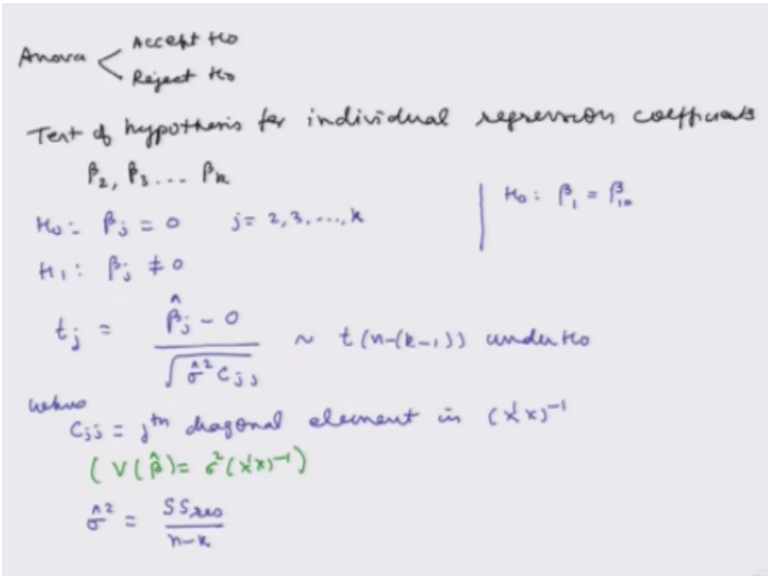
Regression Analysis and Forecasting
Prof. Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology-Kanpur

Lecture – 15
Testing of Hypothesis (continued) and Goodness of Fit of the Model

Welcome to this lecture, you may kindly recall that in the earlier lecture we had discussed the analysis of variance in a multiple linear regression model this is a test for testing the null hypothesis that all the regression coefficient are = 0 or not. So through the test of analysis of variance, we are judging the overall adequacy of the model.

(Refer Slide Time: 00:49)



Now, we have two options, suppose the results from this analysis of variance, they indicate that h-naught is accepted or second option is that h-naught is rejected right. So in case if h-naught is accepted, then there is no issue, we understand that none of the variables are going to contribute in explaining the variation in y. When h-naught is rejected, this indicates that at least there is one independent variable that is explaining the variation in the values of y.

And there is another alternative also, that there can be more than one explanatory variable which are helping in explaining the variation in y. So now the question is this we would like to identify those variables which are contributing and those variables which are not contributing. So in order to do it, we have to proceed one by 1. 1 by 1 means we have to go step wise and we need to test the significance of regression coefficients 1 by 1.

So now we try to discuss here the test of hypothesis for individual regression coefficients, and these regression coefficients are based on our assumption that well, they are responsible for the rejection of null hypothesis. So since, we have considered the regression coefficient beta 2, beta3, beta k. So I would like to develop a test of hypothesis for individual beta So let us try to postulate our null hypothesis, say h-naught beta j = say here 0.

And j goes from here two to k, you may also recall that in the case of simple linear regression modelling, we had discussed the construction of test statistics for testing the significance of slow parameters h-naught beta1 = beta1 naught. So this test of hypothesis now in our case that is going to be base on the similar lines what we did in the case of simple linear regression model, but in this case our alternative hypothesis in which we are interested is h1, beta j is not = 0.

So now, in the case of simple linear regression model, we had two cases, when sigma square is known and when sigma square is unknown. In this case, we can see that we have estimated the sigma square by some of a square due to residual divided by the degrees of freedom. So here we are interested in the case when sigma square is unknown to us and that is being estimated from the given sample of data.

So under social assumption, we can construct the test statistics tj which is beta j hat - its given value which is 0 divided by standard error of beta j hat which I am denoting as a sigma square cjj, where cjj is the jth diagonal element in the matrix x transpose x whole inverse, why if you try to remember, we had obtained that the covariance matrix of beta hat was sigma square x transpose x whole inverse.

And this covariance matrix is giving the variances of beta1 hat, beta2 hat, beta k hat on the diagonal terms and they are covariances on the off diagonal terms. So in order to find out the standard error of beta j, we are picking up the variance of beta j hat. So this is being given by sigma square and the jth diagonal element of the matrix, x transpose x whole inverse.

Now this statistics is going to follow a t distribution with n - k - one degrees of freedom under H nought. And, you can also recall that we are estimating sigma square by ss residual divided by degrees of freedom.

