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 - 50
Multiple Linear Regression Analysis
Analysis of Variance and Implementation in R Software

Hello friends, welcome to the course Essentials of Data Science with R Software-2, where we are going to learn about the topics of Sampling Theory and Linear Regression Analysis. And, in this lecture we will continue with the module on Linear Regression Analysis, we will consider the Multiple Linear Regression Analysis with R Software.

So, you can recall that in the earlier lecture we started a discussion on test of hypothesis and confidence interval. And, we had constructed the test of hypothesis for individual regression coefficients. Now, in this lecture I will continue and I will try to construct the analysis of variance, which is also a test of hypothesis for the equality of more than two regression coefficients, ok.

So, first thing I can share with you from my experience that sometime people get really afraid of the word analysis of variance. But, believe me this is a very simple thing, what we are trying to do. For example, if I take a simple example which we have considered up to now that we have the marks of the students. So, you know that the marks of a student depend on the performance in the exam and the performance in the exam depends on many factor.

For example, how many hours are student has studied, how many assignments the student has submitted, how many hours the student has spent in the library and so on. So, all these variables together they are going to contribute in the variability of the response, variability in the marks.

So, these factors are contributing individually, which all together produces the variation in the values of response variable. So, now, the objective is looking at the total variation in the response variable, I want to divide the total variation into different components,

such that every component is indicating the contribution of the variation due to individual variable.

So, for example, if you say the marks; marks are going to suppose depend on number of hours of study, number of assignments, and numbers of hours of play. Beside those things there will be some random factor. So, we can assume here, that essentially the total variation depends on two parts, broadly two parts; one which we are trying to control through our model $y = X \beta$ and random error ε.

So, broadly there are two groups in which the total variation of y can be divided. One group of variation, which is controlled by the variables which are in our control and second group of variables, which are constituting the random error component, which are not in our control.

Now, there can be another factor, that when we are trying to take the variables, which are in our control, then there are more than one independent variable, which are affecting the marks. That, what is the contribution of each and every variable also. So, if you try to see what are we trying to do?

We are simply trying to consider the variability in the response variable and we are trying to analyze it. How we are trying to analyze it? We are trying to divide the total variation into different components such that, every component is explaining the variation contribute contributed by that respective component. And, that is why this is called as analysis of variant.

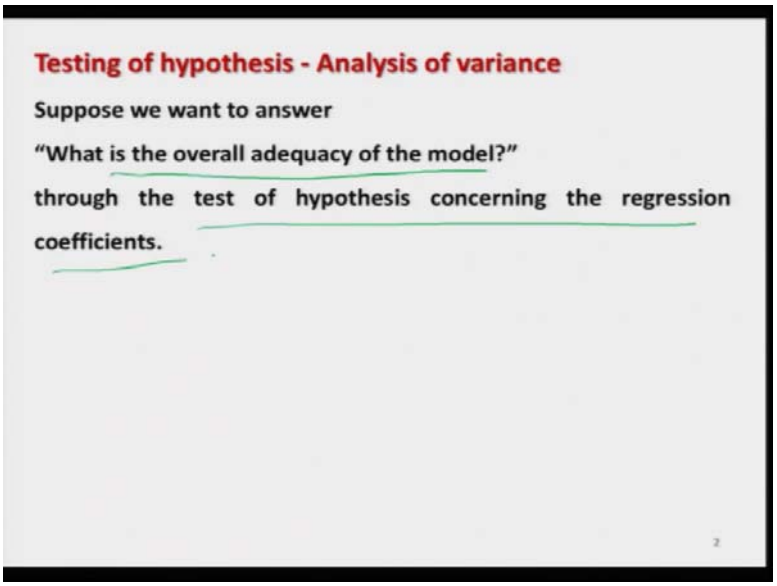
So; obviously, if there is a variable which is contributing very less do you think, that is it going to be an important variable certainly not. Means well, it is contributing so, little that even if we try to ignore it that will not make much impact on the analysis. But on the other hand there will be some variable which are very important.

So, we expect that their contribution in the total variability will be more and hence they are important variable and we cannot ignore them. For example, if I say marks are dependent on the number of hours of study, that we know, that is an important variable. So, if you try to see the total variation in the marks.

And, if you try to see the proportion of the variable total number of hours of studies, that will be very much. For example, you have heard in your family that, if somebody gets a lower marks, the parents first of all may make the first complaint that possibly you have not studied hard. That means, you have not contributed sufficient number of hours in your study.

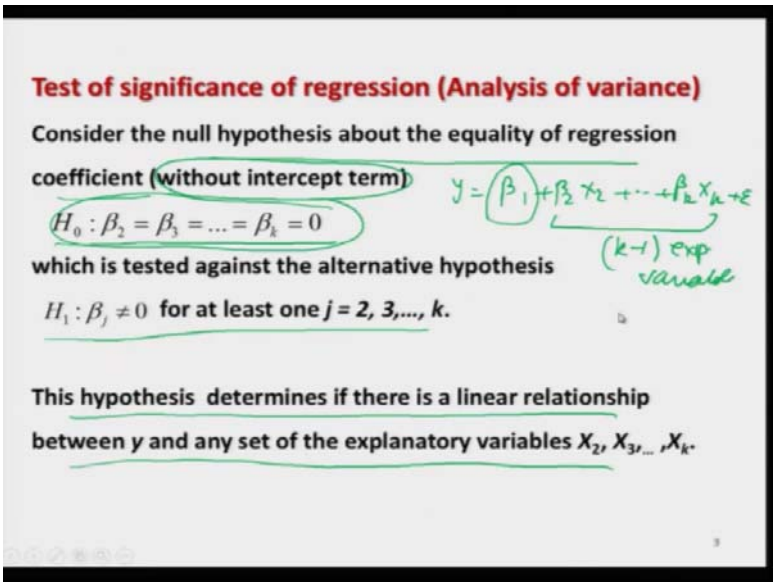
So, this is analysis of variance. So, you can see now, how important it is and I assure you this is a very simple topic. I have tried to make my slides in little bit more detail, but I am sure that you would not mind spending 5 to 10 minutes more, if you can understand it better. So, let us begin our lecture, ok.

