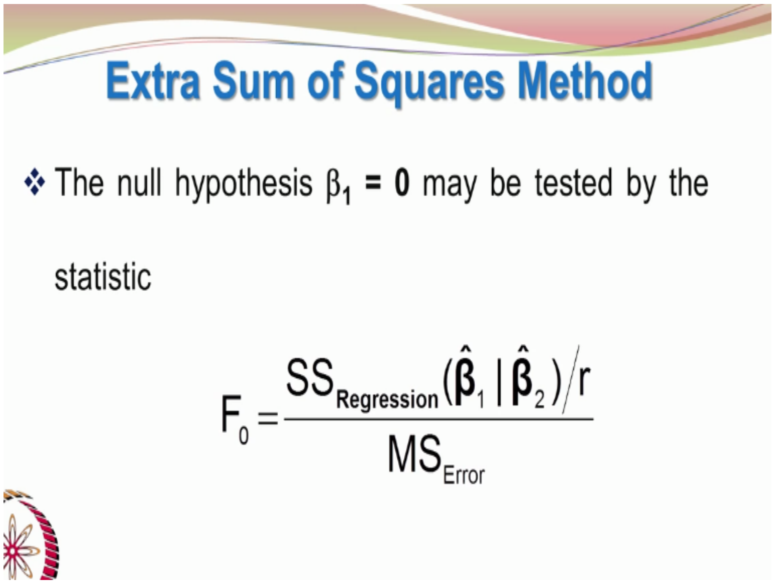
Statistics for Experimentalists
Prof. Kannan A
Department of Chemical Engineering
Indian Institute of Technology – Madras

Lecture – 40
Discussion on Regression Output

Welcome back in today's lecture. We will continue with regression analysis.

(Refer Slide Time: 00:25)



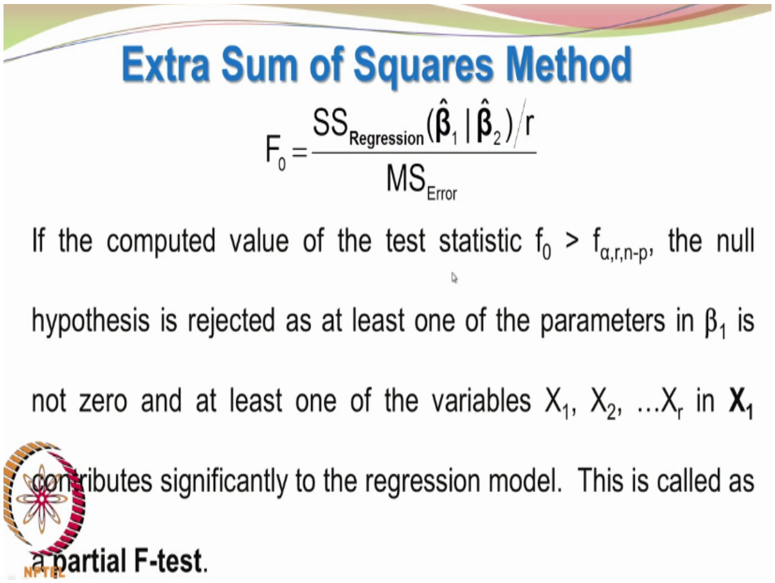
Extra Sum of Squares Method

- The null hypothesis β_1 **= 0** may be tested by the statistic

$$F_0 = \frac{SS_{Regression}(\hat{\boldsymbol{\beta}}_1 \mid \hat{\boldsymbol{\beta}}_2)/r}{MS_{Error}}$$

So we were discussing about the extra sum of squares method and we test the null hypothesis beta1 = 0 using the statistics sum of squares of regression beta hat1 given that beta hat2 is already present in the model/r degrees of freedom for beta hat1/mean square error. So, we know that beta2 hat and beta1 hat are not necessarily single parameters. They represent block of parameters. They are actually column vectors as was discussed in the previous class.

(Refer Slide Time: 01:25)



Extra Sum of Squares Method

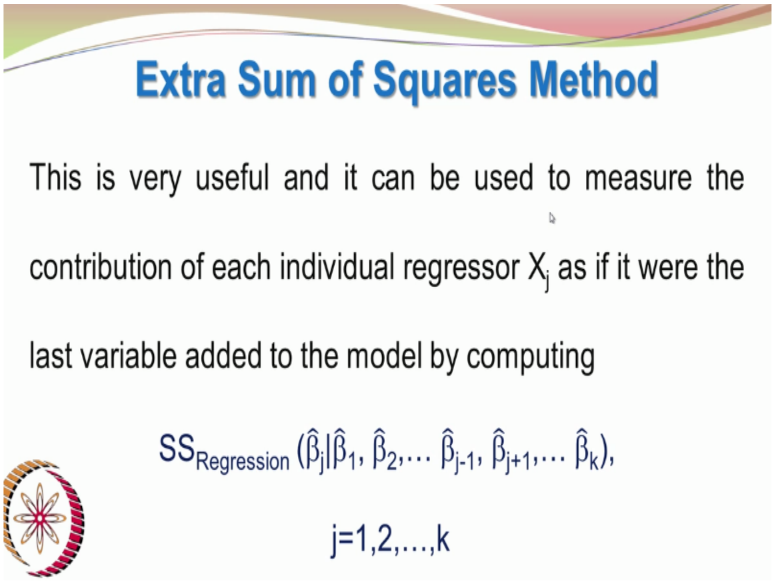
$$F_0 = \frac{SS_{Regression}(\hat{\boldsymbol{\beta}}_1 \mid \hat{\boldsymbol{\beta}}_2)/r}{MS_{Error}}$$

If the computed value of the test statistic $f_0 > f_{\alpha,r,n-p}$, the null hypothesis is rejected as at least one of the parameters in β_1 is not zero and at least one of the variables $X_1, X_2, \ldots X_r$ in $\mathbf{X}_1$ contributes significantly to the regression model. This is called as a **partial F-test**.

So if the computed value of the test statistic F0 is > f alpha r numerator degrees of freedom/n - p denominator degrees of freedom, then the null hypothesis is rejected as at least 1 of the parameters in beta1 is not 0 and at least 1 of the variables x1, x2, so on to xr in x1 contributes significantly to the regression model.

This is called as the partial F test. Beta1 is vector of parameter that is why it is represented in bold and if we reject the null hypothesis at least 1 of the parameters in the newly added model is not 0 and at least 1 of the variables in the fresh set x1, x2 so on to xr contributes significantly to the regression model. So the new variables are considered at least 1 of them brings value addition to the regression. This is also known as the partial F test.