(Refer Slide Time: 06:00)

Decision: Reject H_0 at α level of significance

if $|t_j| > t_{\frac{\alpha}{2}; n-(k-1)}$

Marginal test as $\hat{\beta}_j$ depends on all other explanatory variables also except x_j.

Confidence interval for β_j

$100(1-\alpha)\%$ C.I. for β_j $(j = 2, 3, \ldots k)$

$$P\left[-t_{\frac{\alpha}{2}, n-(k-1)} \leq \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} \leq t_{\frac{\alpha}{2}, n-(k-1)} \right] = 1-\alpha$$

$$\left[\hat{\beta}_j - t_{\frac{\alpha}{2}, n-(k-1)} \sqrt{\hat{\sigma}^2 C_{jj}} \,,\ \hat{\beta}_j + t_{\frac{\alpha}{2}, n-(k-1)} \sqrt{\hat{\sigma}^2 C_{jj}} \right]$$

So now in this case, once we have obtained the data, we have calculated the statistics tj, my decision will become reject h-naught at alpha level of significance if absolute value of tj is greater than the critical value t alpha by two at n minus k - one-degrees of freedom. So this is actually a sort of marginal test. Marginal test means earlier we had analysis of variance, which is testing the equality of all beta two, beta three, beta k together.

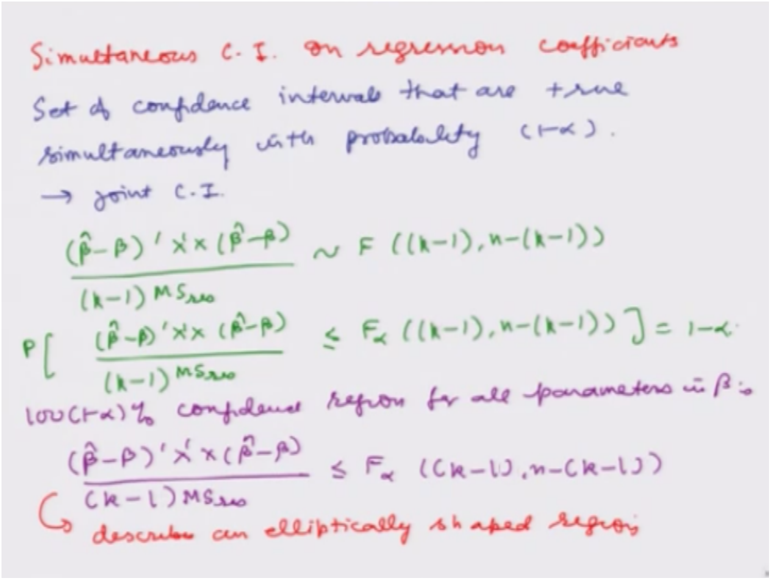
And now we are coming on the aspect of testing one regression coefficient at a time right. So, that are why this is called as a marginal test, why because, also the beta j depends on all other explanatory variables also except xj. So now using this thing, we can now identify that which are the independent variables are explaining the variation in y and which are not. Simultaneously, I would also like to discuss the aspect of confidence interval estimation.

So in this case, if we try to find out the confidence interval for beta j, then the hundred 1 - alpha percent confidence interval for beta j, j goes from two, three, up to here k is given by the expression and if you recall in the case of simple linear regression model, we had obtained the confidence interval for the slow parameter beta one and intercept term beta0. So here also we are going to follow the similar philosophy.

So, I can say here that this beta hat j - beta j upon is the standard that sigma square hat cjj, this lies between - t alpha by 2 and - k - 1 and t alpha by 2 and - k - 1 and this probability is going to be 1 - alpha, and based on this I can find out the confidence interval as beta hat j - t alpha by 2 and - k - 1, sigma hat square cjj and beta j hat + t alpha by 2 and - k - 1, is square root of sigma square cjj.

So this is the hundred 1 - alpha percent confidence intervals for an individual beta j. So now here you can see that how the theory concepts and algebra that we have learnt in the case of simple linear regression modelling is helping us in developing the model for a case when we have more than one independent variables Now continuing on the aspects of this confidence interval, this is a confidence interval for an individual regression coefficient.