(Refer Slide Time: 05:38)



So, suppose we want to answer the question like, what is the overall adequacy of the model through the test of hypothesis, concerning the regression coefficients? So, this can be done by analysis of variance. So, first of all different people have different types of concepts about the analysis of variance.

(Refer Slide Time: 05:52)



But, the most simple thing, the more simple approach through which you can actually thing is the following. Just like that, test of hypothesis $H_0 : \beta_j = 0$, that is for the individual regression coefficient analysis of variance is also a test of hypothesis. Where, we consider the null hypothesis about the equality of regression coefficient.

Remember one thing I am not talking of intercept term, please remember one thing once again, I am considering here the hypothesis about the equality of regression coefficients only without any intercept term. So, I am trying to write down here my model here as, see here $y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + \varepsilon$. So, you can see here there are only k - 1 explanatory variables. And, means why I am trying to do it why I am not including the term like β_0.

Here that will be clear in the next lecture, when we try to address the aspect of goodness of fit. Because, the goodness of fit and this analysis of variance, they are interrelated to each other and there are certain conditions, which are required in the goodness of fit. Because of those reasons, I am considering here an intercept term in the model and the exact reason I will be able to explain you in the next lecture, ok.

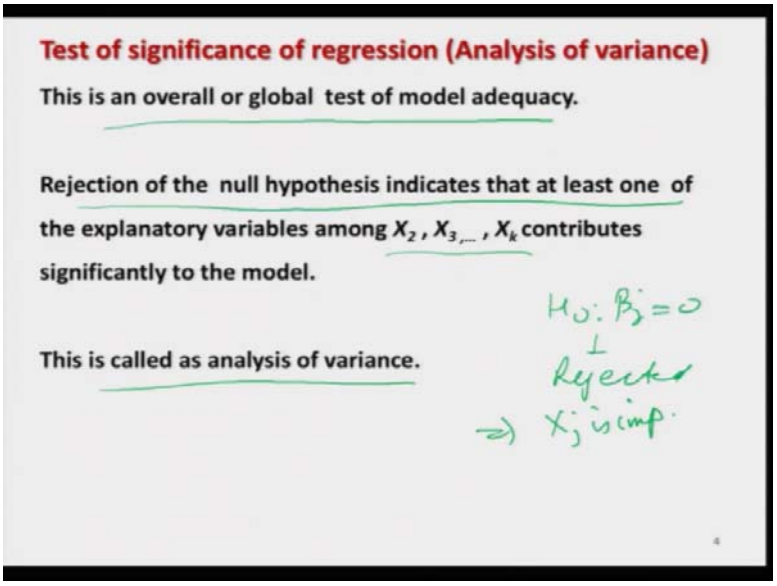
So, now we consider the null hypothesis $\beta_2 = \beta_3 = \ldots \beta_k = 0$. And, the alternative here is this suppose this hypothesis can be rejected, even if there is 1 pair of β_js which are not

equal or any of the $\beta_j \neq 0$. So, the alternative can be framed as H_1 $\beta_j \neq 0$ for at least 1 j going from 2, 3 up to k.

So, if any of the regression coefficient is not equal to 0, then the entire null hypothesis is going to be rejected. And; obviously, if you try to see that, if H_0 $\beta_2 = \beta_3 = \beta_k = 0$ is accepted, what will happen you are simply trying to see that all the variables are not important they are equal to 0. And, so, the only intercept term is there, which is contribution, which is contributing in the response.

So, there will be only intercept term and random error, right. And, actually this hypothesis determines, if there is a linear relationship between y and any set of explanatory variable X_2, X_3,…, X_k or not. So, this will also help you in deciding with the linearity of y with respect to X_1, X_2,…, X_k exactly in the same way, that we have done through the correlation plot and multiple scatter diagrams, ok. So, this is actually an overall or global test of model adequacy, right.

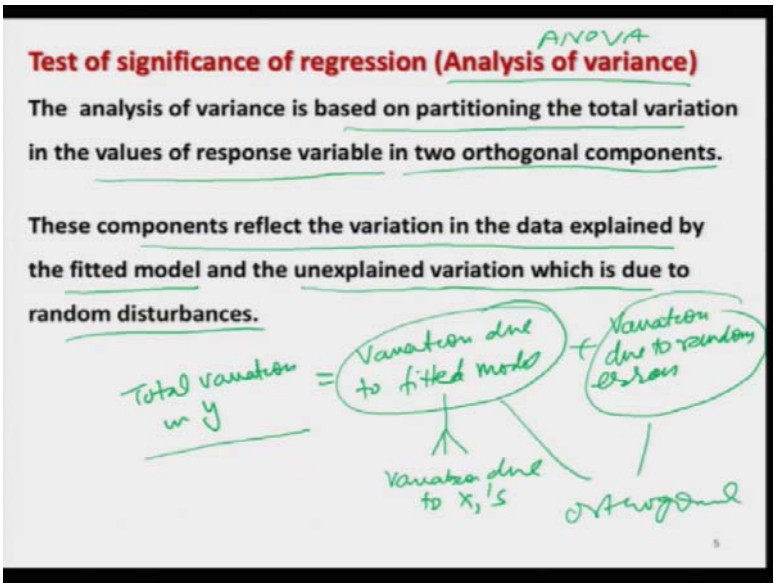
(Refer Slide Time: 09:04)



And, if I say the hypothesis is rejected that the null hypothesis is rejected, that would indicate that there is at least 1 explanatory variable among X_2, X_2, X_3,…, X_k, which is contributing significantly towards the model. And, that is why this hypothesis is getting rejected, because you remember that when we had discussed the test of hypothesis for H_0 $\beta_j = 0$, if this β_j 0 is rejected.

This implies that the corresponding variable X_j is an important variable and it has to be in the model, right. So, similar type of conclusion can be taken here, ok. And, this is what is called as analysis of variance.