(Refer Slide Time: 02:48)



Extra Sum of Squares Method

This is very useful and it can be used to measure the contribution of each individual regressor X_j as if it were the last variable added to the model by computing

$$SS_{Regression}(\hat{\beta}_j|\hat{\beta}_1, \hat{\beta}_2, \ldots \hat{\beta}_{j-1}, \hat{\beta}_{j+1}, \ldots \hat{\beta}_k),$$

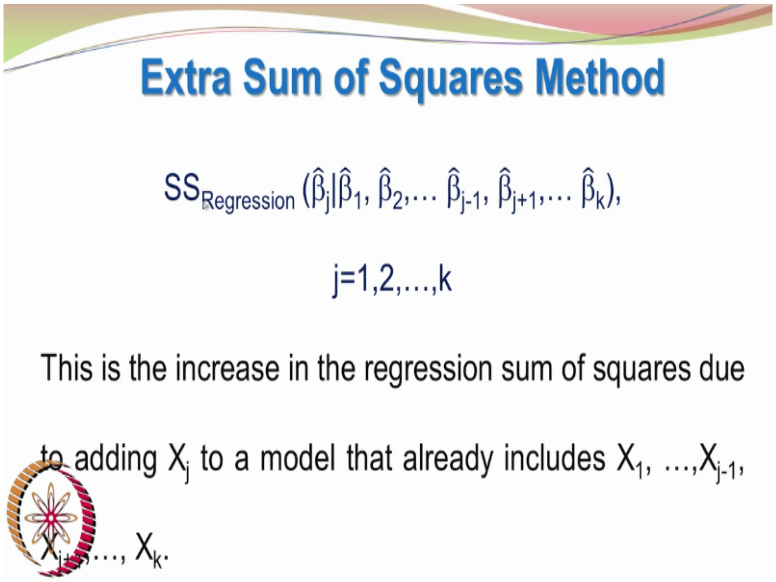
$$j=1,2,\ldots,k$$

So the extra sum of squares method is a very useful technique and what we can do is to use it to measure the contribution of each individual regressor variable xj as if it was the last variable added in the model. So what we do is let us arbitrarily pick up a regressor variable xj. Then we see the impact of adding xj to the modeling process by first developing a model without the xj parameter. So we have a regression model equation.

Now we can see the impact of adding the regressor variable xj to it so we do it by conducting a test of sum of squares of regression brought in by beta hat j given that beta hat1, beta hat2 so on to beta hat j - 1, beta hat j + 1 so on to beta hat k were already present in the model. So this is also a kind of extra sum of squares technique. So instead of adding a block of parameters we are considering only 1 parameter here which means that we are considering the effect of beta hat j. So j can be any value from 1 to 2 on to k.

It need not necessarily be the first parameter or the last parameter all the time. You have to first develop a model without the parameter beta j, so without the regression parameter beta j. In other words you are not accounting for the regressor variable xj. So other than that you consider all the other variables and develop it. Then you see the impact of bringing in the beta j * the regression model equation.

(Refer Slide Time: 05:09)



Extra Sum of Squares Method

$$SS_{Regression}(\hat{\beta}_j|\hat{\beta}_1, \hat{\beta}_2, \ldots \hat{\beta}_{j-1}, \hat{\beta}_{j+1}, \ldots \hat{\beta}_k),$$

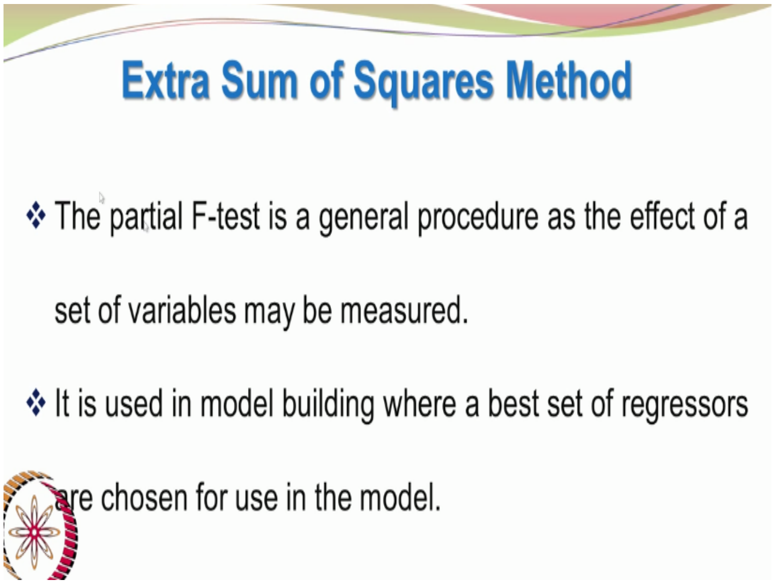
$$j=1,2,\ldots,k$$

This is the increase in the regression sum of squares due to adding X_j to a model that already includes $X_1, \ldots, X_{j-1}$, $X_{j+1}, \ldots, X_k$.

What this means? the sum of squares of regression due to beta hat j given that beta hat1, beta hat2 so on to beta hat j - 1, beta hat j + 1 so on to beta hat k are already present in the model. This means what is the increase in the regression sum of squares due to adding xj to a model that already includes x1 so on to xj - 1, xj + 1 so on to xk. The sum of squares obviously will be positive; it can be never negative.

So when you are considering a new regressor variable xj obviously the sum of squares associated with it will add on to the existing sum of squares due to the other parameters beta hat1 so on to beta hat j - 1, beta hat j + 1 so on to beta hat k. So what is the value addition brought in by this particular parameter beta hat j. It is very interesting and you can see the impact of each and every parameter by doing this exercise.