(Refer Slide Time: 09:43)



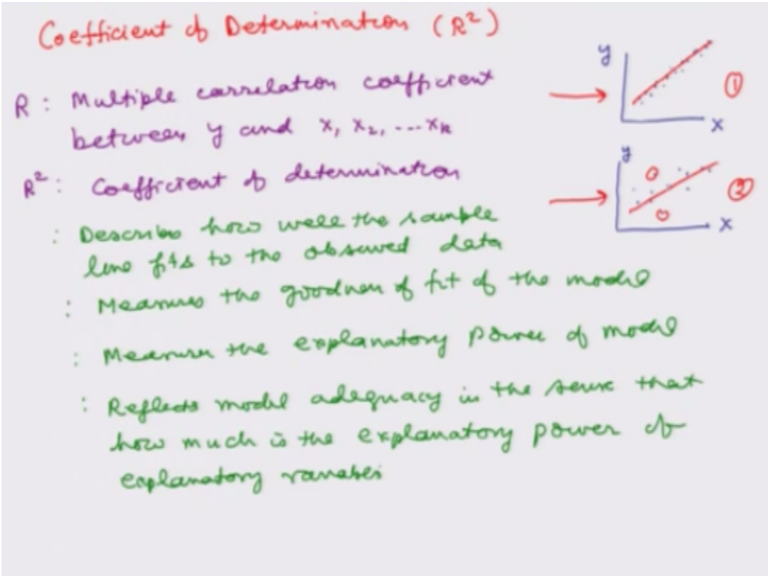
Now we can also construct the simultaneous confidence interval on regression coefficients right. What do we really mean by simultaneous confidence interval on regression coefficient? So this is actually a set of confidence intervals that are true simultaneously with probability one minus alpha right and they are also called as joint confidence interval.

So now in order to construct a joint confidence interval, we can use the result that beta hat - beta transpose, x transpose x beta hat - beta upon k - one MS res this follows a F distribution with k - 1 and n - k - 1 degrees of freedom. Now I can write down that we would like to find out the value of beta in such a way that beta hat - beta transpose x transpose x beta hat - beta over k - 1 times MS res is less than or = F alpha k - 1 and - k - one degrees of freedom.

And the probability of such an event is 1 - alpha. So now hundred 1 - alpha percent confidence region for all parameters in beta is the region which is given by this inequality k - 1 and this describes an elliptically shaped region. You see, when you have only 1 parameter then we have a confidence interval when we have two parameters and we want to find out their simultaneous confidence interval that can be a region in the two dimensional.

Similarly when we go for the ith dimensional, this confidence interval will be transformed into a region. So now, this confidence interval is essentially the simultaneous interval for all the parameters beta1, beta2, beta k, so this is going to be a sort of elliptically shaped region.

(Refer Slide Time: 13:22)



Well, after this we come to another aspect and we talked about coefficient of determination and this is actually denoted by R square. So now the first question is what is this coefficient of determination, now you see, now we have reached to a stage where using the data on independent and dependent variables, we have obtained a model by estimating the parameters beta1, beta2, beta k and sigma square.

Now this estimation technique can be anything, either least square estimation or maximum likelihood estimation, and based on that we have obtained the fitted model. Now basic question is that how do we know that the model which we have got is good or bad or how to judge the goodness of fit of this model. So this coefficient of determination helps us in determining the goodness of fit of a model.

So first question comes, how should I judge it? In case if you try to recall in the case of simple linear regression model what we have done. We had one independent variable x and we had one dependent variable y, we had obtained the data and then we have created a scattered diagram and it was something like this. So suppose if I try to take here two situations something like this x and y and in which the skills of x and y are the same here and we try to fit here a line like this and here this.

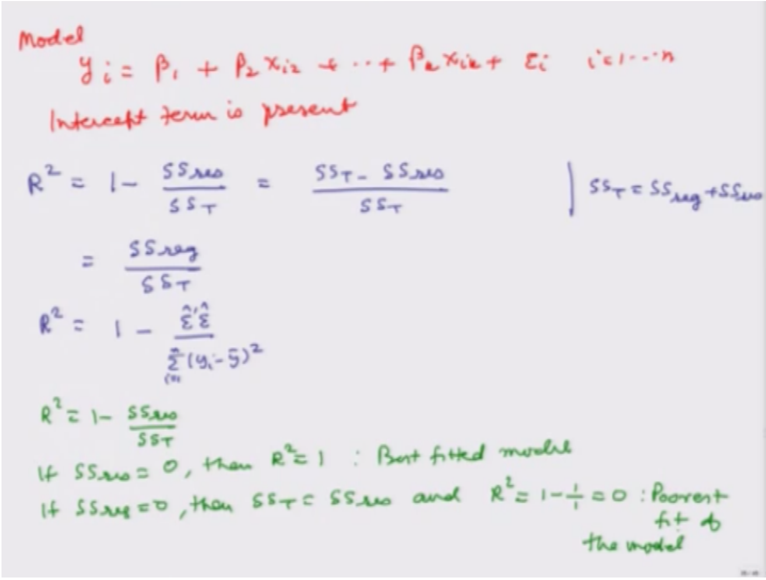
Now what you can say that which of the model is going to be fitted better, so obviously in this figure I can see that the points are lying more closely to the fitted line in comparison to this figure. Here you can see that the points are lying here and here and they are quite far away then the points in figure number one. So this gives us an idea that in case if the points are lying close to the line that means our model is better fitted.

