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 - 46
Multiple Linear Regression Analysis
OLSE, Fitted Model and Residuals

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 and we are going to continue with our module on the Linear Regression Analysis and we are going to continue with our chapter on the Multiple Linear Regression Analysis.

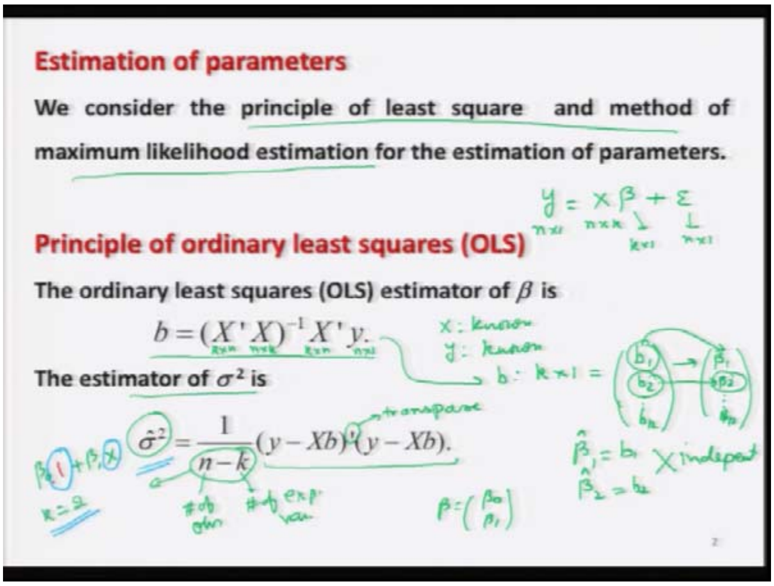
So, you can recall that in the earlier lecture, I have given you a brief background about what is multiple linear regression model and what are the basic assumptions; what is the setup; how are we going to collect the observation; how are we going to represent them in the form of a vectors and matrices and so on.

So, now, in this lecture, I am going to estimate the parameter β but before that, you have to go back to your lectures in simple linear regression model and you have to recall that how we have done the estimation of parameters β_0 and β_1. We had used the principle of least squares and maximum likelihood estimator and we had minimized the sum of a square due to random errors.

So, now exactly the same thing, I am going to follow here in the case of multiple linear regression model. But how to do it, that is my objective in this lecture to explain you because I had told you earlier that, please try to have this simple linear regression model very clearly and try to understand the basic fundamental, they are looking very simple; but they will be used in the case of multiple linear regression model.

So, if you have not done it, so better is to go back and try to have a look and then, you try to watch this video ok. So, let us begin our lecture.

(Refer Slide Time: 02:06)



So, now we are going to consider the principle of least squares and method of maximum likelihood estimation for the estimation of the parameters β and σ^2. So, you may recall our model was y = X β + ε where y is a n x 1 vector of observation on the response. X is a n x k matrix of explanatory variable in which n observation on each of the k explanatory variable have been obtained, β is a k x 1 vector of regression coefficient and ε is a n x 1 vector of say random errors.

Now, in case if I try to employ the principle of least square, then the least square estimator which we are obtaining for β is given by here b. $b = (X'X)^{-1}X'y.$ So, you can see here X is known to us; y is known to us and so, we can estimate this parameter β And you can see here that this expression this is of order here, X is k by n; X is n x k; X k x n and y here is n x 1. So, b is going to be a vector of order k x 1.

Essentially, what b is doing? This b is going to be something like b_1 b_2,…, b_k and what b is doing? b is trying to estimate the parameter vector β which was β_1, β_2,…, β_k right. So, b_1 is the estimator of β_1, b_2 is the estimator of β_2 and so on. So, you can see here now the estimators are in the form of vectors right.

But when I am saying that estimator of β_1 is b_1 estimator of β_2 is b 2; it does not mean that they are independent right. This is a joint of joint estimator of β_1, β_2, β k right. So, because this is a vector quantity. So, all $b_1, b_2,\ldots,b_k$, they are not really independent in general.