(Refer Slide Time: 06:25)

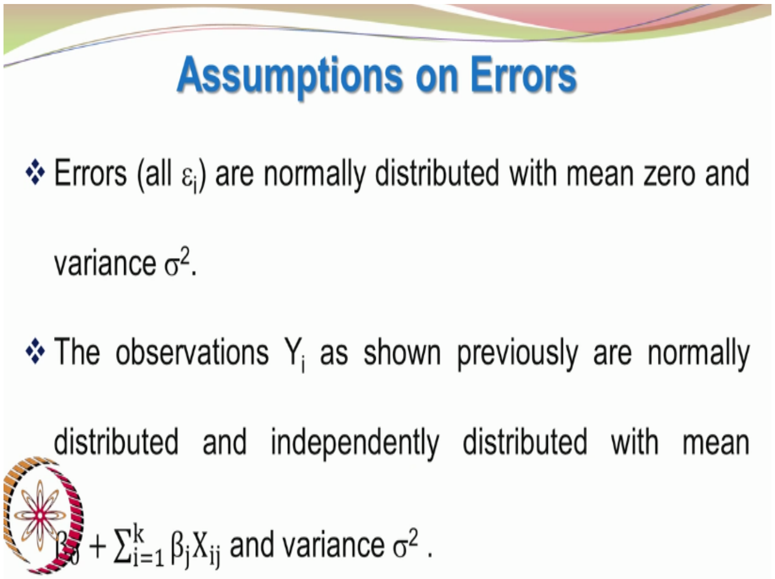


Extra Sum of Squares Method

- The partial F-test is a general procedure as the effect of a set of variables may be measured.
- It is used in model building where a best set of regressors are chosen for use in the model.

So the partial F test is a general procedure as the effect of a set of variables may be measured. It is used in model building where a best set of regressors are chosen for use in the model. So by doing this analysis you can identify the best set of variables which are having maximum impact on the response so that you build a economical compact and efficient model with only the regressor variables actually influencing the process or including the model and the other model terms or other regressor variables are excluded from the model.

(Refer Slide Time: 07:07)



Assumptions on Errors

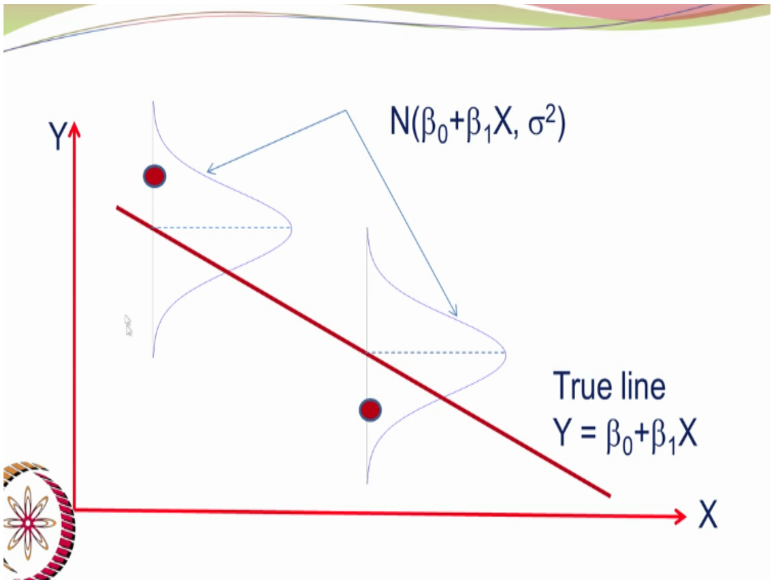
- Errors (all ε_i) are normally distributed with mean zero and variance σ^2.
- The observations Y_i as shown previously are normally distributed and independently distributed with mean $\beta_0 + \sum_{i=1}^{k} \beta_j X_{ij}$ and variance σ^2 .

Now let us look at the errors. The errors you may recall was defined in the regression model, the response is equal to the true value of the response + epsilon, where epsilon is the error. The true value of a response was given as eta and the actual response was given as y. If there were no errors in the experiment, then miraculously all the responses would be equal to the true value and when we repeat the experiments n number of times, we will get the same value eta i.

So we get different values because of the errors and we also noted that the errors are normally distributed with mean 0 and variance sigma square. The errors were normally distributed around 0 with the zero as the average and the constant variance of sigma square. The observations yi as shown previously are also normally distributed and independently distributed with mean beta 0 + I = 1 to k sigma beta I xij and variance sigma square.

So we talked about the errors now what about the response? The response is nothing but, a particular value a constant value eta i + the error. So when your errors are normally distributed when you add a constant to it then the response also will be normally distributed and the true value which is adding to the average of 0 would be nothing but the correct exact model beta0 + i = 1 to k sigma beta i xij. So I am not sure how many of you could follow this verbal statements so I will just show a diagram in the next slide, which we have also seen previously.

(Refer Slide Time: 09:17)



So for easier representation, I am just considering only on regressor variable x1 and here it can be generally called as x and this is the response of y versus x. You can have several points, but I am just showing 2 points for illustration. Looking at this representation, where we have only regressor variables for convenience and that regressor variable is x we are plotting the response y versus x and this solid line here is the true model given by y = beta0 + beta1 x.

So this is != beta0 hat + beta1 hat. It is actually the true model that is why it is called a true line, given by beta0 + beta1 x. So please note the distinction between beta0 and beta1 which are actually the exact or the 2 parameters representing the process whereas beta hat0 and beta hat1 would be the predicted parameters for beta0 and beta1. So having that out of the way we see that we are having these data points scattered around this true line.

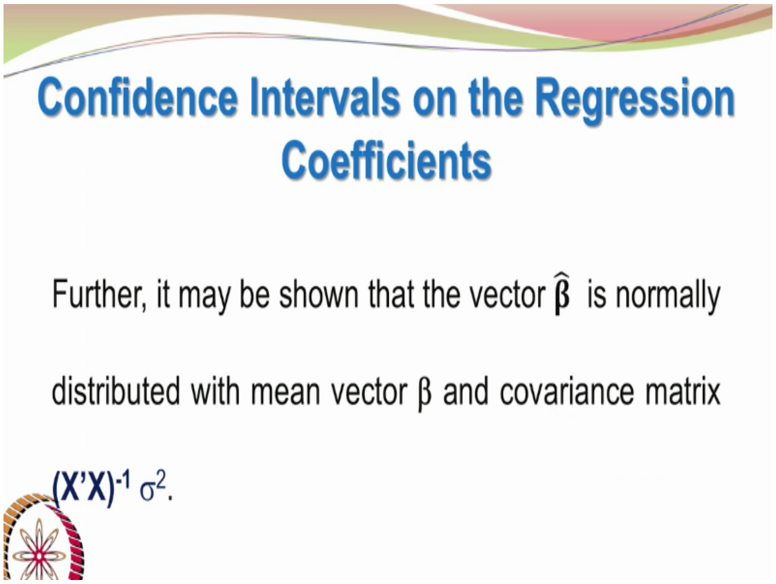
If this experiments were perfect and uninfluenced by errors the 2 dots would have fallen on the solid line, but they are sort of scattered. So these are the responses and these responses are normally distributed and the mean value of the response for example this is response 1 and this is response 2. There can be several such responses. I am just showing 2 for illustration and the line is drawn in a such a way so that the responses are on either side of the line.