(Refer Slide Time: 09:55)



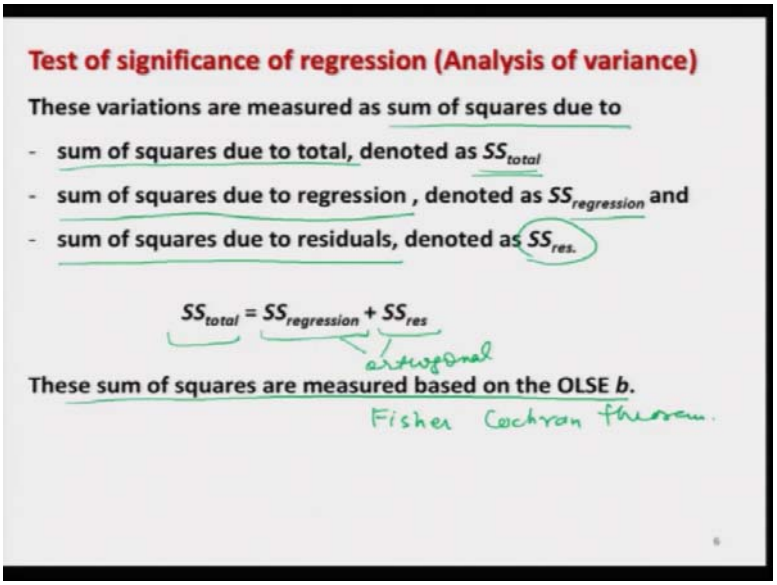
Well, this analysis of variance is based on partitioning the total variation in the values of response variable in two orthogonal components broadly, right. And, these two orthogonal components reflect the variation in the data, which is explained by the fitted model. And, number 2 the second part, which explain the variation in the data due to random error or the unexplained variation, which is due to the disturbance term or random errors, right.

So, essentially the total variation in y in y is contributed into is divided into two parts, say variation due to fitted model and variation due to random errors, right. And, this actually I will show you that in that when you try to employ this, ANOVA in this analysis of variance in R. Then this variation due to the fitted model is further partitioned into variation due to X_js, right. And, one thing this analysis of variance is briefly called as ANOVA, A N O V A that is one of the popular short name of analysis of variance.

So, the total variation is partitioned into two components; one is the variation due to the fitted model and another variation due to the random error, and both of them are

orthogonal, right. They are mutually orthogonal and if there are more than components, then all components are independent of each other.

(Refer Slide Time: 11:53)



So, this variation now actually these variations are measures are measured in terms of sum of squares. And, this sum of squares there are different types of sums sum of squares which are involved, which is one is the total variation in the values of y, which is indicated by sum of square due to total and this is denoted at SS total.

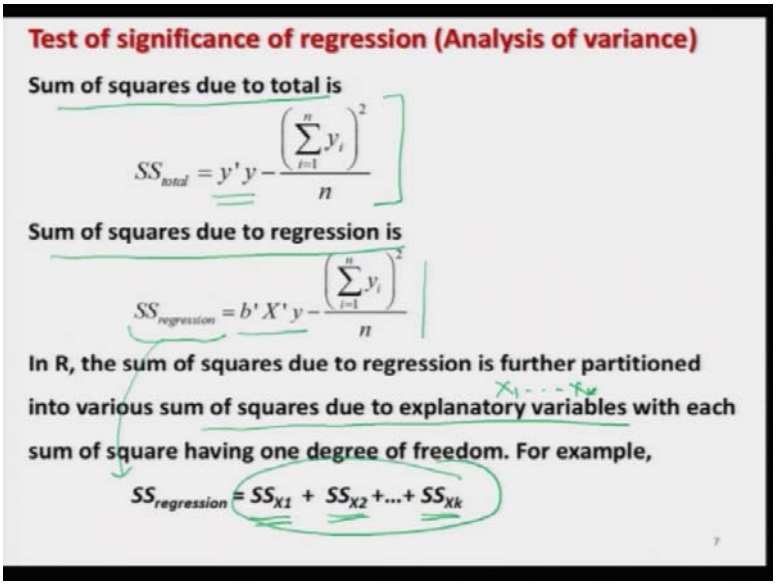
And, then the sum of square due to the fitted model, which is called a sum of square due to regression this is denoted as SS regression. And, the third component will be the sum of square due to the random errors. So, this is actually the sum of square due to residuals and which is indicated by SS_{res}, ok.

So, you can see here I can write that the sum of square due to total is partitioned into two parts some of squares due to regression and sum of square due to residuals and both of them are orthogonal to each other, right. Actually, here just for your information we try to use the Fisher Cochran theorem. Well those who are from statistics they might be knowing, those who are those who do not know they need not to bother about it they say just for a information.

And, these sum of squares they are measured using the ordinary least square estimator of β which is b, right. If you try to use any other estimator that will create different types of

problem. So, whatever analysis of variance we are considering here, that is based on the ordinary least square estimation.

(Refer Slide Time: 13:32)



So, now there are two aspects; one aspect is the complete derivation of this analysis of variance. Well, I personally feel that there is no problem in doing it, but those students who are those candidates, who are not from the main statistics background, it will become difficult for them to understand and possibly as far as that data sciences is concerned.

This ANOVA is going to be more useful actually. So, from so, keeping in mind the utility of ANOVA, I am skipping the proof and I am giving you the details of the steps and details of the expressions, which are used in the analysis of variance. Those who are interested I recommend them that they can go to the lectures on regression analysis well. I had given some lectures and which are available on the NPTEL site in the PDF format.

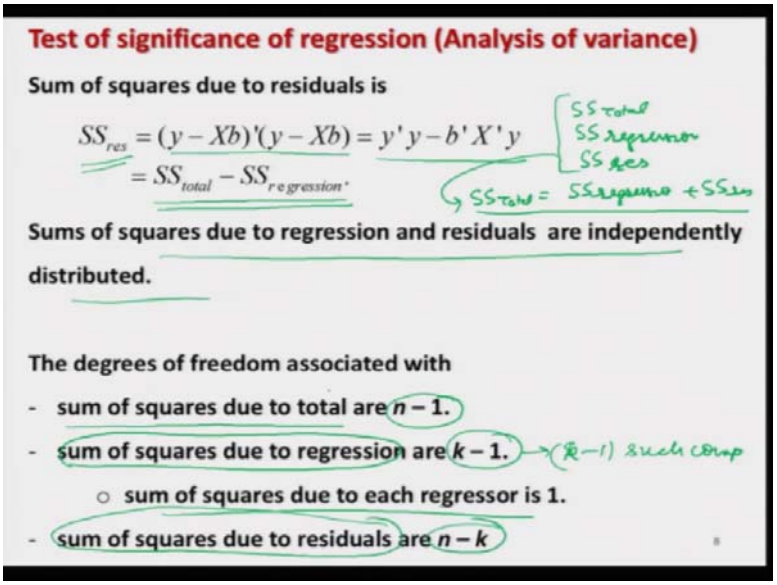
So, you can look into those lecture notes, where I have given the complete detail or complete proof of the analysis of variance step by step, right. Now, we come back to our this lecture. So, the sum of square due to total this is obtained by this expression, right.