So, next is our estimator of σ^2. Once you have obtained the β vector by b, then we have to replace $\hat{\beta}$ by b and we can write down the $\hat{\sigma}^2$ as (y - X b)' (y - X b). This is the symbol. Here is this is'you know it;'of a matrix or the'of a vector.

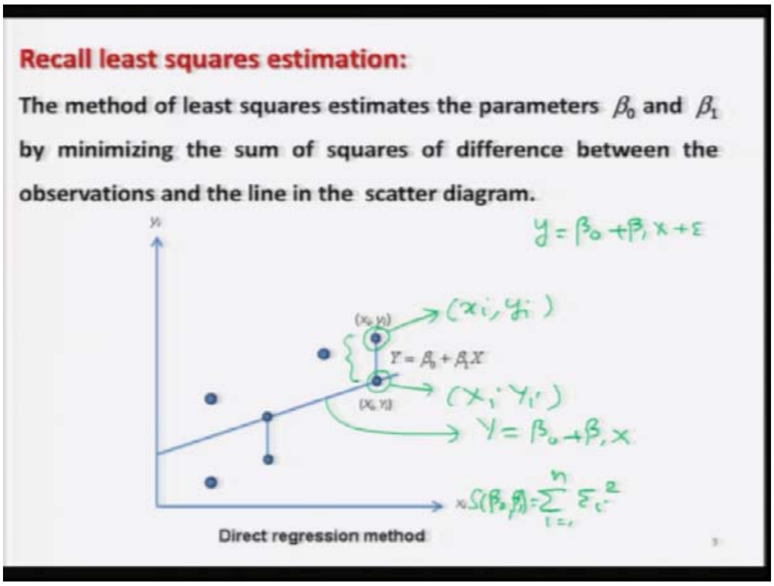
And this is divided by here by n = k. So, n is the number of observations and k is the number of explanatory variables. So, now, you know y, you know X, you know b, you know n, you know k; so you can estimate this σ^2 and the estimated value will be indicated by results we which we are going to consider and yeah, I will try to show you the proof also. But what you have to consider that when you consider the simple linear regression model, then your β vector was something like β_0 and β_1 right and similarly, you had obtained this here $\hat{\beta}$ 0 and $\hat{\beta}$ 1 as b_0 and b_1 respectively.

And when you consider the estimate of σ^2 in the case of simple linear regression model, then this divisor was actually 2; means here k was equal to 2. Why k was equal to 2? Because you were writing the model $\beta_0 + \beta_1$ X as β_0 x 1.

So, that means, you had two explanatory variable; one was variable which always takes the value 1 and say another was your X. So, that is why it was k + 2. So, remember, the point which I am trying to emphasize which sometime create a confusion among the students is that the divisor in the case of estimator of σ^2 is given by 1 upon number of observations = number of explanatory variable.

Now, in any case if the number of explanatory variables are more or less, they are de converted by different formula, different expression; then, we can adjust them accordingly right ok.

(Refer Slide Time: 07:16)