So this response is above the line and this response is below the line. So you can see that the responses because of the error are normally distributed with the mean value given by the true line beta0 + beta1 x and the variance of this distribution is sigma square. So what it means is, because of random effects the points here my fall anywhere in this region. Of course, it may even go beyond that, but the probability of that occurrence would be very less.

This is the normal distribution and it depends on the value of sigma square. If the sigma square is pretty high, then there is a possibility that the point may be even further off because the distribution would be more broad and on the same line if the sigma square is very small then this distribution would be narrow and the points would be lying closer to the line. So this is the value x1, let us say the first setting of the regressor variable x or xA.

For example, and this is the response. So the true value would be y at 0.A = beta0 + beta1 xA. Similarly, this is xB for the regressor variable x. So then the yB the response at B would be beta0 + beta1 xB that would be the mean value or the true value, but the actual value may be somewhere away from the mean value.

(Refer Slide Time: 13:18)



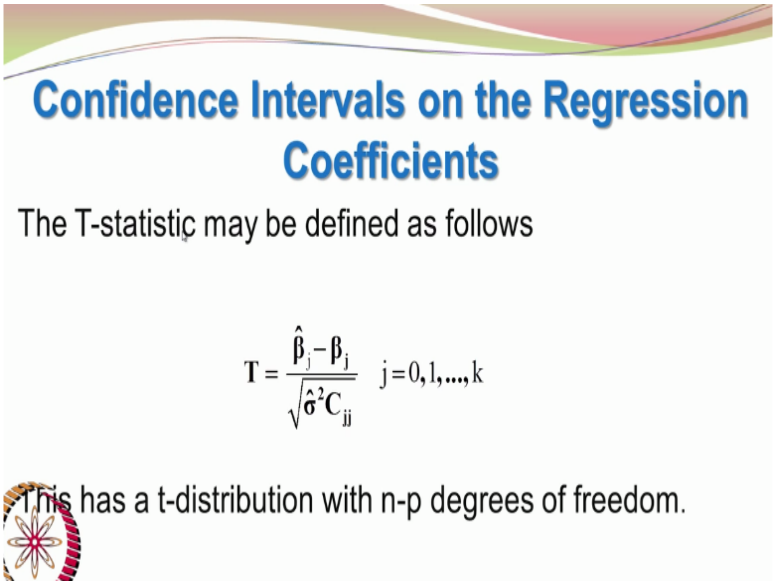
Confidence Intervals on the Regression Coefficients

Further, it may be shown that the vector $\hat{\boldsymbol{\beta}}$ is normally distributed with mean vector β and covariance matrix $(\mathbf{X'X})^{-1}\sigma^2$.

So now, we have discussed about extra sum of squares and the T test and so on. So now let us look at the confidence intervals on the regression coefficients. So the vector beta hat may be shown to be normally distributed with mean vector beta and covariance matrix x prime x inverse sigma square. Now we are not dealing with the individual entities, but we are actually dealing with the collection of regression coefficients and they are given is the column vector.

And so that would have a mean vector beta and a covariance matrix x prime x inverse sigma square. We have already seen what is the covariance matrix in one of our earlier lectures?

(Refer Slide Time: 14:10)



Confidence Intervals on the Regression Coefficients

The T-statistic may be defined as follows

$$\mathbf{T} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} \quad j = 0, 1, \ldots, k$$

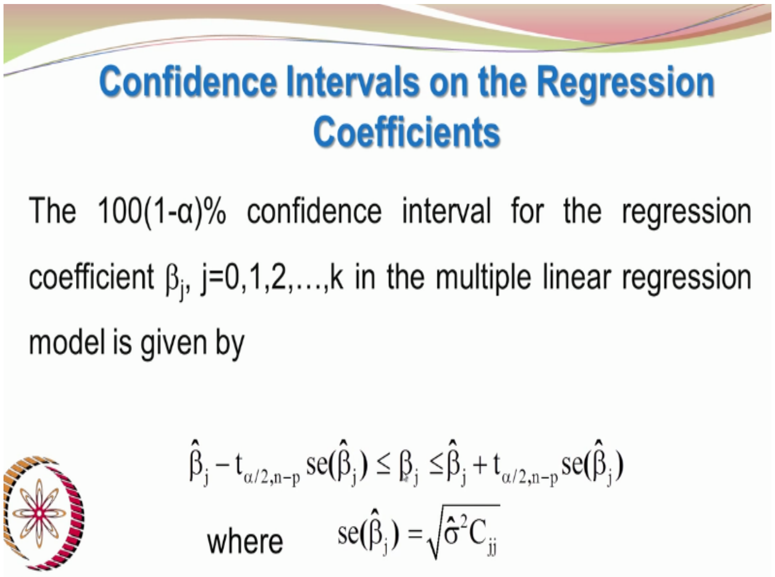
This has a t-distribution with n-p degrees of freedom.

Now we have T statistics which may be defined as T = beta hat j - beta j. This is the actual regressed value of the parameter and this is the true value and we also have the sigma hat squared because sigma square is not known so we need to have an estimate of the sigma square for which we use if you recollect the residual sum of squares/n - p where n is the number of data points and p is the number of parameters and Cjj is the x prime x inverse matrices diagonal coefficient corresponding to j.

The x prime x inverse matrix is matrix which may comprise of off diagonal term 0 or nonzero, but we are not interested in the off diagonal terms, we are only interested in the diagonal term and we pick up the diagonal corresponding to j for example if we are looking at beta1 then we will be looking at c11, first row first column element. If you are looking at beta2, then j will be = 2 and so we will be looking at C22, second row second element.

So we will be looking at the value of the variance covariance matrix along the diagonal and since the sigma hat square was based on n - p degrees of freedom, the n - p degrees of freedom were associated with the residual sum of squares. The T distribution is also associated with n - p degrees of freedom. n is the number of data points and p is the number of parameters.

(Refer Slide Time: 16:15)



Confidence Intervals on the Regression Coefficients

The 100(1-α)% confidence interval for the regression coefficient β_j, j=0,1,2,…,k in the multiple linear regression model is given by

$$\hat{\beta}_j - t_{\alpha/2,n-p}\, se(\hat{\beta}_j) \le \beta_j \le \hat{\beta}_j + t_{\alpha/2,n-p}\, se(\hat{\beta}_j)$$

where $se(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 C_{jj}}$

So now we can define the 100 * 1- alpha % confidence interval for the regression coefficient beta j, j running from 0, 1, 2, so on to k in the multiple linear regression model. So we have this beta hat j - T alpha/2 n - p standard error of beta hat j <= beta j, <= beta hat j + t alpha/2 n - p standard error of beta hat j. I think this should look very familiar to you in our phase 1 of the lectures where we discussed about T distributions, the hypothesis testing, confidence intervals.

