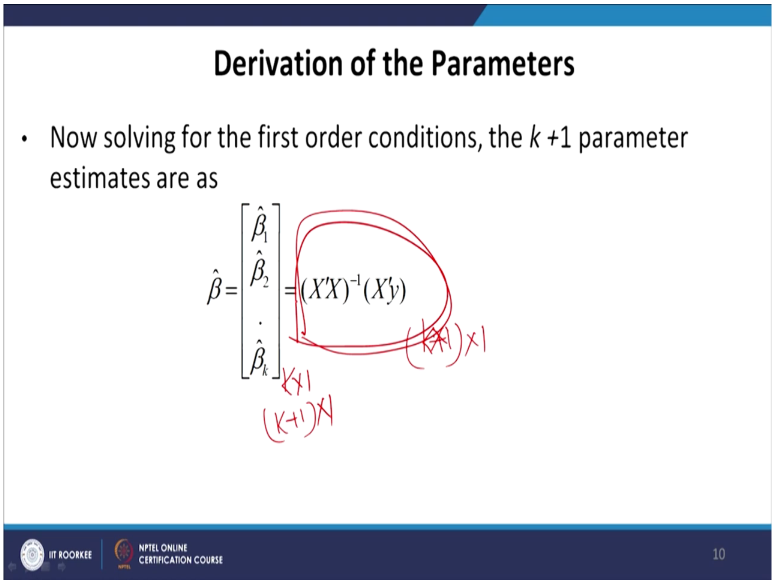
(Refer Slide Time: 09:58)




So, we are now going to minimize this expression. So, first of all, what we are going to do is that we would be taking derivative of this expression with respect to all the betas. So, since there are k plus 1 beta, we will be arriving at k plus 1 linear equations in k plus 1 unknowns. And these unknowns are beta naught hat beta 1 hat and so on, up to beta k hat. *(See the slide above)*

And these are the equations, that we arrive at or obtain when we take the first derivative of this expression with respect to individual betas. These are called the OLS first-order conditions, very similar to the OLS first-order conditions that we had obtained in the case of simple regression. There, we had only two first-order conditions, nowhere, we will be having k plus 1 first order conditions.

(Refer Slide Time: 10:55)

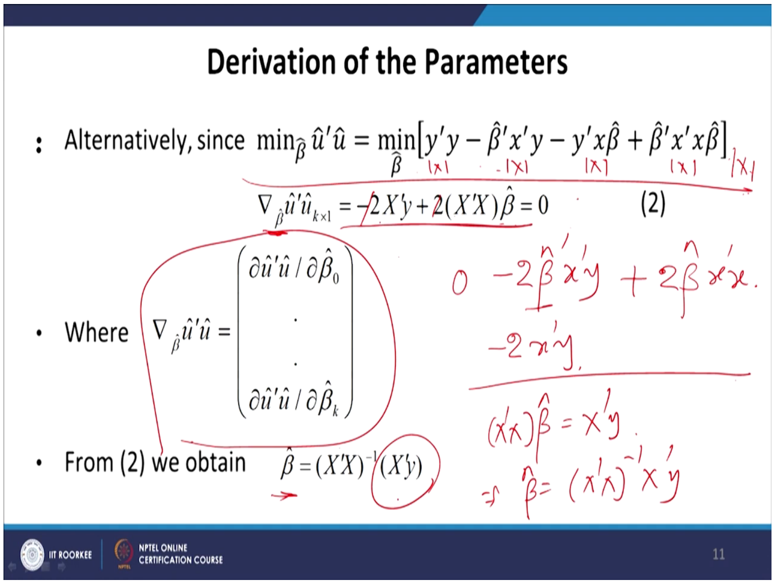



Now, solving for the first-order conditions, the k plus 1 parameter estimates are

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ . \\ \hat{\beta}_k \end{bmatrix} = (X'X)^{-1}(X'y)$$

This is a very crucial, derivation where we say that beta hat is X prime X inverse X prime y, this is a k by 1 vector. This also has to be a k plus 1 by 1 vector. So, once we have this expression, then equating each row with the corresponding row, we can have values of beta, betas that is starting from beta 0 hat to beta k hat.

(Refer Slide Time: 11:41)




Now, we get into the derivation. So, how do we arrive at X prime X inverse X prime y? Now, if you remember in the previous to previous slide, we had this expression minimizing u hat prime u hat is equal to minimizing this expression. These are of dimension 1 by 1. So, by taking the first derivative with respect to beta hat, where this is actually the true operation, I am taking the first derivative of u prime hat, or u hat prime u hat with respect to beta hat individuals, individual betas.