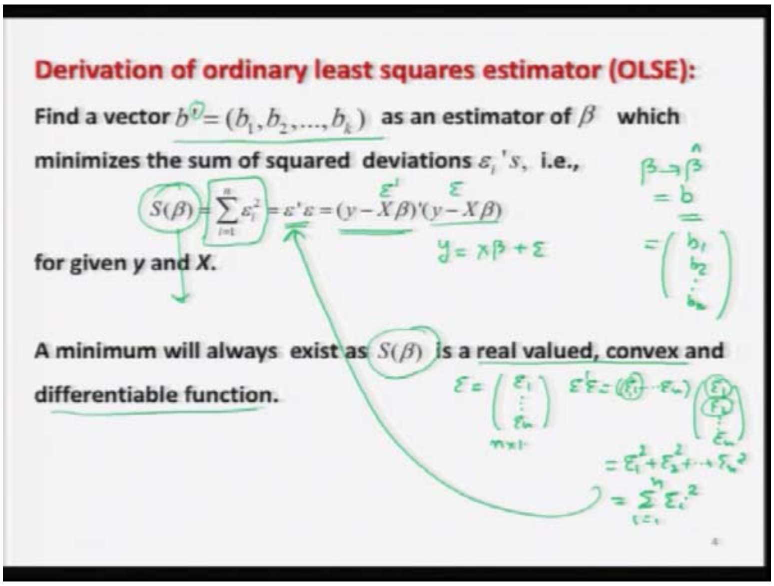


So, now let me try to give you here the basic idea behind the formulation of this estimator and its proof. So, you can recall the least square estimation in the case of simple linear regression model; $y = \beta_0 + \beta_1 X + \varepsilon$ right. So, so you may recall that we had a line, which we expected to fit here, like this and then, we had the observation which are lying on this line.

They were denoted by capital X_i Y_i. All these observations will satisfy the line $y = \beta_0 + \beta_1 X$ and then, we had observed those points at the same value of X and the observed value was obtained somewhere here, which was on the upper side of the line and the observed points were denoted as x_i and y_i right.

So, and then what we had done? We had minimized the difference between the observed and true values or we had minimized the vertical distance between the line and the observations and we had minimized the sum of squares due to error. If you remember, you had minimized $\sum_{i=1}^{n} \varepsilon_i^2$ and you had denoted by $S(\beta_0, \beta_1)$ and so on. You had differentiated use the principle of maxima and minima and then, you had obtained the values of or the estimators of β_0 and β_1.

(Refer Slide Time: 08:53)



So, now the same thing is going to be followed here also. Now, the estimator of β suppose I denote this is my parameter and this is being estimated by $\hat{\beta}$ and the ordinary least square estimator of this β which is $\hat{\beta}$ is going to be denoted by symbol b. I am using here the symbol b, not $\hat{\beta}$ because I want to extend the same symbol that you have used in the case of simple linear regression model right.

So, suppose the estimator of β obtained by the method of least square estimation is obtained and this is denoted by here like this. I am using here because now, I am writing it here as a row vector. So, essentially here this b is something like $b_1, b_2, \ldots, b_k$.

So, I am denoting by b, the k x 1 vector of the ordinary least square estimator of $\beta_1, \beta_2, \ldots, \beta_k$ which are in the form of a vector ok. So, now, essentially, we have to obtain this value of b. So, how to obtain this value of b? We will again use the same principle that we try to minimize the sum of square due to error and sum of square due to error is $\sum_{i=1}^{n} \varepsilon_i^2$.

Now, you have to observe one thing that my ε here is something like ε_1, if you remember $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$. This was your here n x 1 vector. Now, if I try to write down here ε'ε what will be this thing $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ and $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$?

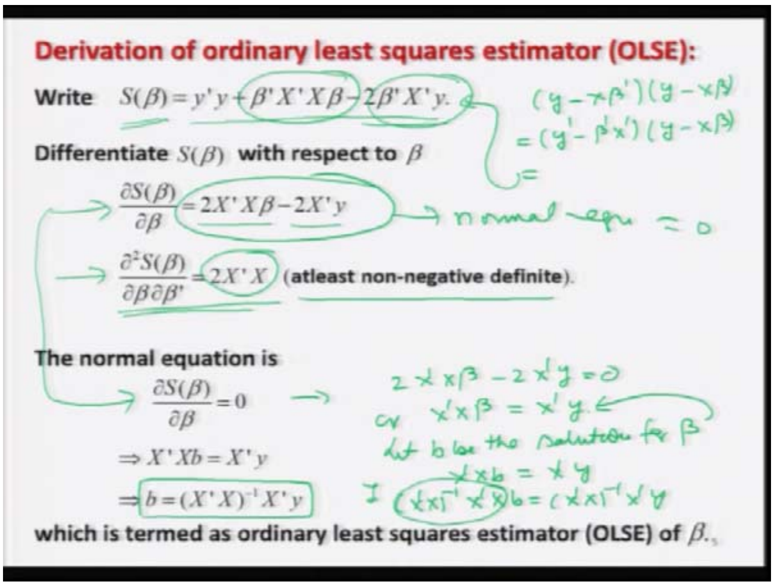
Now, if you try to multiply it, this will be ε_1 into ε_1, that is ε_1^2; then ε_2 into ε_2, that is ε_2^2 and their sum. So, this will basically become $\varepsilon_1^2 + \varepsilon_2^2 + \varepsilon_n^2$. So, this is nothing but your $\sum_{i=1}^{n} \varepsilon_i^2$.