We had if you recollect x bar - t alpha/2 s/root n where s is the standard deviation of the sample s/root n <= mu <= x bar + t alpha/2 n - p * s/root n. In some cases, we had sigma/root n, in some cases we had s/root n so it depends on whether we use the T distribution or the z distribution if the population variance was known then and the population was normal distribution or the sample size was pretty high greater than 30 so that we can bring in the central limit theorem into play.

Then we can use z alpha/2 * sigma/root n in the cases where the parent distribution is normally distributed and the variance sigma squared is not known which is usually the case then we have to make do with the sample standard deviation and so we have s the sample standard deviation and so we put x bar - t alpha/2 * s/root n. so whatever we have studied earlier is making perfect sense now, now are developing the confidence intervals for the regression coefficient beta j.

And that is why we have the predicted or the sampled if you want to put in that way value of beta j so that would be beta hat j and then you have the T distribution value corresponding to the chosen level of significance of alpha and n - p degrees of freedom. You can always read up this value from the T tables or go t a spread sheet and then calculate the T value and then you also have the standard error of beta hat j.

So this will give you the confidence interval and what do you use with this confidence interval. If the confidence interval is such that you have a negative lower limit and positive upper limit, then the beta j is pretty much worthless. On the other hand, if the lower limit of beta j is let us say very close to the upper limit of beta j the lower limit would be on the left hand side and the upper limit would be on the right hand side so if the upper limit and the lower limit are pretty close to each other they can be negative or positive.

But if they are very close to each other, then that parameter beta j has been precisely identified, but if you have a case where the left hand side is negative and the right hand side is positive then what do you make out of that beta j is it acting towards increasing the response when the regressor variable xj increases or is it acting towards decreasing the response when the regressor variable xj increases.

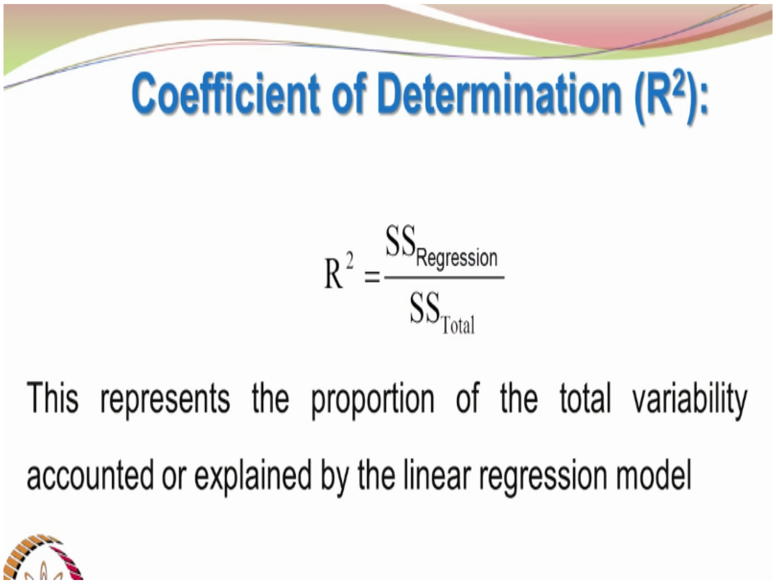
So under such a scenario we cannot make any definitive conclusion about the beta j and we pretty much say that it is insignificant. So the moral of the story is the lower limit and the upper limit of beta j should bare the same sign. If they have a negative value, then that beta j is acting towards decreasing the model response when xj increases. If the beta j has both positive lower limit and positive upper limit then the beta j is taking a positive value and when xj increases the effect of xj is to increase the process response.

(Refer Slide Time: 20:54)

Regression Analysis: A Few Terms Explained

Now whenever you do regression analysis either manually which is very rare or you do with the help of software or a spread sheet the program through a lot of results and sometimes we do not really know what those results mean, the most popular of that would be r square and if the r square value is let us say 0.99 we feel very happy and we feel the achievement of fitting an excellent model to the given data. Actually let us see there a few pit falls in this kind of feeling when you have a very high value of r square. Let us see what those are.

(Refer Slide Time: 21:40)



Coefficient of Determination (R^2):

$$R^2 = \frac{SS_{Regression}}{SS_{Total}}$$

This represents the proportion of the total variability accounted or explained by the linear regression model

Now the coefficient of determination R square, now we have a name for it instead of just a symbol coefficient of determination I do not know how many of you were aware of it previously. The coefficient of determination is simply the ratio of the regression sum of squares the total sum of squares. The regression sum of squares is a very valuable entity. It sorts of gives you the effective worth of the regression model.

We also looked at the extra sum of squares and the partially of test and so on. So we were always talking about the sum of squares of the regression brought in by a particular parameter or a set of parameters. So collectively they represent the total sum of squares of the regression and we compare the sum of squares of the regression with the total sum of squares in the model and see what fraction of the total sum of squares is contributed by the sum of squares of regression.

If miraculously you have a situation where the regression entirely contributes to the total sum of squares, then R square will be = 1. So you would like by looking at this equation, the R square value to be as close as 1 to be good enough, but I have seen papers especially in the biological sciences where people report values of R square of 0.68, 0.7 and so on. So it all depends upon the application, what would be an acceptable value. So what exactly is R square.

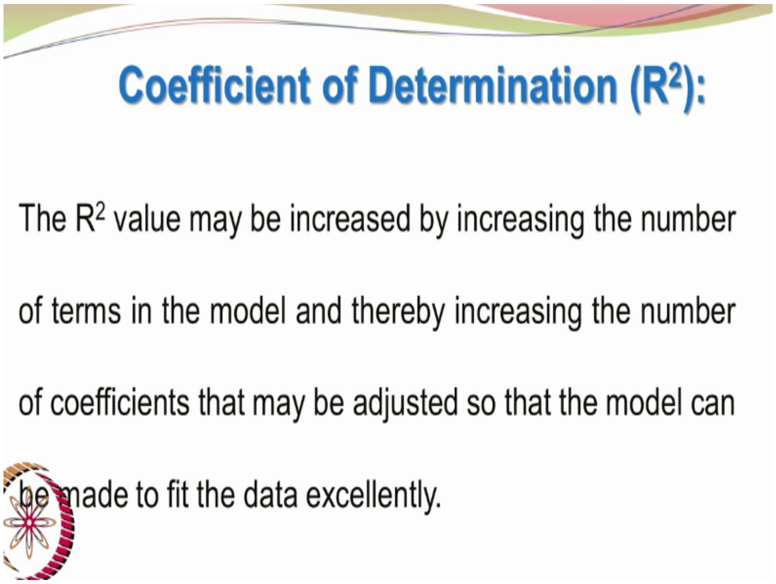
R square other than being the sum of squares of regression, the total sum of squares which does not really make sense to somebody who is not familiar with the subject R squares represents the proportion of the total variability accounted or explained by the linear regression model. So you have certain amount of variability in your process and what fraction or portion of the variability may be explained by your developed regression model.