One simple option to major this quantity is to find out the correlation coefficient between x and y. So obviously, in case of figure number 1, the correlation coefficient will have a higher value than in figure number 2. This concept is extended and this is used to judge the goodness of fit in a multiple linear regression model. Now we try to extend the concept of a simple correlation coefficient and we try to use the concept of multiple correlation coefficient.

So in case if I define R be the multiple correlation coefficient between y and x1, x2, here say x k. Then the square of this multiple correlation which is denoted as R square, this is called as coefficient of determination and the utility of R square is this it describes how well the sample line fits to the observed data, and in some sense this measures the goodness of fit of the model.

And it also measure the explanatory power of model and this reflects the model adequacy in the sense that how much is the explanatory power of the explanatory variables. So in simple sense I would say R square is a measure that will give us an idea that the model which we have obtained whether this model is good or bad and if good how much this is good, and if bad how much this is bad.

(Refer Slide Time: 19:06)



So we try to consider here the model and we consider here a model yi = beta1, beta2 xi 2 + up to here, beta k xik = epsilon i, i goes from here 1 to n. One very important thing we have to keep in mind that we are assuming here that intercept term is present. This R square has a limitation that it assumes that the intercept term is present in the model and the value of R square can be obtained only in such a condition.

If you do not consider the intercept term in the model, then the definition of R square which we are going to consider here, this will not remain valid. So now in this case, we try to define the R square has one minus say sum of a square due to residual divided by sum of a square due to total. Now if you see the interpretation of this R square under what condition you would call that a model is good fitted.

The model is good fitted when the contribution of the random error component in the model is as small as possible, and ideally this should be = 0. So now this can be return as SST minus SS res divided by SST. Now, you may recall that in the case of analysis of variance we had proved that SST = SS reg and SS res. So now if I try to use this relation over here, this can be written as SS reg or SST.

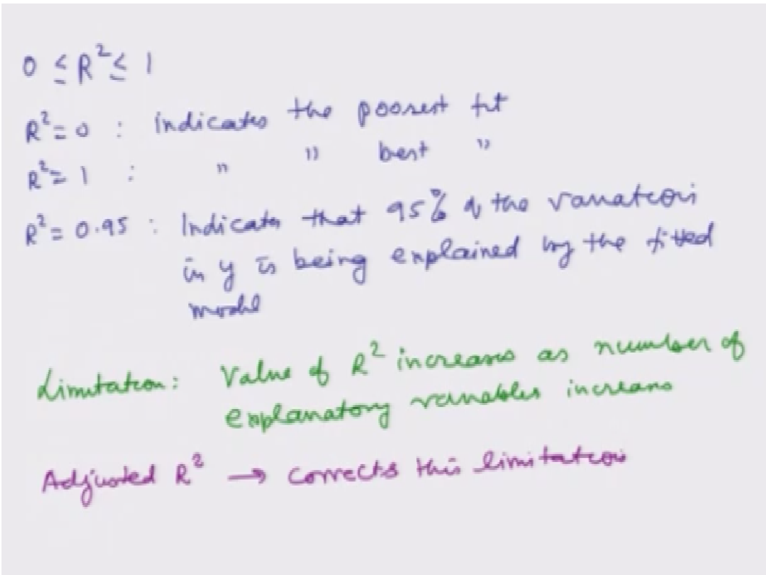
So now, the expression for this R square will simply now here one - SS res that is epsilon hat transpose epsilon hat divided by summation i goes now one to here n, yi - y bar, whole square. Now the question comes, how do we ensure that this definition of R square is giving us what we want? If you see in a good model, what will happen, for example, if I try to consider this expression R square = 1 - SS res over SST.