And that is how we have this expression. Now, by doing so, we will be arriving at this expression. Why this is so? Because, you see that since this is a 1 by 1 dimension, individual components are also 1 by 1. So, they are all numbers. Now, when we take the first derivative of y prime y with respect to the beta hat, this is 0 because beta hat does not exist here. And these two can be written as minus 2 beta hat prime X prime y, or it can also, it could have been also written as minus 2 y prime x beta hat. *(Refer to slide time 11:41)*

Now, when I take the derivative with respect to the beta hat, then I will be arriving at minus 2 X prime y, which is this expression or this part. And this is actually equivalent to obtaining a plus 2 beta hat X prime x *(Refer to slide time 11:41)*. Now, from these two, we obtain the below equation:

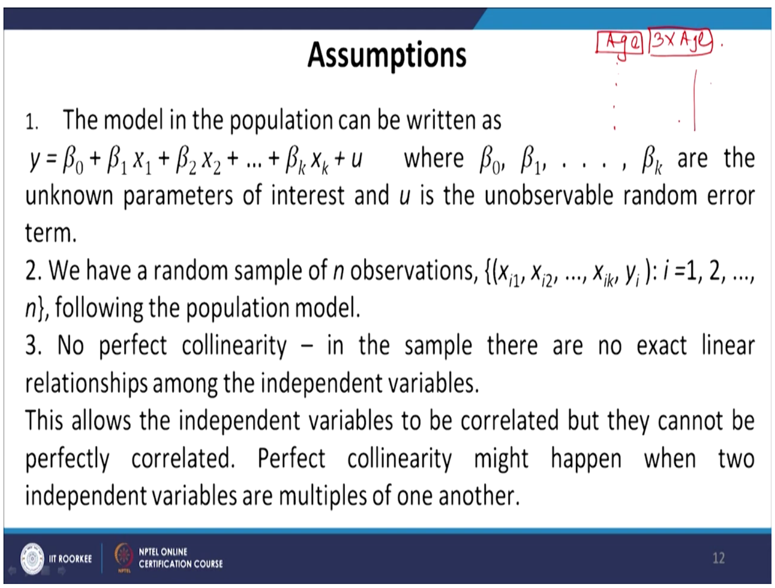
$$\nabla_{\hat{\beta}} \hat{u}'\hat{u}_{k\times 1} = -2X'y + 2(X'X)\hat{\beta} = 0$$

And then from here, we arrive at the expression for beta which is

$$\hat{\beta} = (X'X)^{-1}(X'y)$$

So, when I take this to the other side, this becomes 2 cancels out. So, I have X prime x beta hat equals X prime y, which implies beta hat equals X prime X inverse X prime y *(Refer to slide time 11:41)*.

(Refer Slide Time: 14:08)




Now, we talk about the assumptions of multiple linear regression, they are exactly the same as that of simple regression. So, first of all, the model in the population can be written as

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$ where $\beta_0, \beta_1, \ldots, \beta_k$ are the unknown parameters of interest and $u$ is the unobservable random error term. So, the first assumption relates to, the model specification in the population.

Second, we have a random sample of $n$ observations, $\{(x_{i1}, x_{i2}, ..., x_{ik}, y_i): i = 1, 2, ..., n\}$, following the population model.

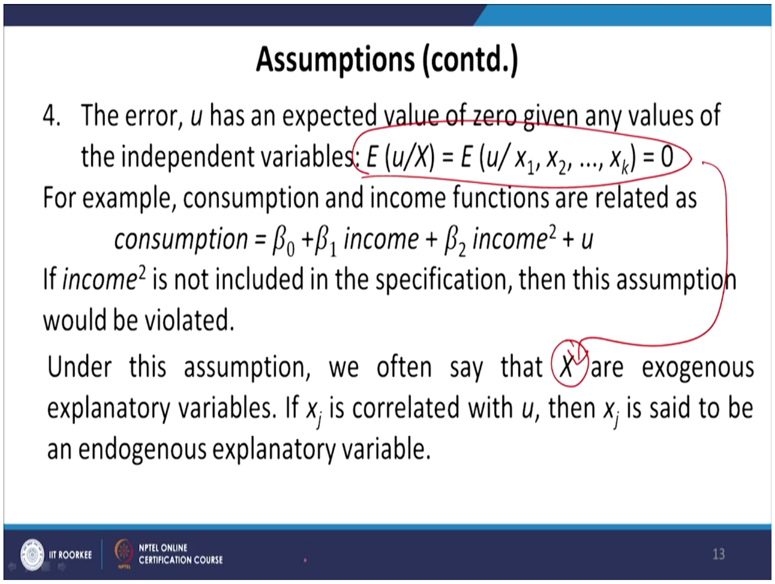
The second assumption states that we pick up a sample of size and observation from the population. The third assumption states that there is no perfect colinearity. In the sample, there are no exact linear relationships among the independent variables. This allows the independent variables to be correlated, but they cannot be perfectly correlated,