So, what I try to do here? I try to write down the ε transpose ε here as in place of $\sum_{i=1}^{n} \varepsilon_i^2$ and you will see that once I try to write down these quantities in the form of vectors and matrices, then many of my algebra will become very simple to understand and the expressions are easy to found.

So, now, so I will write here like this and now, I am writing ε here as a y - X β right. You remember y = X β + ε and then, I try to write down here this y - X β in place of ε So, this is your here ε' and this is your ε and the values of y and X are given to us.

So, what we try to now do? Our objective is this we would like to minimize this function S(β) which is $\sum_{i=1}^{n} \varepsilon_i^2$ and we would like to obtain the value of β such that the value of β minimizes this sum of squared deviations right. So, the question comes what is the guarantee that the minimum will always exist. The minimum will always exist because as β is a real valued convex and differentiable function right.

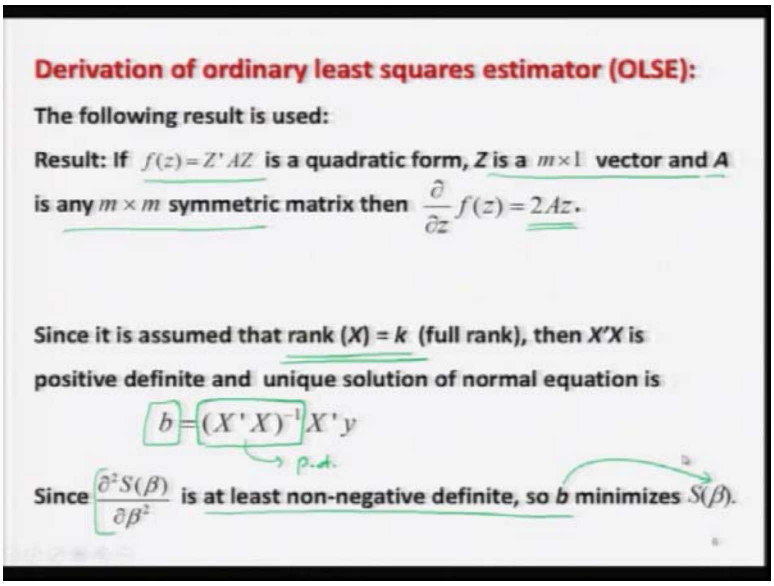
(Refer Slide Time: 12:27)



So, what I can do? Exactly on the same lines as we had used the method of maximum minima in case of simple linear regression model, here also I will use the same principle. So, I try to expand this (y - X β)' (y - X β) which is equal to here y'- β'X' and y - X β So, I try to multiply it and whatever is my outcome this is given here.

So, you will see here that S(β) will come out to be like this y'y + β'X'X β = 2 β'X'y. Now, what we have to do? I simply have to use the principle of least square. So, I try to differentiate this S(β) with respect to β vector. Now, remember one thing, here we are using the differentiation with respect to a vector.

(Refer Slide Time: 13:21)



So, in order to do it, first I try to show you the result which I have used with which is available in all the results related to the linear estimation. So, if there is a quadratic function, say f(z). and my variable here is capital Z. So, if the quadratic function is of the form Z'AZ, where Z is a m x 1 vector of real values and A is any m x m symmetric matrix.