So a model will be good in case if the contribution due to random error is as small as possible and in that case, ideally we would assume that sum of square due to residual should be 0. So in this case, if I say that if sum of a square due to residual is zero then R square = 1 and this is a best fitted model and that would be an ideal condition in which we are all interested.

Now on the other hand, in case if the model is not at all good fitted, that means the contribution of the sum of square due to regression is 0 that means none of the x1, x2, x k variables are helping us in explaining the variations in the values of y and in that case when the sum of a square is due to regression become zero then SST = SS res and R square in this case becomes one minus one over one that = 0.

So this would be indicating the poorest fit of the model or rather I would say this is the worse fitted model.

(Refer Slide Time: 23:23)

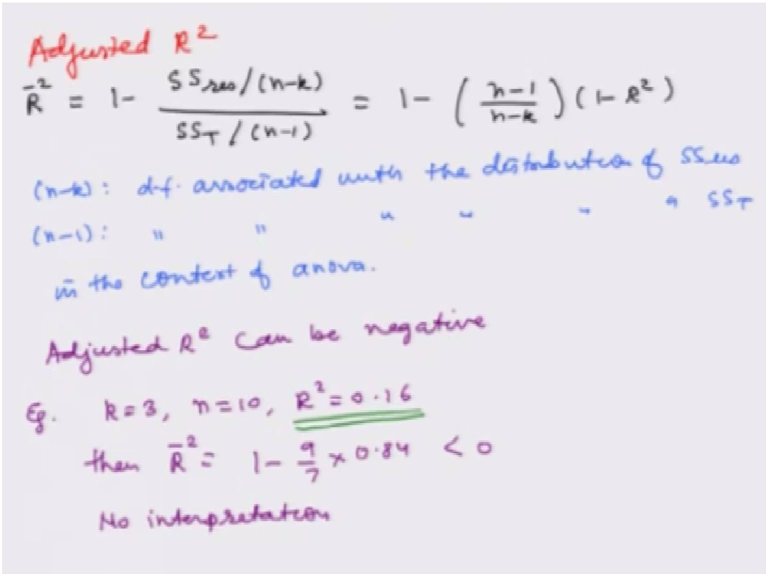


So now we see that the value of R square is lying between zero and one and R square = 0, this indicates the poorest fit, worse fit and R square = 1, this indicates the best fit. Now, if I try to take any other value of R square say for example, 0 point nine five, then this indicates that 95 percent of the variation in y is being explained by the fitted model or the independent variable x1, x2, x k.

This R square has one limitation that value of R square increases as number of explanatory variables increases, this is a limitation. Now suppose somebody is trying to fit a model with certain number of explanatory variables. Now some more variables are added in the model, which are not relevant, they are simply useless variable. They are not affecting at all the values of y.

So this does not indicate that the model will get better by using those irrelevant variables, but and we try to do so, the value of R square will increase and that would indicate my model is getting better and better. So in order to handle this limitation, we can define variant of R square which is called as adjusted R square. So this adjusted R square corrects this limitation.

(Refer Slide Time: 26:17)



So now we discuss about the adjusted R square. The adjusted R square, this is denoted by R bar square and this is defined as 1 - sum of square due to residuals divided by n - k upon sum of squares due to total and divided by n - one. So this can be further simplified as one - n - one over n - k times 1 - R square. Now if you try to observe, what are we going to do here, we are trying to divide sum of square due to residual by n - k.

So if you try to recall what was your n - k in the context of some of the square due to residual, this was actually the degrees of freedom associated with the distribution of SS res in the context of analysis of variance and similarly, if you see we are trying to divide the total sum of a square SST by here n - 1. So here again, this n minus one is simply the degrees of freedom associated with the distribution of SST in the context of analysis of variance.