Perfect colinearity might happen when two independent variables are multiples of one another. So, this is an assumption which we will be dealing with in the next part, when we talk about the problem of multicollinearity. For the time being, this simply assumes that there cannot be perfect collinearity among the independent variables, which essentially implies that one variable lies completely in the column space of another variable, or maybe one variable is hundred percent or fully explained by the other variable.

So, for example, in the case of independent variables, suppose I have one independent variable age, so, I jot down the age of individual respondents. And if I have another variable, which is 3 multiply by 8, this is a case of perfect collinearity. Because you can see that, this independent variable lies completely in the column space of this independent variable or vice versa.

So, we cannot have perfect collinearity among the independent variables, this is the basic idea. And what are the problems that arise? And why cannot we have it? That will be discussed at length when we when we discuss the problem of multicollinearity.

(Refer Slide Time: 16:42)

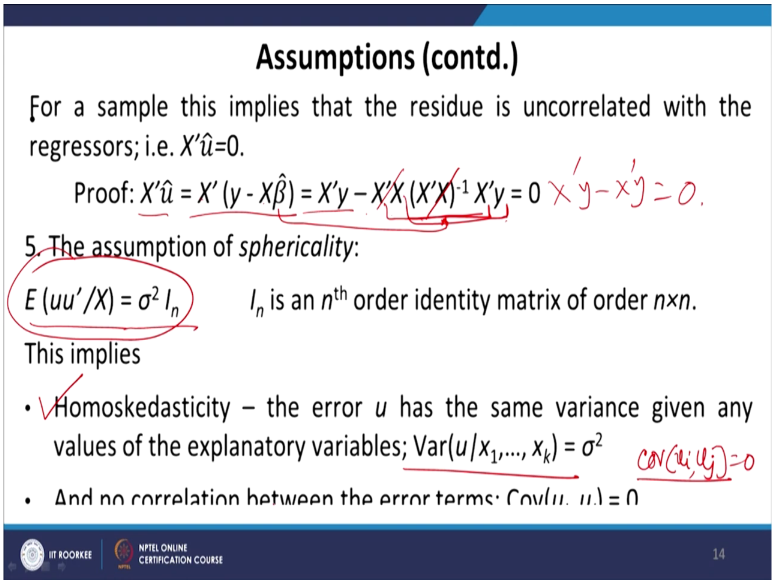



So, continuing with assumptions, the fourth assumption is that the error u has an expected value of 0, given any values of the independent variables. That is expected value of u conditional upon x is actually equal to 0. This is also an assumption we have dealt with at length while discussing the assumptions of simple regression analysis. For example, consumption and income functions are related as, $consumption = \beta_0 + \beta_1\ income + \beta_2\ income^2 + u$

So, if income square is not included in the specification, then this assumption would be violated. So, you can see that u is not only independent of the observation x, but it is also independent of any functional form of x. So, under this assumption, we often say that x is exogenous explanatory variable. So, when this assumption is fulfilled, then x's are called exogenous.

If x j is correlated with u, then x j is said to be an endogenous explanatory variable. So, first of all, the entire matrix x, that is all the independent variables need to be exogenous. If any one of them is not exogenous, then we call it an endogenous variable.

(Refer Slide Time: 18:13)