So, you can see here $SS_{total} = y'y - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$. And, the sum of square due to regression, that is obtained by the expression $SS_{regression} = b'X'y - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$, right.

And, you will see that, when we try to implement this concept in R then this sum of square due to regression can be further partitioned into the sum of square due to explanatory variables. Since, there are k explanatory variables X_1, X_2,…, X_k. So, this sum of square due to regression can be further partitioned into, sum of square due to X_1, sum of square due to X_2, and sum of square due to X_k.

So, this individual sum of squares are going to indicate the variation contributed by the individual explanatory variables and each of this sum of square will carry 1 degrees of freedom. What is the concept of degrees of freedom? This I will try to show you, but here at least I can inform you that these sum of squares they will be carrying 1 degrees of freedom, ok.

(Refer Slide Time: 15:58)



So, now the third component of the analysis of variance, this is sum of square due to residual SS res. This is the same thing what we have done in the case of simple linear

regression model also many times, this is $(y - Xb)'(y - Xb) = y'y - b'X'y$. And, this is actually also obtained by the subtraction. That is you try to subtract sum of square due to regression from the sum of square due to total.

So, it is advisable, in the case of analysis of variance that you have three component. SS total, $SS_{regression}$, and SS_{error} or say a res. And, there is a relationship $SS_{total} = SS_{regression} + SS_{res}$. So, we recommend that you try to compute any two components from the data directly using the expressions using the formula and try to obtain the third component by subtraction using this relationship.

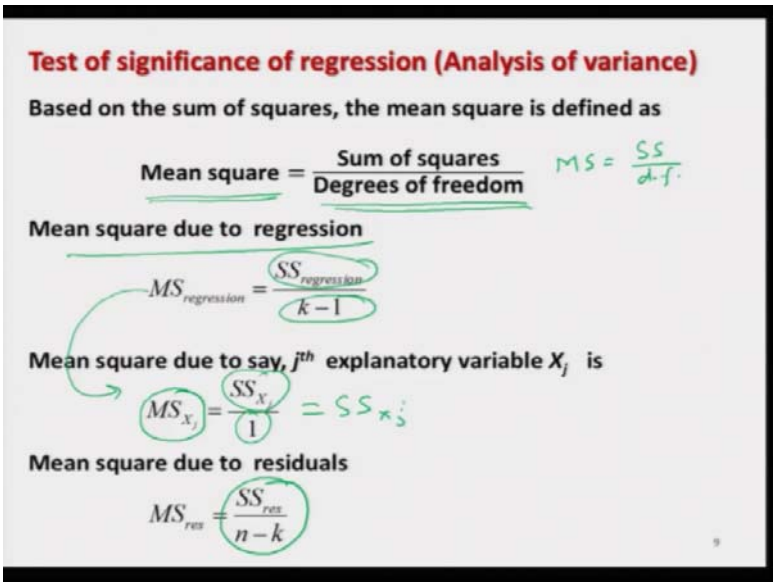
The reason is that the reason behind this computation that you try to obtain the third component by subtraction is the Fischer Cochran theorem, right. Because in the development of analysis of variance we try to use the Fisher Cochran theorem in finding out the distribution of the sum of squares. And, so, the, so, for the validity of the Fisher Cochran theorem, it is important that you try to compute two expression directly using the formula or the expression and third expression has to be found using this relationship.

So, you may recall that when we had considered such sum of squares in the case of simple linear regression model, then we also had associated chi square distribution. And, the chi squared distribution was specified by the name chi square and the associated degrees of freedom.

So, similarly in the case of analysis of variance in multiple linear regression model, we have this several types of sum of square and this sum of square also carries degrees of freedom. So, the degrees of freedom which are carried or which are associated with the sum of square due to total are n - 1.

And, the degrees of freedom associated with the sum of square due to regression they are k - 1, right. And, these k - 1 sum of squares they are further partitioned into sum of square due to each regressor. So, there will be say k - 1 such components. And, each component will indicate the contribution due to an individual variable X_j and each of the sum of square will carry a degrees of freedom 1. And, finally, the last sum of square due to residuals this will carry a degrees of freedom n - k.

(Refer Slide Time: 19:14)



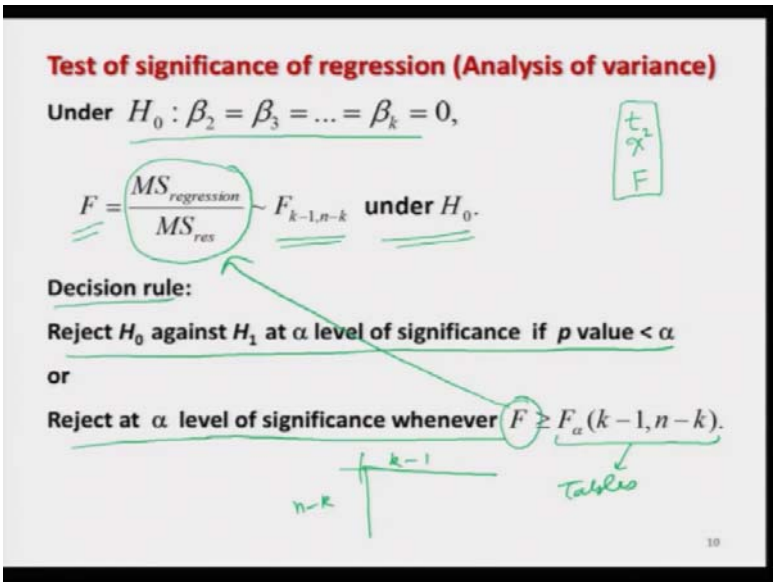
So, this is how we try to move forward and after that we try to define the quantity mean square. Mean square is actually divided in is defined in general as sum of squares divided by degrees of freedom. So, it is actually briefly also indicated as MS. So, MS is equal to SS upon degrees of freedom.