If the variability is predominantly explained by your regression model, then the R square value would be quite close to 1 and you may have the satisfaction of developing a reasonably good model. So but there is a word of caution that may be added when we use R square. For example, if you have 5 data points and you fit a model with 5 parameters R squared will be = 1. All the variability would have been explained by the regression model.

Here you are not doing regression or linear regression curve fitting, in fact you are trying to fit 5 unknowns and you are having 5 equations and so essentially solving for 5 equations and 5 unknowns and obviously the 5 unknowns you are estimating should actually satisfy all the 5 equations. So all the variabilities account for and the R square value will be equal to 1 that is not acceptable.

We normally work with large data set let us say 40 or 50 data points and we try to fit only a few parameters 3 or 4 parameters. Unless the model is exact miraculously and the data have been generated with exactly no error which is very unlikely you will not have a situation where the 5 parameters in the regression model will be able to account for the responses of the 40 experimental sets. So there will always be some discrepancy.

(Refer Slide Time: 25:31)



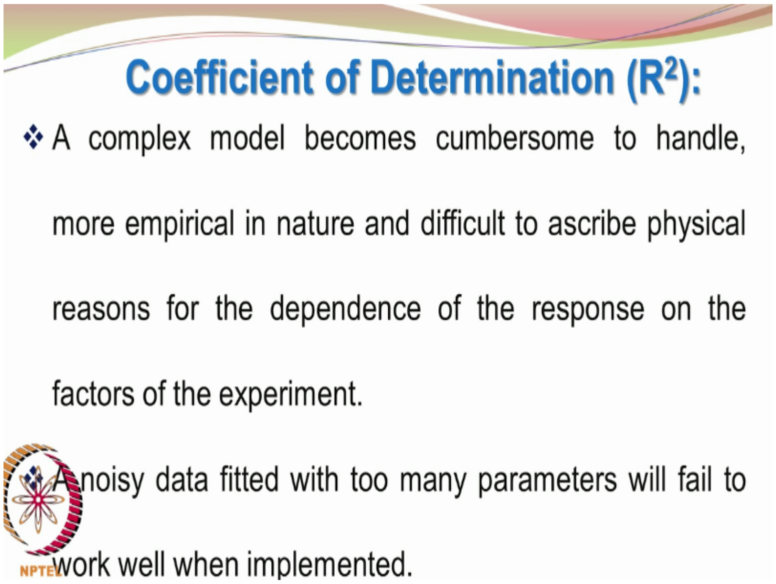
Coefficient of Determination (R^2):

The R^2 value may be increased by increasing the number of terms in the model and thereby increasing the number of coefficients that may be adjusted so that the model can be made to fit the data excellently.

So, how to increase the value of R square? The R square value may be increased by increasing the number of terms in the model and thereby increasing the number of coefficients that may be adjusted so that the model can be made to fit the data excellently. In the extreme case, if you fit a model for 40 experimental points with 40 parameters then you will get an exact fit.

But imaging having a model with 40 parameters it will run to half a page or full page and that model will look really ugly. We have to see what would be the best set of parameters which will give you a reasonably high value of R square. So what is the reasonable high value of R square? How do you quantify it? So again to sort of summarize what I have said so far.

(Refer Slide Time: 26:22)



Coefficient of Determination (R^2):

- A complex model becomes cumbersome to handle, more empirical in nature and difficult to ascribe physical reasons for the dependence of the response on the factors of the experiment.
- A noisy data fitted with too many parameters will fail to work well when implemented.

A complex model running to half a page or a single page becomes cumbersome to handle more empirical in nature and difficult to explain physically as to why this model is able to fit the data. What is the physical meaning of the model and so on? For example, if you have a temperature to the power of 3 or temperature to the power of 4 what is the physical reason that the response is affected by the fourth power of temperature? Is it radiation?

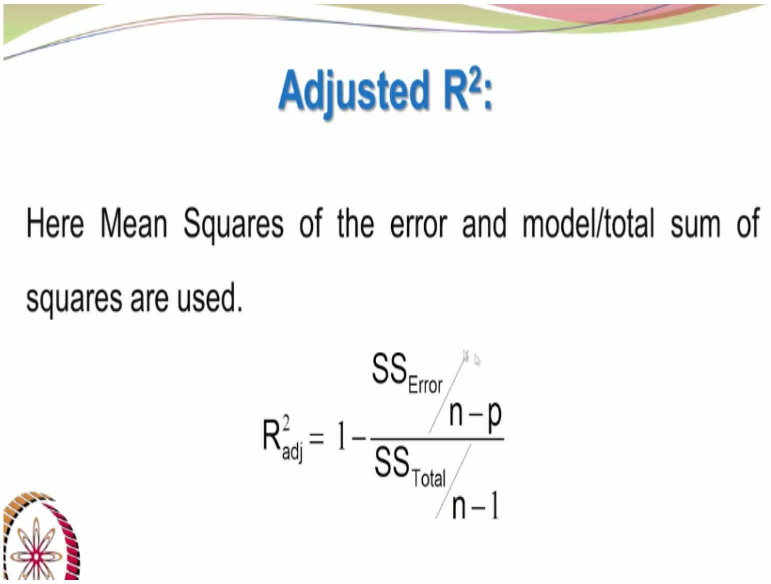
If it is not a radiation if it is a simple reaction problem why do you have power of T to the power of 4 as a simple illustration and also if your data is very noisy it is subject to lot of error then when you fit a model with many parameters to it, it may not be really successful when you slightly change the value of the regressor variables. For example, model was developed for the certain set of values of xj and why some model develops in the first place so that you do not have to keep on doing experiments time after time. Once you have a develop model, you can use it to represent the process in future design calculations or simulations and so on.

So you do not have to resort doing experimentation every time, but when you are having a noisy data and you have fitted a model with too many parameters, then when you change the value of x slightly or even use the same values of x you will find to a surprise that the model which was developed with so many parameters and worked well with those set of data may not be doing a good job with the new set of data. So this is a problem you may encounter often because your experiments are very variable.