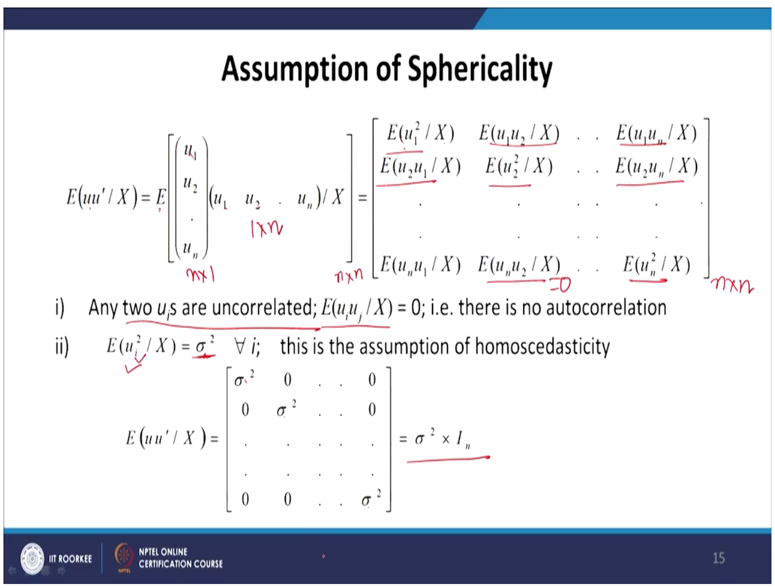

We talk a little more about the previous assumption. For a sample this assumption implies that the residue is uncorrelated with the regressor, that is X prime u hat is equals to 0. This can be also proved very easily, that X prime u hat is actually explained y minus x beta hat, which is if I now expand X prime y minus X prime x, this is X prime x and then this is the expression for beta hat, X prime X inverse X prime y.

So, we have X prime x from here and X prime X inverse X prime y. Now, you can see that X prime x and X prime X inverse, they cancel out and I have X prime y minus X prime y which also cancels out and we have 0. So, if beta had, has this formula, then we can prove that X prime and u hat, are independent of each other or uncorrelated with each other. So, this actually follows, also implies that the independent variables are exogenous, that is they are not correlated with the error term. (*Refer to slide 18:13)*

Now, talking about the last assumption of sphericality, this is, actually houses two assumptions, the first assumption is of homoscedasticity. That is, the error u has the same variance given any values of the explanatory variables. So, various of u given x1 to xk is equal to sigma square. This assumption also encompasses another assumption and that is there is no correlation between the error term. So, the covariance between ui uj is equal to 0.

So, we do not assume any correlation among or between the error terms. And we also assume a constant error variance, that is variance of u equals sigma square. Now, this is written as the expected value of u prime conditional upon the values of x equals sigma square I n. Where In is an nth order identity matrix of order n by n, and the implications have already been mentioned. Now, I expand this expression specifically expected value of u u prime given x, and how we arrive at the value or the expression sigma square In under these assumptions(*Refer to slide 18:13)*.

(Refer Slide Time: 21:04)




So, the assumption of sphericality, if expanded we will have;

$$E(uu'/X) = E\left[\begin{pmatrix} u_1 \\ u_2 \\ . \\ u_n \end{pmatrix}\begin{pmatrix} u_1 & u_2 & . & u_n \end{pmatrix} / X\right] = \begin{bmatrix} E(u_1^2/X) & E(u_1u_2/X) & . & . & E(u_1u_n/X) \\ E(u_2u_1/X) & E(u_2^2/X) & . & . & E(u_2u_n/X) \\ . & . & . & . & . \\ . & . & . & . & . \\ E(u_nu_1/X) & E(u_nu_2/X) & . & . & E(u_n^2/X) \end{bmatrix}$$

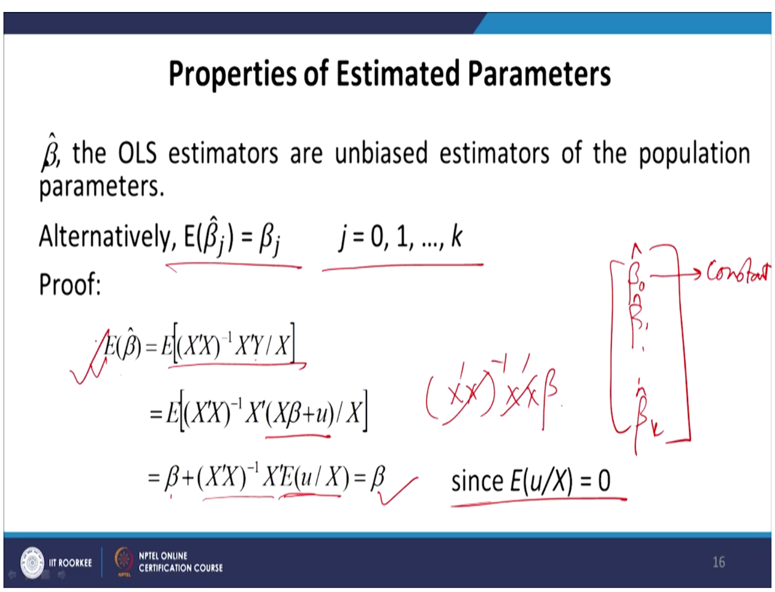
Now, since any two u is are uncorrelated, the expected value of $E(u_i u_j / X)$ is 0, this is called the assumption of no autocorrelation. And the second thing that $E(u_i^2 / X) = \sigma^2$ for all i. This actually makes these expressions equal to sigma square. So, the diagonal, diagonal terms are all sigma squared, sigma is independent of the observation.