So, when you are trying to take different types of components. So, you will have 1 mean square corresponding to each of the sum of a square. For example, you have considered here the mean square due to regression, mean square due to say this error, random error, and mean square due to regression. And, to, ok total, but that is not actually used.

So, essentially we try to define here two mean square - one for the regression and one for the residual. So, the mean square due to regression is defined; obviously, by say sum of square due to regression and divided by the degrees of freedom. And, similarly the mean square due to residual is defined by SS_{res} divided by degrees of freedom. You can see here and for this mean square due to regression. Since, the sum of square due to regression has been further partitioned into individual independent variable.

So, for individual explanatory variable we can define the corresponding mean square as say mean square due to X_j as sum of square due to X_j divided by 1. So, this is essentially the same as sum of square due to X_j, right, ok.

(Refer Slide Time: 20:54)



So, now the we define the test statistics for testing the null hypothesis under analysis of variance, which is $H_0 : \beta_2 = \beta_3 = \ldots \beta_k = 0$. And, the test statistics is defined as the ratio of mean square due to regression and mean square due to residuals. And, this is indicated by F, which follows a F distribution with k - 1 and n - k degrees of freedom, when H_0 is true that is under H_0, right.

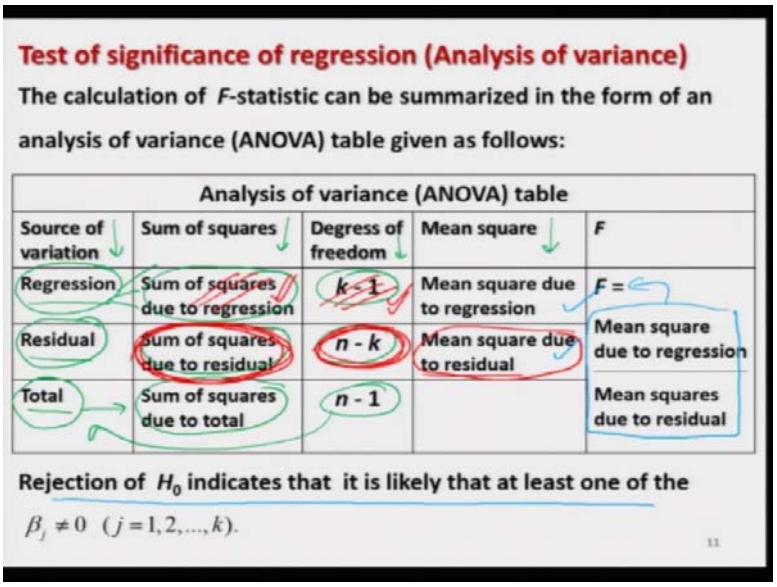
So, you have seen now we have used our distribution t we have used a distribution chi square and now we are going to use here distribution F. And, these are our three popular sampling distribution and that is why they are called as a sampling distribution, because they are trying to indicate the distribution of a sample statistic like this one, ok.

So, now, the story finishes here. Now, you simply have to define the decision rule, which is the same what you have defined in the earlier case, that reject H_0 against H_1 at α level of significance if p value is smaller than α, or if you want to go with the go with the critical value.

So, what you have to keep in mind that reject H_0 at α level of significance whenever the calculated value of F which you have obtained from here is greater than that, tabulated value of here F. This value can be obtained for the tables.

The tables are available and this tables are something like this two way tables, once there will be a degrees of freedom k - 1, and say on the column there will be degrees of freedom n - k. And, you can choose with the proper value, but definitely in our case we are going to use the R software. So, we need not to worry about it.

(Refer Slide Time: 22:48)



So, finally, all these calculations whatever we have done. They can be combined in a table which is called an ANOVA table, analysis of variance table. And, my idea of giving you all those expression was that you must understand what is happening in this analysis of variance, because in the case of a software you will get only this table.

So, you can see here that this table has a very nice structure very well structured table. So, first column is the source of variation, that how the variation is coming into the response variable? So, I can broadly define it into divide it into three components; regression, residual and total. So, regression is going to contribute the variation, which is being explained by the fitted regression model, residual is going to explain the variation due to the random error component. And, total is going to give us the sum of square due to total.

Yes, this sum of square due to regression will be further classified into different orthogonal components, which are corresponding to individual variables. But, here I

would like to make a point here as long as we are using here the R software this sum of square due to regression is partitioned.

But, in some other software also you will find that this sum of square is not partitioned and it is given simply as sum of square due to regression. So, please do not get confused. After, this corresponding to each source of variation we have to find the sum of squares. So, corresponding to regression we have sum of square due to regression, corresponding to residual, we have sum of square due to residuals, and regarding total we have sum of square due to total.

And, then the third column is degrees of freedom, which are here the associated degrees of freedom with sum of square due to regression k - 1, n - k degrees of freedom associated with sum of square due to residuals. And, n - 1 degrees of freedom associated with the sum of square due to total.

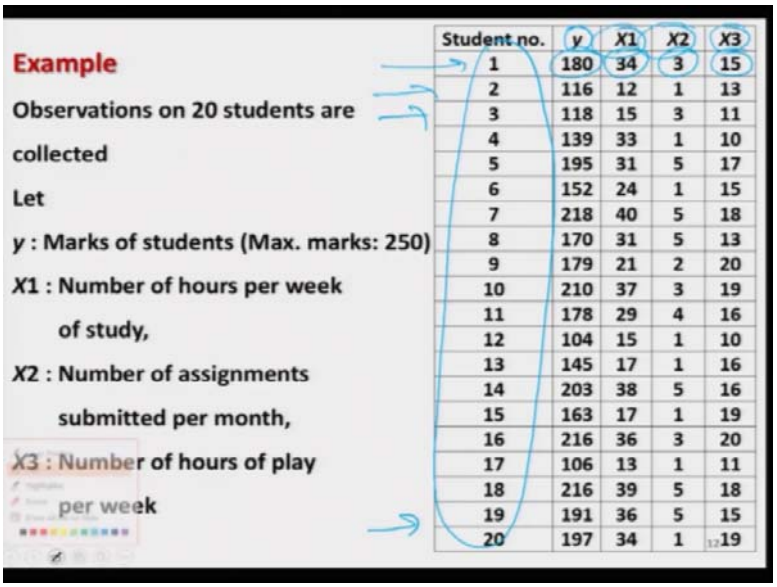
And, after this we create here 1 more column which is indicated the mean square. So, mean square due to regression, which is simply the ratio of these two column try to observe my here pen, red in color reds. If you try to see here I am trying to highlight these two. So, you simply try to obtain the sum of squares from this column, try to obtain degrees of freedom from here and take the ratio.