Then, the $\frac{\partial}{\partial z} f(z) = 2Az$ right. So, this is the same result which I have used here and after, so then. So, the $\frac{\partial S(\beta)}{\partial \beta} = 2X'X\beta - 2X'y$ and the $\frac{\partial^2 S(\beta)}{\partial \beta \partial \beta'} = 2X'X$.

And yeah, means just I am trying to obtain the second differential also because that will be used to determine whether the obtained equation is providing the maxima or minima. So, just to be; so, I am just finding out the $X'X\beta - X'y = 0$.

So, you can see that this quantity is independent of β and this X'X matrix, this is going to be at least non-negative definite always right. So, that will ensure that whatever the value of β which we are going to obtain by solving this normal equation, this is my here normal equation when I try to put it equal to 0 right. Whatever the solution, I obtain from this equation that is going to minimize the function S(β) that is the sum of squared deviations due to random error, right.

So, if you try to put this partial derivative of S(β) with respect to β equal to 0, then we get the normal equation and if you try to solve it, this is trying to give you here like this 2X'X β - 2X'y = 0 or I can write down here X'X β = X'y. So, I can now say here, let b be the solution of solution for β from this equation. So, I can write down that solution will satisfy this equation. So, X'X b = X'y. Now, I need to find out the value of b.

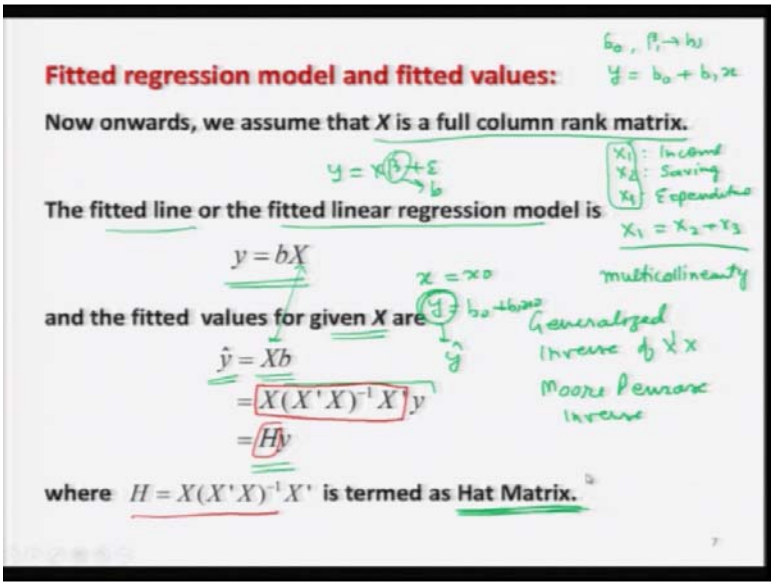
So, what I can do? I can pre multiply by $(X'X)^{-1}$ on both the sides. So, this will become here like this X'y. So, now this part becomes here identity matrix I and so, this b will come out to be like this; $(X'X)^{-1}X'y$ which we have obtained which is called as an ordinary least square estimator right ok.

So, now, if you try to look at the structure of this ordinary least square estimator, you can see here this is involving a quantity $(X'X)^{-1}$. Now, under what type of condition, this inverse will exist? So, we know that this inverse is going to exist only when rank of X matrix is equal to k, that is it is full column rank.

Because only under that condition this X'X matrix will become a positive definite matrix right and that is the reason if you remember, we had made one assumption that rank of X = k, when we made different assumption for the multiple linear regression model. So, by assuming that rank of X = k, we can ensure that X'X matrix is a positive definite matrix and hence, the inverse will always exist.

And thus, the least square estimator of β can be obtained by here b and we already have seen that the second partial derivative of S(β) with respect to β this is also at least non negative definite. So, now, I can assure that b minimizes S(β) So, this is our ordinary least square estimation.