So, sigma is just a constant number, it is not dependent on the ith observation or jth observation. So, we have a sigma square along with the diagonal terms. As a result of which finally we have sigma squared multiplied by In,

$$E(uu'/X) = \begin{bmatrix} \sigma^2 & 0 & . & . & 0 \\ 0 & \sigma^2 & . & . & 0 \\ . & . & . & . & . \\ . & . & . & . & . \\ 0 & 0 & . & . & \sigma^2 \end{bmatrix} = \sigma^2 \times I_n$$

where In is an n by n identity matrix. So, this is how we present the assumption of sphericality, which encompasses two assumptions of homoscedasticity and no autocorrelation among the error terms.

(Refer Slide Time: 22:59)

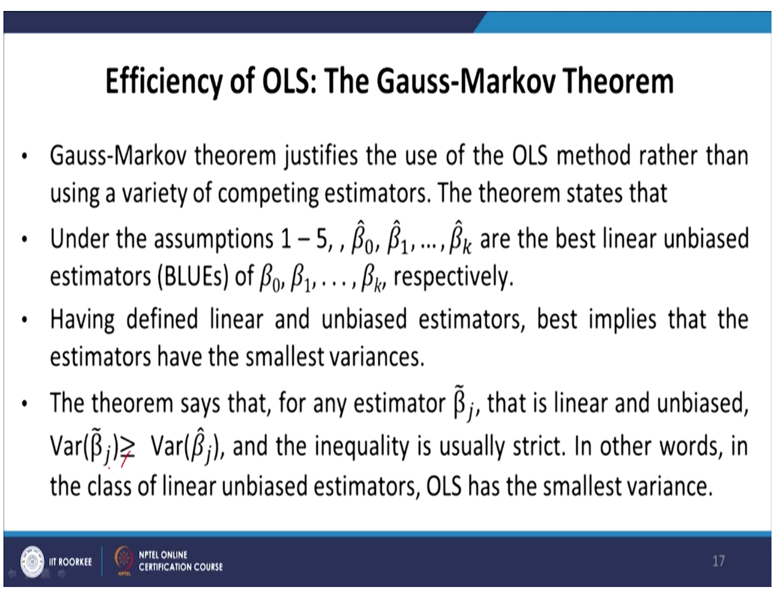


Now, talking about the properties of estimated parameters, beta hat, the OLS estimators are unbiased estimators of the population parameters. So, we again stick to the concept of best linear unbiased estimators. Alternatively, the expected value of beta j equals beta j hat is equals to beta j, for any j 0 to k. Now, we just prove it, that expected value of the beta hat is expected value of X prime X inverse X prime Y conditional upon the values of X. *(Refer slide 22:59)*

Now, y can be expanded as X beta plus u. Now, getting into the multiplication or further expanding it, I will be having X prime X inverse X prime x beta. So, these two terms cancel out and I am only left with the beta, plus X prime X inverse X prime u, which actually remains here. So, X prime X inverse X prime u, and I take the expected value of u only, because X prime X inverse X prime conditional upon X is non-random. *(Refer slide 22:59)*

So, we cannot have any expected value operating on that expression. So, the expected value of u given x is 0, as a result of which the expected value of beta hat is equal to beta. And now, beta is a vector consisting of all the terms like beta naught hat beta 1 hat and so on, where beta naught hat is the constant term. So, we do not need any separate proof for individual beta parameters. This itself implies that all the estimated parameters are unbiased.

(Refer Slide Time: 24:43)