So, you can obtain the mean square due to regression. Similarly, for the mean square due to residual, you try to obtain the sum of square due to residual from here degrees of freedom from here and try to take the ratio and you will obtain the mean square due to residual. Once, you have obtained the mean square due to due to regression and mean square due to residuals, you simply have to take their ratio. And, you will find the value of F statistic.

And, based on that you can decide whether this is accepted or not by looking in the at the p values. So, when H_0 is rejected this will indicate that it is likely that at is 1 of the $\beta_j \neq 0$. And, actually that is a good thing for us no experimenter would not try to do a modeling in which all the variables are not important. It is possible that out of a group of large number of variables some important variables are there and maybe some unimportant variables are there.

But, if definitely if the experiment has been started in which most of the variables are not significant, then possibly 1 has to lo, ok into the experiment itself, that what type of data has been collected well that will be my view.

(Refer Slide Time: 26:39)

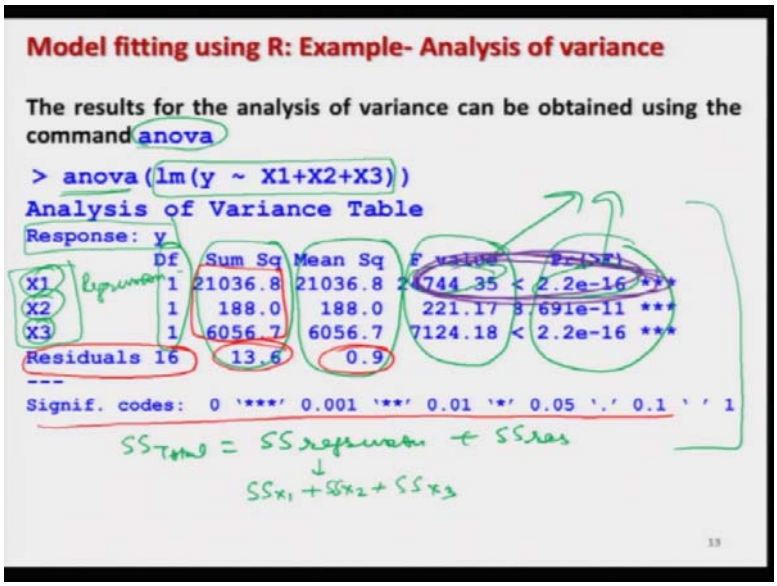


Student no.	y	X1	X2	X3
1	180	34	3	15
2	116	12	1	13
3	118	15	3	11
4	139	33	1	10
5	195	31	5	17
6	152	24	1	15
7	218	40	5	18
8	170	31	5	13
9	179	21	2	20
10	210	37	3	19
11	178	29	4	16
12	104	15	1	10
13	145	17	1	16
14	203	38	5	16
15	163	17	1	19
16	216	36	3	20
17	106	13	1	11
18	216	39	5	18
19	191	36	5	15
20	197	34	1	19

Now, I try to show you how you can implement it on the R software? So, I am using here the same data set, which I used in the earlier lectures, that we have a data of 20 students. In which we have collected their marks, which are indicated by y and then these marks are supposed to be dependent on 3 variables. Say, X1, X2, X3, which are the number of hours per week study, number of assignments submitted per month, and number of hours of play per week.

So, this student number 1 and the student has got 180 marks and the student has studied 34 hours in a week. The student has submitted 3 assignments in a 1 and the student has played 15 hours in a week. And, the similar data is obtained for 20 students. So, this is the data, that we already have used couple of time. So, I try to obtain the analysis of variance result for this data.

(Refer Slide Time: 27:37)



So, in order to obtain the analysis of variance result you simply have to use the command anova a n o v a on the lm object. That means, if you remember, we already had used this command lm to obtain the fitted model between y and X1, X2, X3. Now, you simply try to use the outcome and try to execute the command anova. Inside the parenthesis you write the lm object and you will get this type of outcome, right.

So, you can see here, it is trying to give you here the response, which is your here y. And, then whatever is the sum of square due to regression which is contained here. So, sum of square due to regression. So, this has been further partitioned, because there are 3 explanatory variables. So, this has been partitioned into sum of square due to X1, sum of square due to X2, and sum of square due to X3, right.

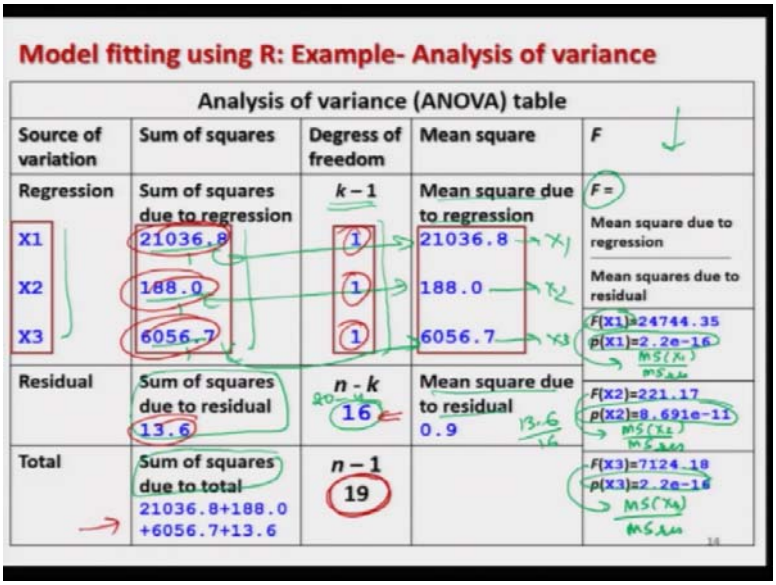
So, and then these are the corresponding degrees of freedom, these are the corresponding sum of squares, these are the corresponding mean squares, these are the corresponding F values and these are the corresponding p values, right. So, if you remember we had defined the sum of square due to total, as say sum of square due to regression, and sum of square due to error which is residual.