And every time you cannot be fitting a new regression model to explain the particular set of experimental data. You would have a experimental data set collection that is the reason why you should do the experiments as carefully as possible trying to minimize the errors and unwanted errors or unavoidable errors you have to live with, but you should deliberately not introduce any systematic error in your experiments.

So you have to collect the data properly and fit a satisfactory regression model. You should not try to aim for regression coefficient of 1 all the time. Now that brings us to the concept of adjusted R square.

(Refer Slide Time: 29:09)



Adjusted R^2:

Here Mean Squares of the error and model/total sum of squares are used.

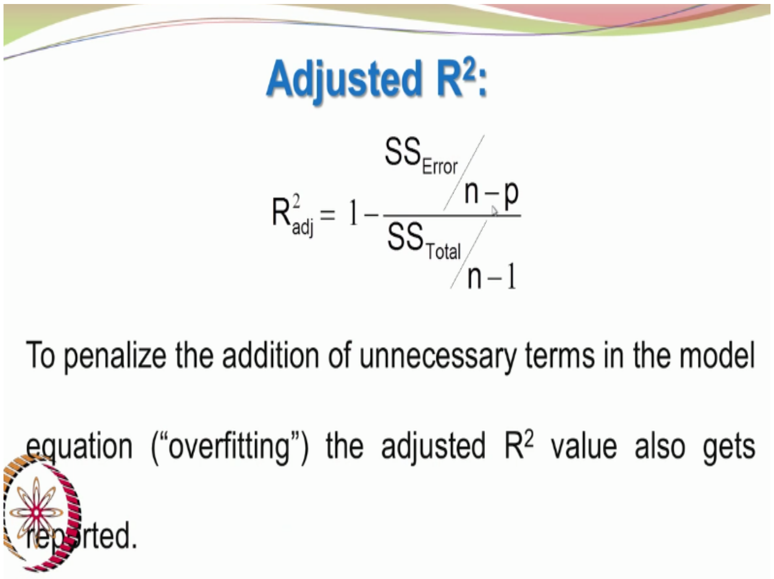
$$R^2_{adj} = 1 - \frac{SS_{Error}/n-p}{SS_{Total}/n-1}$$

And what we do is the concept is pretty much the same, it is the regression sum of squares/total sum of squares and the regression sum of squares may be written as total sum of squares - error sum of squares/total sum of squares so that is why you will get 1 - error sum of squares/total sum of squares. just back to the equation. This can be written as total sum of squares - error sum of squares and when you divide by sum of squares of total you will get 1 - sum of squares of error/sum of squares of total.

I think you can figure it out. This is total sum of squares - error sum of squares/total sum of squares. You just make the division and you will find it is equal to 1 - error sum of squares/total sum of squares and that is similar to what we have written here, but here we have scaled the sum of squares of error, by sum of squares total with the associated degrees of freedom. The degrees of freedom for sum of squares of error is n - p and the degrees of freedom for sum of squares of total is n - 1.

So here we scale sum of squares of error/n - p sum of squares total/n - 1. So rather than using sum of squares we are using mean square. There is a strong justification for scaling the sum of squares by the degrees of freedom. What is it?

(Refer Slide Time: 30:57)



Adjusted R^2:

$$R^2_{adj} = 1 - \frac{SS_{Error}/(n-p)}{SS_{Total}/(n-1)}$$

To penalize the addition of unnecessary terms in the model equation (“overfitting”) the adjusted R^2 value also gets reported.

So if you want this R square adjusted to be as close as 1 to be possible, then this term, the numerator term should be as small as possible. How will the numerator term be as small as possible? When either the sum of squares of error is very very small, or this n - p is quite high, but when you keep on adding more and more parameters this n - p term will become smaller and since this becomes smaller the numerator term will start to increase.

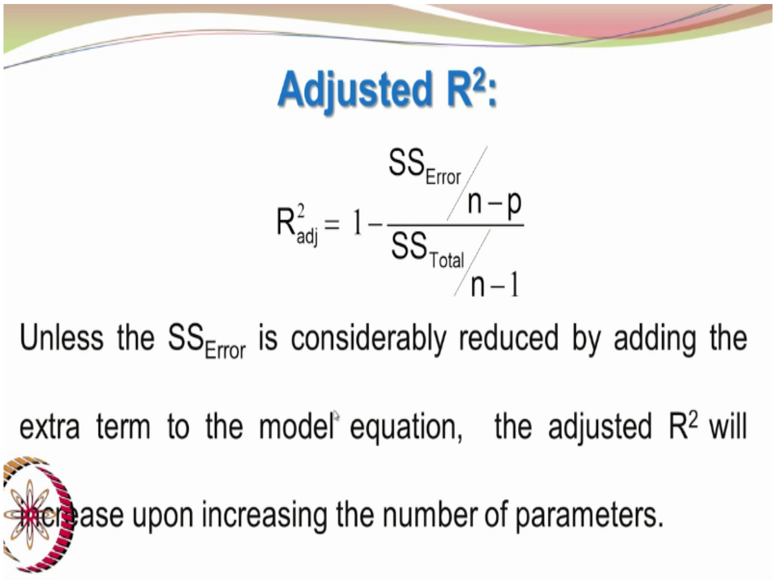
And so the adjusted R square will start to decrease, because when the numerator starts to increase, the R square adjusted will start to decrease so that is penalty for adding more and more parameters. Suppose you add a parameter which is having a strong influence on the process the sum of square of error will drastically reduce so even though the n - p has decreased by 1 the sum of squares of error has decreased even more considerably and so the overall effect would be to reduce the term on the other side of the negative sign.

So this term will decrease so the R square adjusted would be quite high, but on the other hand if the sum of squares of error decreases only by a small amount and you are adding many parameters to this, n - p will start to decrease very quickly and this will start to increase the numerator term and then the numerator term increases R square adjusted will decrease. So that is why we should not be in a hurry to keep on adding more and more regressor variables to our model just to get the R square to be 1.

So it is a good practice to look at the R square adjusted, also and see, whether it is satisfactory. Sometimes, I have seen cases where the regression R square value is 0.97 or 0.98 whereas the R square adjusted would be only 0.84 or 0.85. Suppose when you add the additional parameter, the regression coefficient goes to 0.975 or so, but the R square adjusted reduces to 0.83, then really the effect of the additional parameter is pretty much worthless. It is useless.

So please look at the variation of R square adjusted when you add more parameters rather than only looking at R square.