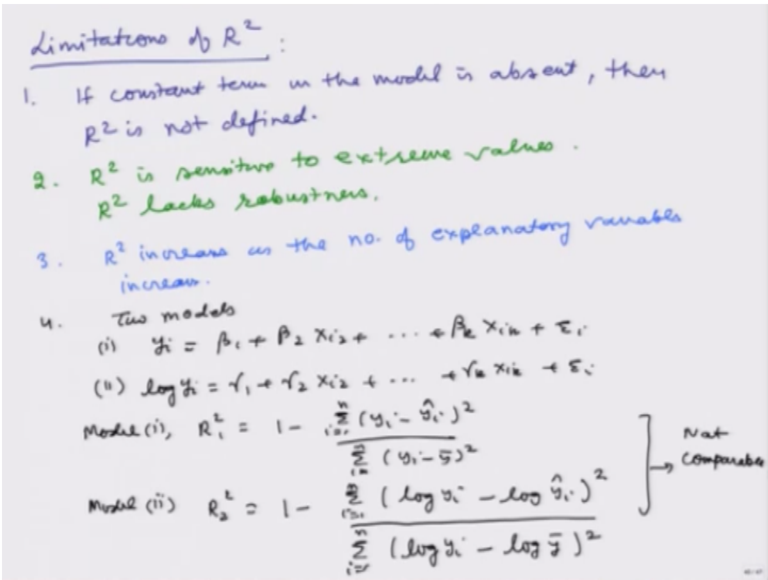
So this adjusted R square helps us and it does not increase has the number of independent variables are added in the model, but on the other side, adjusted R square also has certain limitation. So, first limitation is this, that adjusted R square can be negative. With this is difficult to believe that well, that is a square quantity but if you try to see with this is a function of n k and see this here R square.

Yes, R square cannot be negative. For example if I try to illustrate this limitation, let me take a hypothetical example that k = 3, n = 10 and R = 0 point 1, 6 and then R bar square will be one - 9 over 7 into 0 point 8 4 and this value turns out to be less than 0. So obviously now you can see here that R bar square has no interpretation.

On the other hand, if you try to look at this situation from the application point of view, I would argue that in practice such situations are very rare to occur, why because if you see we are here getting a value of R square which is 0 point 1, 6 that means the fitted model is explaining only 16% of variability using the variable x1, x2, x k. So obviously this is already indicating that the linearity of the model is questionable.

So in this situation, even I would not like to use the multiple linear regression model, but rather I would try to fit some other model.

(Refer Slide Time: 30:09)



So now after this, let me try to explain that what are the limitations of R square, well R square is a very popular goodness of fittest statistics among all the user in experimental sciences, but it has got some serious limitations also. First limitation, we already have discussed that if constant term or the intercept term in the model is absent, then R square is not defined.

This can also be shown mathematically, but I am skipping the proof here, and in case if someone is considering a model without the intercept term and still if he or she tries to find out the value of R square, there is a risk that the R square value cannot be negative. The next question comes, then if someone is trying to fit a model without intercept term, then how the goodness of fit of a model can be judged.

Well, there is a unique answer in the literature some adopt measures have been defined, but definitely there is no guarantee that those adopt measures will give us a good statistical outcome, but any way this is the limitation of the R square and this is how it has to be used okay. The second limitation of R square is this that R square is sensitive to extreme values. So I can say in simple word that R square lacks robustness.

Now coming on the issue from the application point of view that we know that when we are going to fit a model to a given set of data, we need to first make sure that there are no extreme values in the given set of data and in case if they are present, there are some other ways to handle them. So I really, this condition will not happen in practice if somebody is carefully making a model, but definitely in case if somebody is ignoring this aspect then R square will lack the robustness.

And earlier, we had also discussed the third limitation that R square increases as the number of explanatory variables increases. Now, let me come to a different type of situation where we are interested in comparing two different models. Suppose, there is a situation and there are two models which are fitted. Suppose the model number 1 is say yi = beta1 + beta2 xi 2 + beta k xik + epsilon i.