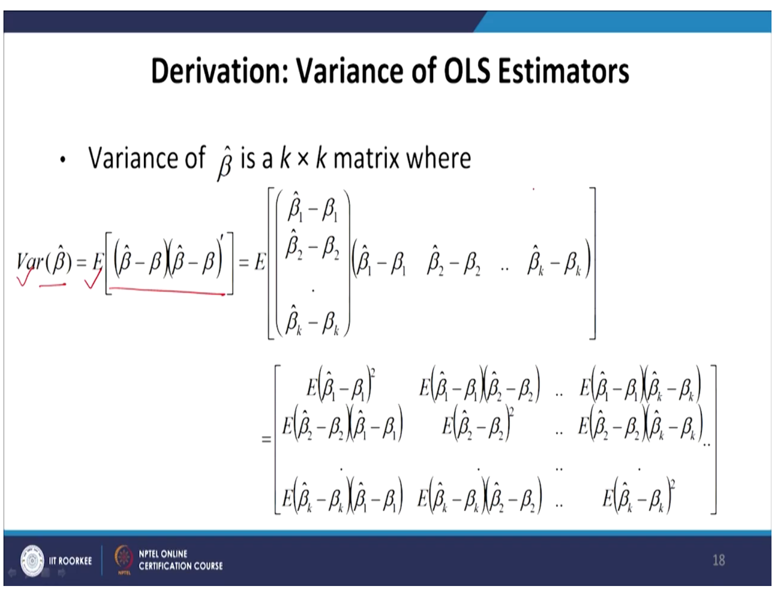


Now, talking about the other property, that is the OLS estimators, or best estimator or minimum variance estimator or when we talk about the efficiency of OLS estimators, we use the Gauss Markov theorem. Gauss Markov theorem justifies the use of the OLS method rather than using a variety of competing estimators. The theorem states that under the assumptions one to five the assumptions that have just been mentioned, beta 0 hat to beta k hat are the best linear unbiased estimators or BLUEs of beta 0 beta 1 to beta k respectively.

That is, they are the best linear unbiased estimators of the population parameters. Having defined linear and unbiased estimators, best implies that the estimators have the smallest variances. The theorem says that for any estimator say beta j tilde, that is linear and unbiased. So, that has other properties of the OLS estimator, various of beta j tilde will be greater than equal to the variance of beta j hat, that is the OLS estimator.

And the inequality is usually strict. So, we would often come across variance of beta j tilde are greater than the variance of beta j hat. In other words in the class of linear unbiased estimators, OLS has the smallest variance.

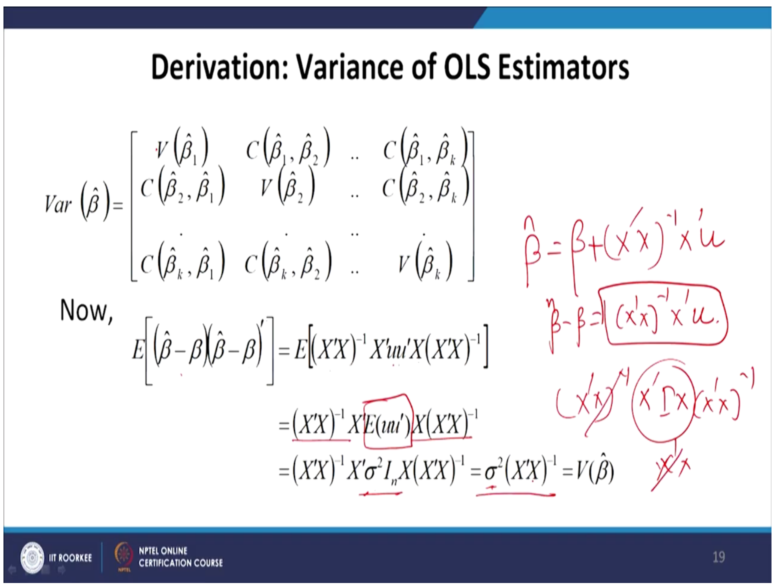
(Refer Slide Time: 26:21)

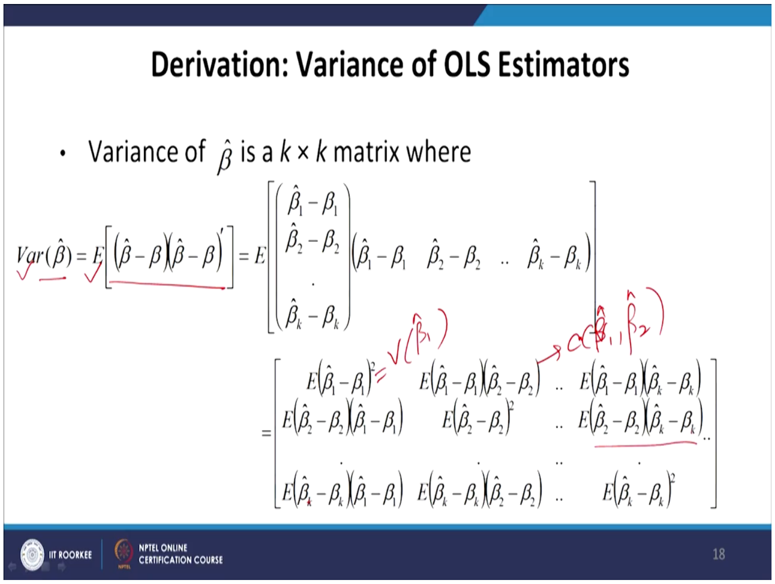



So, now, we derive the variance of OLS estimators. The variance of the beta hat is a k by k matrix, where the variance of beta hat would be written as beta hat minus beta multiplied by beta hat minus beta prime. And then, instead of variance, we have the expectation operator. Now, I multiply them and then we arrive at:

$$Var(\hat{\beta}) = E\left[\left(\hat{\beta}-\beta\right)\left(\hat{\beta}-\beta\right)'\right] = E\left[\begin{pmatrix}\hat{\beta}_1-\beta_1\\ \hat{\beta}_2-\beta_2\\ .\\ \hat{\beta}_k-\beta_k\end{pmatrix}\begin{pmatrix}\hat{\beta}_1-\beta_1 & \hat{\beta}_2-\beta_2 & .. & \hat{\beta}_k-\beta_k\end{pmatrix}\right]$$

(Refer Slide Time: 26:50)






So, we have variance of beta hat. So, this expression, this matrix can be rewritten as

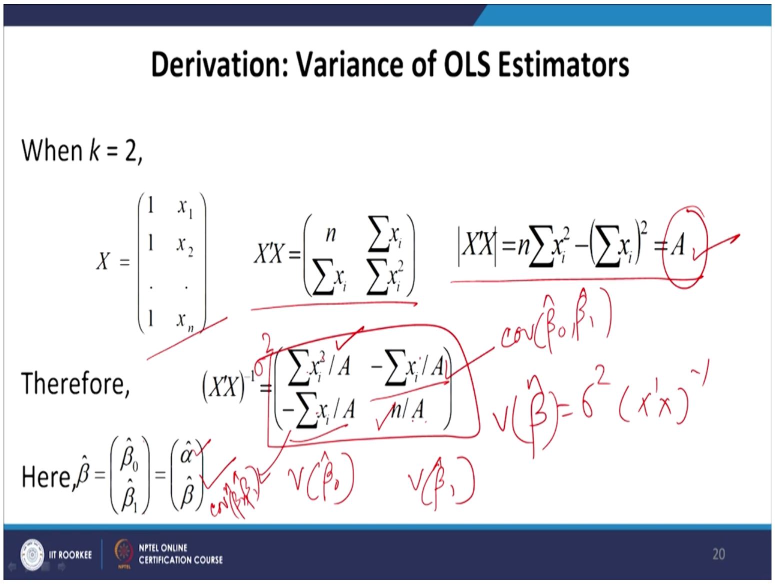
$$Var\left(\hat{\beta}\right)=\begin{bmatrix} V\left(\hat{\beta}_1\right) & C\left(\hat{\beta}_1,\hat{\beta}_2\right) & .. & C\left(\hat{\beta}_1,\hat{\beta}_k\right) \\ C\left(\hat{\beta}_2,\hat{\beta}_1\right) & V\left(\hat{\beta}_2\right) & .. & C\left(\hat{\beta}_2,\hat{\beta}_k\right) \\ . & . & .. & . \\ C\left(\hat{\beta}_k,\hat{\beta}_1\right) & C\left(\hat{\beta}_k,\hat{\beta}_2\right) & .. & V\left(\hat{\beta}_k\right) \end{bmatrix}$$

Now, we get into an alternative procedure, where we have just shown you that beta hat is equal to beta plus X prime X inverse X prime u. *(Refer to slide time 26:50)*

This implies that beta hat minus beta is X prime X inverse X prime u. So, now, beta hat minus beta is replaced with X prime X inverse X prime u, again beta hat minus beta prime is replaced with u prime X into X prime X inverse. And what happens? These are actually non-random entities. So, they are, we do not have expectations operating on them. So, expectation actually is working only on u u prime.