(Refer Slide Time: 18:23)



So, now onwards, whenever we are trying to do anything in this course, we will always assume that X is a full column rank matrix. Well, just for your information, I can also share this information that it is possible that X may not be a full column rank matrix and yeah, this also happens.

That means, all the columns of X_1 , X_2,,., X_k are not independent; they have got certain relationship. For example, if you try to take here 3 variables here X_1 , X_2 and X_3. So, X_1 is suppose the income and X 2 here is saving from that income and X 3 here is expenditure right from this income. Then, you know that $X_1 = X_2 + X_3$ right, ok.

So, now, under this thing this X_1 X_2, X_3, three variables which you are taking here k equal to 3, all these variables will not be independent and the rank will be only here equal to 2. So, in that case, X will not remain as a full column rank matrix and this problem is actually termed as problem of multi-collinearity.

So, I am not saying that this problem does not occur in practice. So, I am assuring you yes, it may happen that the data may have such an issue. Under those conditions, we have two options that either we try to use the tools for the multi-collinearity or we can obtain the generalized inverse of X'X.

So, the generalized inverse of X'X is obtained in a different way than the unique inverse of X'X, I am not going into those detail; but I am just informing you that it is possibility. That this possibility is there, but definitely the generalized inverse can be obtained in different ways and except one which is called as Moore Penrose inverse; Moore Penrose inverse, all other inverses are not unique.

So, it is possible that for the same data set, different people may obtain different values of β using the generalized inverse. Well, I am not going into those detail, but my objective was simply to inform you, to update you. Now, you remember that in the case of simple linear regression model, once we had obtained the value of β_0 as b_0 and value of β_1 as b_1 then, we had obtained the fitted model just by replacing β_0 by b_0 and β_1 by a b_1 x.

Now, the same thing, I try to do in the case of multiple linear regression model. Now, you see now you cannot blame me that I am not explaining anything because I already had told you during the lecture of simple linear regression model, that please try to understand these concepts very carefully because they because I will be using them there directly right.

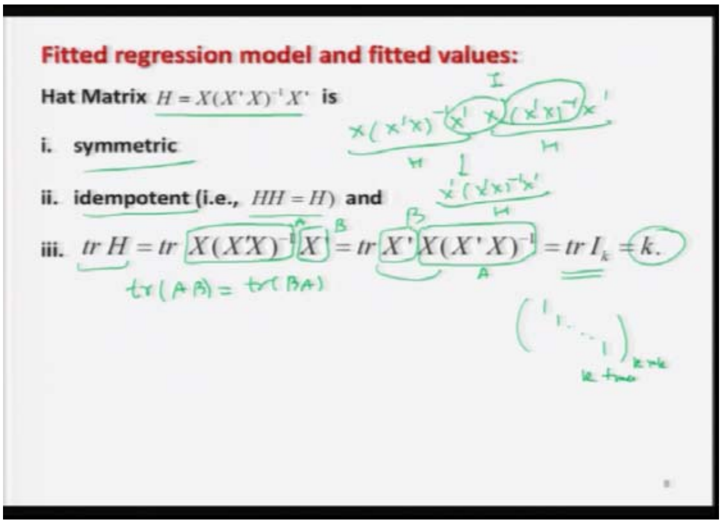
So, working on the same line, if you remember we had simply replaced the estimated value of the parameter in case of the or in place of the true parameter. So, same thing I am trying to do it here. So, our model was $y = X\beta + \varepsilon$ Now, I am trying to estimate this β by b.

So, I can replace this β by b and I can write down y = b X. So, this is our fitted line or this is the fitted linear regression model right. And after this, if you remember in the case of simple linear regression model, we had taken the values of x and then, using the fitted model, we had replaced the value of x say by x_0 in the fitted model $y = b_0 + b_1 x_0$ and this value was called as fitted value, $\hat{y}$.

You can recall that we have taken several examples to explain you. So, now, the same concept, I am trying to extend it here and the fitted values for a given of given X here, they are denoted by here $\hat{y}$ and they will be X b right. So, I am using the same X over here; but now I am trying to obtain the value of y which are obtained through the fitted multiple linear regression model.