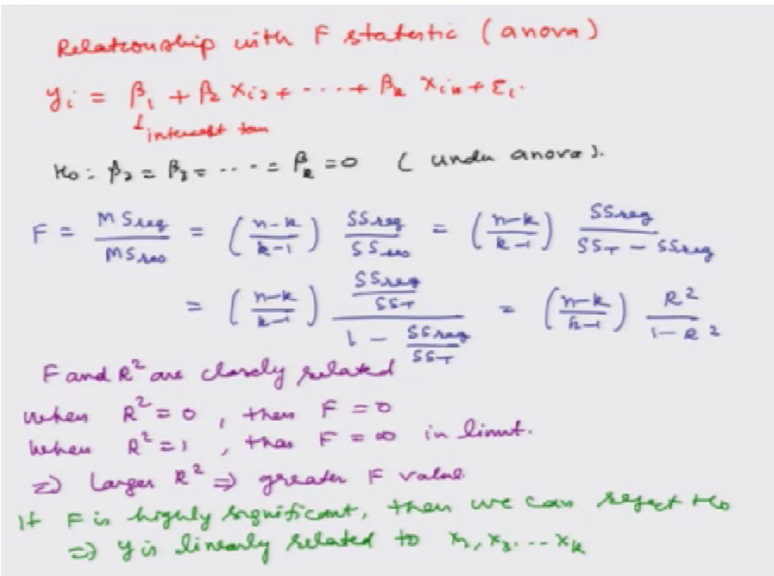
Second model is that, the model is fitted using the same data, but by taking the log transformation, that all the values on the response variable, there log is taken and then the model is fitted. So in this case, I would like to denote the regression coefficients by gamma1, gamma2, gamma k in place of beta1, beta2, beta k, so the model can be written as here xi 2 + here gamma k xik + some suppose some random error component epsilon i.

In this case, if we try to define the R square then for model number one, the R square is suppose, R1 square that is defined as one minus summation i goes now 1 to n, yi - yi hat whole square divided by i goes from 1 to n. Say yi - y bar whole square. Now we have to define the R square for model number two, now there can be various possibilities. For example, one simple option, I am not saying this is the only option there can be many other option.

One option is that 1 - summation i goes from 1 to n, log of yi - log of yi hat square divided by i goes from one to n, log of yi - log of y bar. Here also someone may argue that instead of taking log of y bar we would like to consider the arithmetic mean of log of yi's that first we take the transformation and then finding out the sample mean but anyway that is not my objective way to discuss those things, but it is clear from the values of R one square and R two square that these two values are not comparable.

So now the issue is very, very simple that two different persons or the same persons have obtained two different models and he wants to know out of this model number 1 and 2 which is a better fitted model, so this cannot be obtained using the definition of R square. So I am not saying at all that R square is a bad measure but R square is a very good measure, R square measures the goodness of fit but it has some limitation and it has some nice properties. So I would suggest that use R square but be careful, handle with care.

(Refer Slide Time: 36:51)



This R square has also got a relationship with F statistics that we had obtained in the case of analysis of variance, we see how, if you recall that in the case of analysis of variants, we had considered the model yi = beta1, beta2 xi 2 + beta k xik + epsilon i and at that time also I had told that we are going to consider here the presence of intercept term in the model because that we will use later on to established relationship between R square and this F statistics.

So now, this is the situation we are going to use it here. So if you remember that over a null hypothesis in the case of analysis of variants was h-naught beta2 = beta3 = beta k = 0. This was your under analysis of variance, and based on that we had finally obtained the F statistics which was obtained as mean square due to regression divided by mean square due to residuals and this was actually n - k upon k - 1, SS due to regression and SS due to residuals.

Now this can further be expressed as a n - k over k - one and SS regression and now using the relationship that total sum of square is equal to sum of square due to regression plus some of square due to residuals, I can rewrite some of square due to residual as total sum of squares - sum of square due to regression. So this can be written as here, n - k over k - 1, SS reg divided by SST divided by one - SS reg divided by SST.

So this comes out to be nothing but n - k over k - one and then R square over one - R square. So you can see here that there is a close form relationship between the F statistics of analysis of variance and R square and both are very closely related and then they have a very close interpretation also. So F and R square are closely related. We will see how? When I say, suppose R square = 0.

Then in this case F also becomes zero and when we say that R square = 1 then that becomes infinity in limit. So this implies that larger the value of R square this implies greater the value of F. So now what is the interpretation of this thing, so I can conclude that if F is highly significant that means the test of hypothesis based on this F statistics is indicating that all the regression coefficients are significant, then we can reject h naught.

And when we reject the hypothesis h-naught, when all the variables x2, x3 up to x k they are relevant variable and they are helping in explaining the variation in y. So I can conclude that y is linearly related to x2, x3, x k and that is what we had also said in case of analysis of variance that this is a test of overall adequacy. So now one can see that there is a close connection between F statistics of analysis of variants and this R square and their interpretations are also related.

So when we are going for the software issues then we will see that the software outcome consist of R square values, adjusted R square values as well as the analysis of variance stable. So looking at the outcome of software, we try to make different types of conclusions for the fitted linear regression model. So now we stop here the topics of multiple linear regression smodel, and in the next lecture, we will come off with some other issues, till then good bye.