So, you can see here sum of square due to regression here has been partitioned into sum of square due to X_1, + sum of square due to X_2 + sum of square due to X_3. And, you can see here this is here residual, I am trying to use here a red color, right. So, this is the

degrees of freedom corresponding to sum of square due to residual, this is sum of square, this is mean square, and these are the different values of α which are used here.

So, if you can see here, if you want to obtain the sum of square due to regression, then you need to sum up these three components, right. So, this is actually what you have to learn that, how to interpret the software outcome, right. But at least one good thing is this at least now you know that, how this individual values have been obtained. So, I will try to show you it in more detail. So, that you do not get confused.

(Refer Slide Time: 30:28)



First of all I try to compare it with the and with this analysis of variance table, which I shown you here, which I explained it and this software outcome, right. So, here I have written all the values of the ANOVA table. So, sum of square due to regression this is obtained by here X1, X2, X3 these are the sum of square due to regression and you can see here that there are total number of variables are here, there are four values related to X1, X2, X3 and β 0. So, so k is here 4.

So, k - 1 becomes here 3 and this 3 has been divided into 1, 1, 1, right. And, now corresponding to this each of this X1, X2, X3, this mean square due to regression is obtained this is for X1, this is for X2, and this is for here X3. And, since sum of squares have been divided only by here 1.

So, that is why sum of squares and mean square, you can see they remain the same, right. And, sum of square due to residual, you can see here this is what is there and then - here k and is here 20 k is here 4. So, this is here 16. And, how to obtain the mean square due to residual, this is simply sum of square due to residual, which is 13.6 divided by 16, which is here 0.9, right?

And, if you want to obtain although we do not need this, if you want to obtain the sum of square due to total, you simply have to sum these sum of square, which I am trying to highlight in color red, this one, this one, this one, and this one, which I have written here. And, same as for the degrees of freedom also, this degrees of freedom here corresponding to the total, this is the sum of all the degrees of freedom.

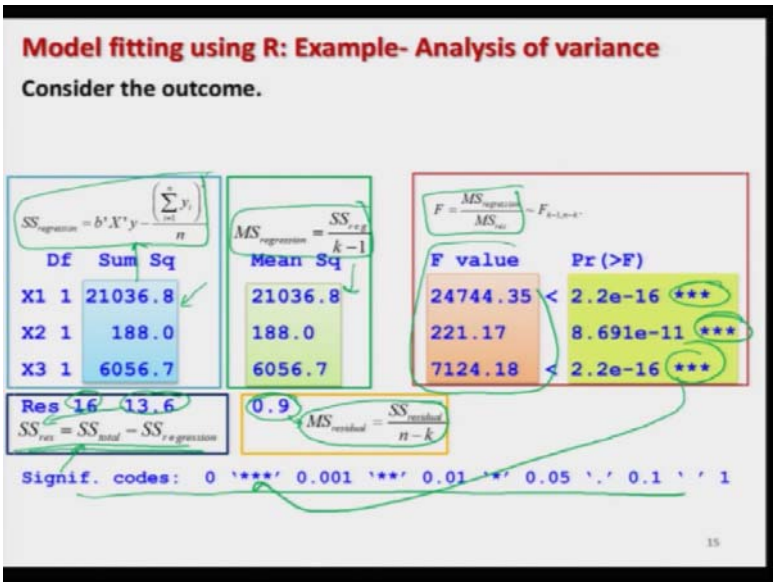
So, you can see here, this is here 16, this is here 17, 18 and here 19. So, the sum of squares are additive, similarly the degrees of freedoms are distributed exactly in the same way as the sum of squares have been distributed. And, this is precisely the Fischer Cochran theorem, right. Now, after this you come to the last column, that here we are getting the values of the F statistic.

So, now, you can see here there are 3 possible values X1, X2, X3, which are trying to indicate the variation due to these 3 variable. So, you have so, this F of X1 is indicating the F statistic corresponding to the sum of a square due to X1. So, this has been obtained as mean square due to X.

So, this is obtained here as say mean square due to X1 divided by mean square due to residual, right. And, this is here the corresponding p value, which has been obtained here in the table you can see this is here and here, or I will try to use here a different pen, so, that you can see it here. These are the value which is indicated here, right.

Similarly, if you go for this here second row, so, this is actually the F statistics about X2. So, this is mean square due to X2 divided by MS_{res}. And, similarly here the third one this is the F statistics for X3, which is here mean square due to X3 divided by mean square residual due to residuals. And, similarly the corresponding p values are here like this.

(Refer Slide Time: 34:28)



So, you can compare these p values and you can take a call whether the null hypothesis is accepted or not. So, you can compare here that, this is the comparable. And, here I am just trying to give you the same outcome, which I have rearranged. So, that I can give you that, what are the expressions which are computing these values?

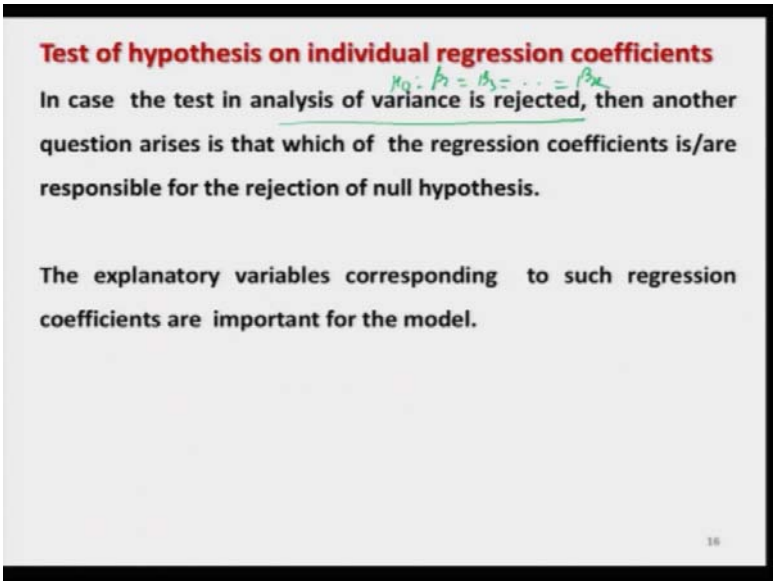
So, if you try to see the first column, which you have here sum of square, which I have highlighted in blue color, all these values are obtained from here, using this formula sum of square due to regression. And, similarly the mean square due to regression in the second column they are obtained here by s expression.