We know that u u prime, the expected value of u prime is sigma square In. So, sigma square being constant actually comes out. And what I am left with is X prime X inverse X prime InX X prime X inverse. So, X prime Ix is actually X prime X, because I is an identity matrix. So, these two things cancel out and I have left with sigma squared X prime X inverse, which is the variance of beta hat. *(Refer to slide time 26:50)*

(Refer Slide Time: 28:42)

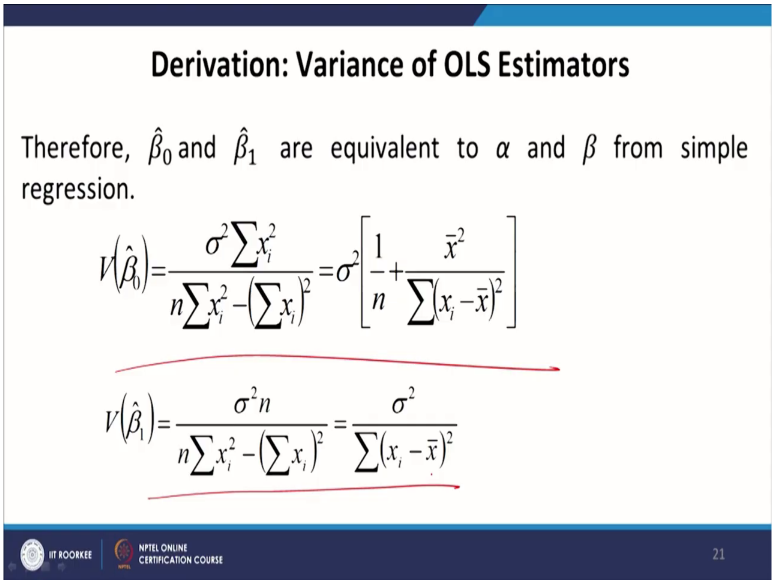



If you remember while working with simple regression analysis, I did not derive the variance of alpha hat, I only derived the variance of beta hat. And I mentioned the derivation of, variances are easier when we work with matrix algebra. So, in the case of k variable case that is sorry, two variable cases that when k is equal to 2, then we have this matrix. And our beta hat is actually beta 0 beta 1 hat, which is equivalent to our alpha hat beta hat with the simple regression case.

So, X prime x is this, this is X prime X, the determinant of X prime X. So, once I obtain the determinant of X prime X and denote it by a, I arrive at X prime X inverse as this expression for the two-variable case. So, this is my X prime X inverse. Now, you remember that beta hat, the variance of beta hat is sigma squared X prime X inverse. So, I have sigma squared here. And what does this give me? This gives me the variance of beta naught hat. *(Refer to slide time 28:42)*

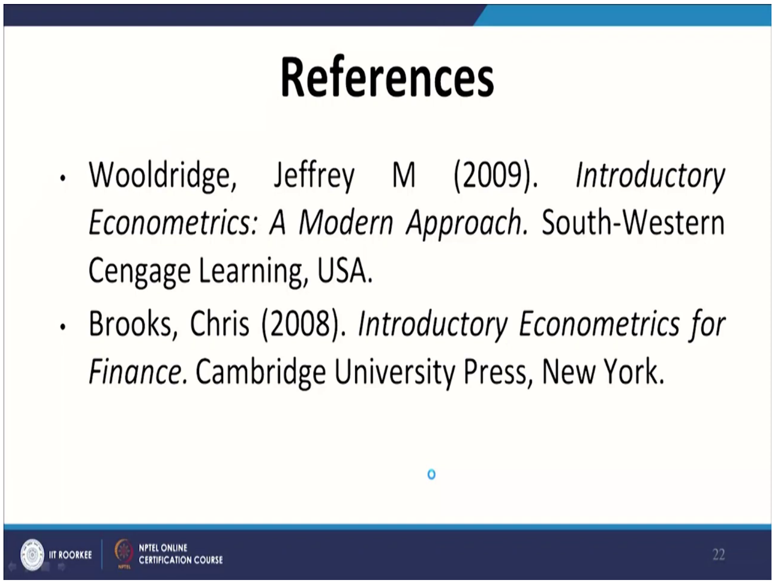
This gives variance of beta 1 hat. This is a k valiant to the covariance between beta naught hat and beta 1 hat. And this is equivalent to again covariance between beta naught hat and beta 1 hat, you can see they are the same terms. So, to our covariance between beta naught hat and beta 1 hat this is variance of beta naught hat or alpha, and variance of beta 1 hat or variance of beta hat in case of simple regression.

(Refer Slide Time: 30:31)




So, these are the expression for the variance of beta 0 hat *(Refer to slide time 30:31)*, which can come directly from the expression of the matrix, that we have just arrived at or derived in the previous slide. And this is the variance of the beta hat. Or here in this case beta 1 hat.

(Refer Slide Time: 30:58)




And these are the books which you can follow for further understanding. In the next also, we will continue with the derivation of some important theorems that will be discussed in the next module, along with the derivation of unbiased estimators of the residual or the population variance. So, we will be deriving the residual variance as an unbiased estimator of the population variance. Thank you.