And if you try to do it, if I try to replace here the value of b which is $(X'X)^{-1}X'y$ like this, then I can write this $\hat{y}$ something like this H y. So, what I am doing here that I am writing this matrix, this component of this expression $X(X'X)^{-1}X' = H$, right. Actually, this H is a very important matrix in regression analysis; particularly, in multiple linear regression analysis, this matrix is called as hat matrix right. This is called as a hat matrix.

(Refer Slide Time: 23:35)



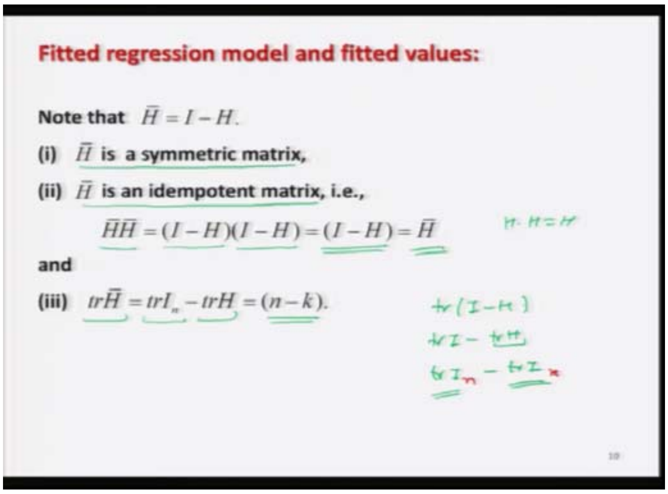
And hat matrix has a couple of properties which helps us in making very useful statistical inferences. This matrix, H which is hat matrix is symmetric; this is idempotent. That means, if you try to multiply H into H, you can see here if you try to see write down here $X(X'X)^{-1}X'$which is your here H and if you try to repeat it here again, $X(X'X)^{-1}X'$ this is here another H.

So, you can see here X'X into X'X whole inverse this becomes here I. So, now, what are you getting here in the next step? $X(X'X)^{-1}X'$and this is same as here H. And similarly, if you try to find out the trace of the matrix H, then this is obtained by using the result. If you remember there is a result in matrix theory, which is something like trace of A B = trace of B A right.

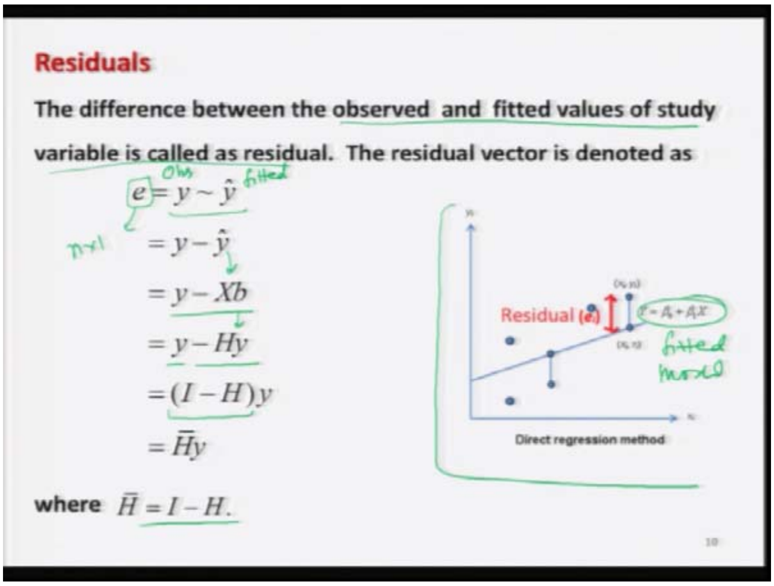
So, in those cases, what I am trying to do here that I am trying to find out the trace of here matrix H. So, I write it here say here H and then I try to write down this matrix here as say A and this matrix here say B and then, I try to use this result trace of A B = trace of B A. So, you can see here, this is your here B and this was your here A.