And, then finally, the F value they have been obtained here using the this expression. And, then here you can see in the this box, the sum of square due to residual. It is a degrees of freedom and sum of squares they have been obtained by this formula and, right. So, this is the value which is obtained here by SS_{res}, this is the degrees of freedom, right.

And, similarly in the case of mean square due to residual, which is here this value; this is obtained by this expression. And, these are the different values of α, which are indicating the level of significance and they are corresponding to these stars. So, you can see here these there are 3 stars. So, they are corresponding to this value, ok. So, now, you can see

there means I have tried my best to explain you each and every step and how these values have been obtained and how to take a final conclusion?, right.

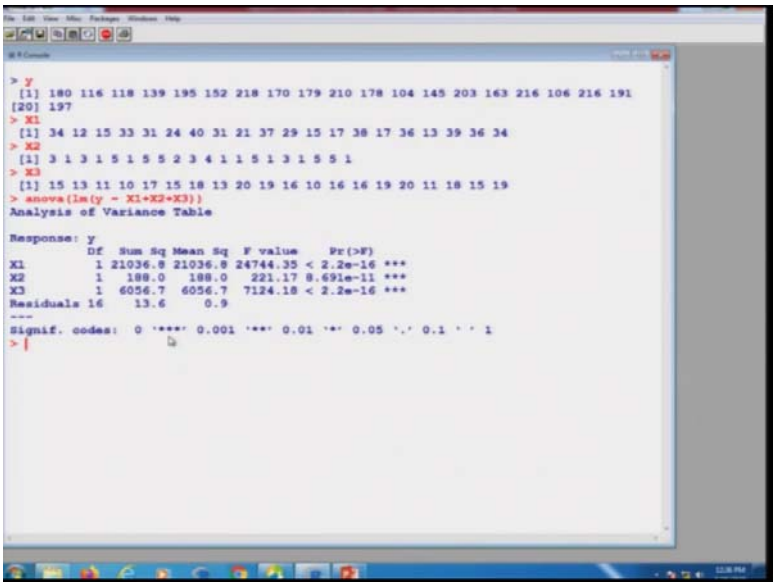
(Refer Slide Time: 36:12)



So, now the next question comes, in case if this hypothesis is rejected. Suppose, $H_0 : \beta_2 = \beta_3 = \ldots = \beta_k$, β_k is rejected; that means,, ok I agree that they are not equal. But, which are the components, which are trying to create this table, that which are the regression coefficient, which are not equal to 0, and which are not equal to others, which are making this hypothesis to be rejected. So, for that then we have one option that we can try to test the individual regression coefficients.

And, then we try to use the t statistic that we have used in the earlier lecture. And, they will try to help you in identifying that which are the variable corresponding to which the regression coefficients are significant. And, which are the important regressor or the important explanatory variables.

(Refer Slide Time: 37:30)



So, now we come to the r console and we try to show you that, how these things can be done? So, I already have entered this data you can see here y is your here like this X1 here is like this, X2 here is like this, and X3 here is like this. So, I already had entered this data, because, I had used it in earlier example.

Now, I want to I try to use the command for this anova and you can see here that this is the outcome which you have obtained. Here you are trying to write down the object that you try to use in the lm command and then you are simply using the anova. So, remember one thing spelling is a n o v a, that is this short form of analysis or variant and you will get here the same outcome which I have reported in the slides.

So, similarly you can take any number of variables, whatever you want and then you can conduct such anova, which is not difficult at all, ok. So, now, we come to an end to this lecture. But, this is not really actually an end the analysis of variance is itself a huge topic. And, this is very important from different perspective.

You might have heard the name of designs of experiment. And, design of experiment also plays a very important role in the this is sciences that, how should you get the data, how should you collect the data, so, that you get a correct outcome. It is not only in the sampling theory, but also design of experiment also plays this role. And, the data of the design of experiment is mainly analyzed using the analysis of variance table.

So, depending on different type of situations, we have different types of analysis of variance, but those type of analysis of variance are not really possible for me to cover under the same topic. You might have heard the names like, 1 way analysis of variance, 2 way analysis of variant, 3 way analysis of variance, or you might have heard the names of designs of experiment like, completely randomized, design randomized block design, Latin square design, balance incomplete block design and so on, right.

So, you can now see that, once you are trying to do something fruitful you cannot depend only on 1 topic, but all the topic they are trying to contribute in making a good decision. And, that is the role of decision sciences. That to integrate all the things together, but in order to integrate you need to learn all the things individually also and this is what I am doing here.

So, once you have learned the sampling theory, once you have learned the multiple linear regression analysis. At least now you can combine those things and you can take a much better decision. Sometime in future you can have a course on design of experiments also. Then whatever you are learning there they can be integrated with the multiple linear regression modeling.

And, all these analysis of various technique, whatever you have learnt here they can be extended. Actually, one the null hypothesis of analysis of variance is rejected, then we also go for multiple comparison test. And, there are different types of multiple comparison test, which help us in identifying that which are the variable which are causing the H_0 to be rejected. Different type of tests will give us different type of conclusion and this is our experience our knowledge based on which we take the correct decision.

So, I hope that after this lecture you will not have fear of analysis of variance or anova table in your life. And, if you have I will say please try to read the book, try to clarify your concept, and I hope there should not be any fear in your heart for anything in the life. So, you try to take an example practice it and I will see you in the next lecture with on with the lecture on goodness of it, till then goodbye.