So, now, you can see here that this matrix here is X'X. So, X'X $(X'X)^{-1}$, this is identity matrix of order k by k. So, we have a trace of I_k and trace of I_k is something like you are trying to write down 1, 1, 1, 1, 1 here k times. So, k into k matrix. So, this is going to be 1 + 1 + 1 … k times which is k.

(Refer Slide Time: 25:39)



(Refer Slide Time: 25:41)



So, these are very important result that you will see that we will be using them and you will also use them in further lectures. Now, from the results that we obtained during the simple linear regression model, you had obtained the residuals. If you remember we had constructed this type of figure in which we fitted this model right.

This is now under this circumstances, this is your here fitted model and using the value of using the given value of X, you obtain the value of y. So, and the difference between the observed values which is here y and the values which you have obtained after fitting the model the which are your fitted values.

So, the difference between the observed and fitted values of the variable is called as residuals. So, now earlier, you had residuals which were scalars right because you are trying to obtain for every X_i, here also they are going to be scalars; but I can express them very nicely in a form using the vectors and matrices.

So, if you try to see, I am using the same result which I used earlier. So, the residual vector e is now defined as y - $\hat{y}$; the only difference with respect to the simple linear regression model is that now this is a n x 1 vector right. So, means I am now fixing my notation that this difference is y - $\hat{y}$.

So, now, $\hat{y}$ here is y is here y = X b. So, y - X b can be written as y - this X b can be written as H y. So, this is here y - H y. So, I can take out this y common on the right hand side and so this becomes I - H y. Now, I am using another notation $\bar{H}$. $\bar{H}$ is denoting this matrix I – H, right. So, you can see here $\bar{H}$ is a matrix which is related to your hat matrix and this hat matrix also has the similar properties what we had in the case of H.

Yeah, for example, you can see here these are the property in case of H. Now, in the case of $\bar{H}$ also, we have a similar properties $\bar{H}$ is a symmetric matrix and $\bar{H}$ is also an idempotent matrix just like H and you can see here what I have shown you here is bar into $\bar{H}$ is I = H into I - H and if you try to open it and use the property that H into H = H, you can see here this will again come out to be I - H which is same as $\bar{H}$.

And if you try to find out here trace of H matrix so, this is simply trace of the matrix, identity matrix of order n - trace of H. Trace of H you already have obtained. So, so, trace of this I_n - trace of H is simply your here n - k right. Sometimes, you have to be careful people try to make a mistake that you try to find out the trace of I - H.

And since in the case of they try to write out trace of I and minus trace of H and if you remember that in the case of finding out the trace of here matrix H, we also had here say trace of I right. So, people try to people sometime get confused that this is also trace of I and this is also trace of I; but what you have to keep in mind? This was an identity matrix of order n and now, this is a matrix of this is an identity matrix of order k.

So, do not get confused that trace of I = trace of I is 0 which is wrong. So, this is equal to n = k ok. So, now, we come to an end in this lecture. My idea was very simple. Using the concept that we had earlier discussed in the case of simple linear regression model, I wanted to extend them to a multiple linear regression model; the only thing is this there are small differences in the algebra, the way we have obtained the expression; otherwise, the concepts and the definition they remain the same.

And you will see that these things are going to work exactly on the same line at as b_0 and b_1 worked in the case of simple linear regression model. So, now the next question is you have obtained this value, you obtain the model, you have obtained the fitted values, you have obtained the residuals; now, how to obtain them on R console in the R software that is our next question.

So, you will see that I am going to use the same commands which I used in the case of simple linear regression model right. So, that lecture is going to be very simple and straight forward provided. You come after a quick revision of the lecture in the chapter of simple linear regression model, where we had obtained the regression coefficient, fitted values and residuals in the R console.

So, you please have a look and I will see you in the next lecture; till then, good bye.