(Refer Slide Time: 33:50)



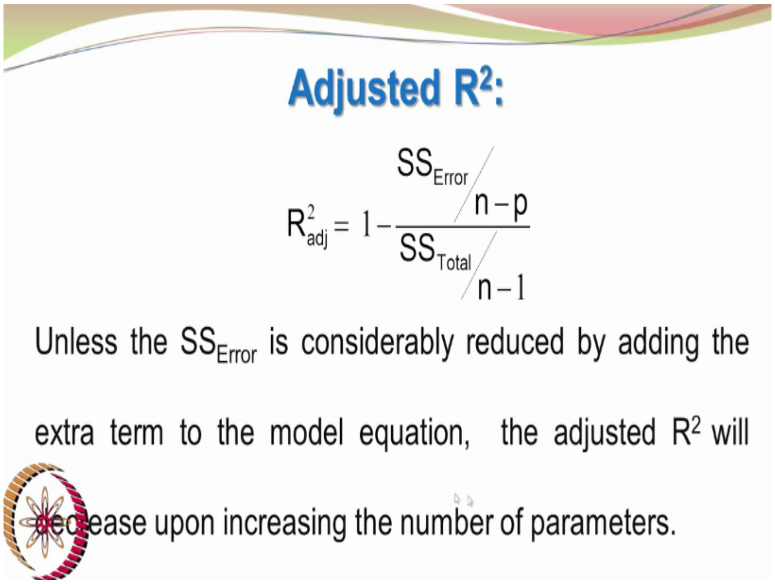
Adjusted R²:

$$R^2_{adj} = 1 - \frac{SS_{Error} / n-p}{SS_{Total} / n-1}$$

Unless the SS_{Error} is considerably reduced by adding the extra term to the model equation, the adjusted R^2 will increase upon increasing the number of parameters.

So as I said earlier, unless the sum of squares of error is considerably reduced by adding the extra term to the model equation, the R square adjusted will increase in the number of parameters. So there is a type here I will just correct it. The adjusted R square will decrease.

(Refer Slide Time: 34:19)



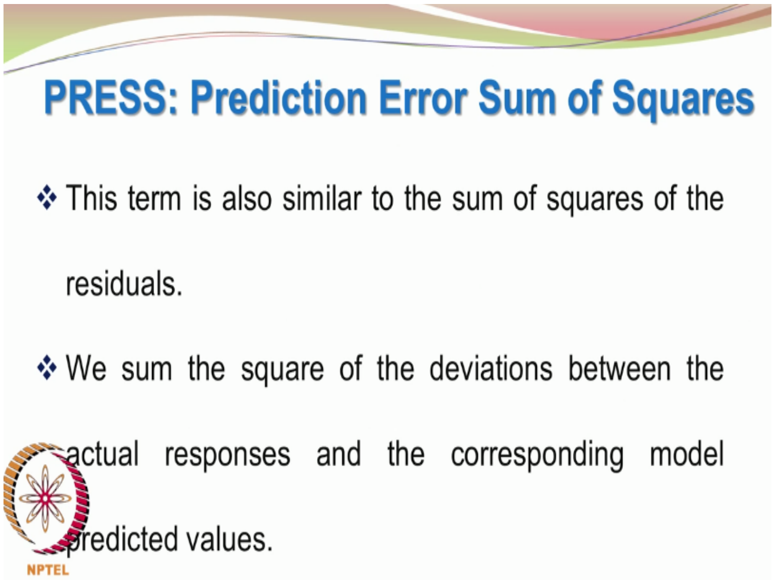
Adjusted R^2:

$$R^2_{adj} = 1 - \frac{SS_{Error}/n-p}{SS_{Total}/n-1}$$

Unless the SS_{Error} is considerably reduced by adding the extra term to the model equation, the adjusted R^2 will decrease upon increasing the number of parameters.

So, unless the sum of squares is considerably reduced by adding the extra term to the model equation, the adjusted R square will decrease upon increasing the number of parameters.

(Refer Slide Time: 34:28)

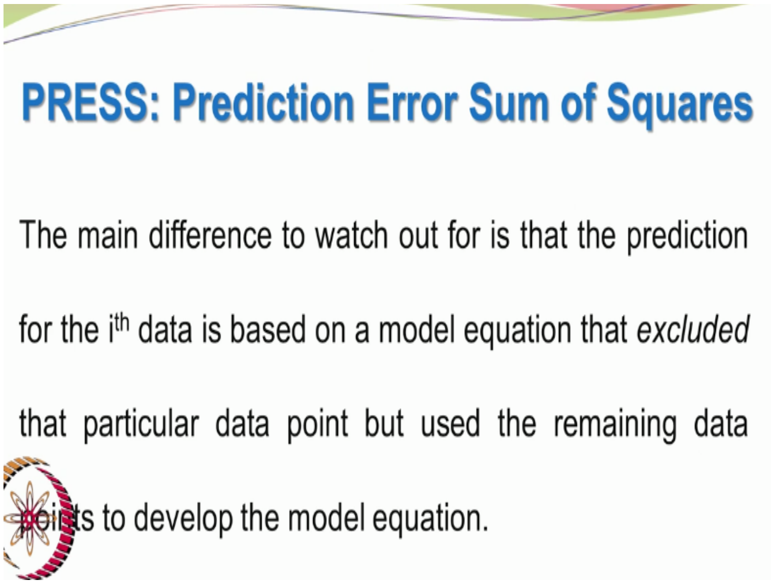


PRESS: Prediction Error Sum of Squares

- This term is also similar to the sum of squares of the residuals.
- We sum the square of the deviations between the actual responses and the corresponding model predicted values.

Now we come to another term called as the prediction error sum of squares and this is called as press, sum of the computer outputs also report this value and this term is also similar to the sum of squares of the residual what we do is we sum the square of the deviations between the actual responses and the corresponding model predicted values. What is the difference here? You are having the actual response in the corresponding mode prediction, the difference between the 2 is squared. So this is also looking like some residual sum of squares. So what is nu here impress?

(Refer Slide Time: 35:16)



PRESS: Prediction Error Sum of Squares

The main difference to watch out for is that the prediction for the i^{th} data is based on a model equation that *excluded* that particular data point but used the remaining data points to develop the model equation.

We will see, so the main differences that the prediction for the ith data is based on a model equation that excluded that particular data point, but use the remaining data points to develop the model equation. So when you are considering the error sum of squares or the residual sum of squares for a particular ith data, obviously you are going to subtract the response with the model predicted value and square it. So what is the difference here?

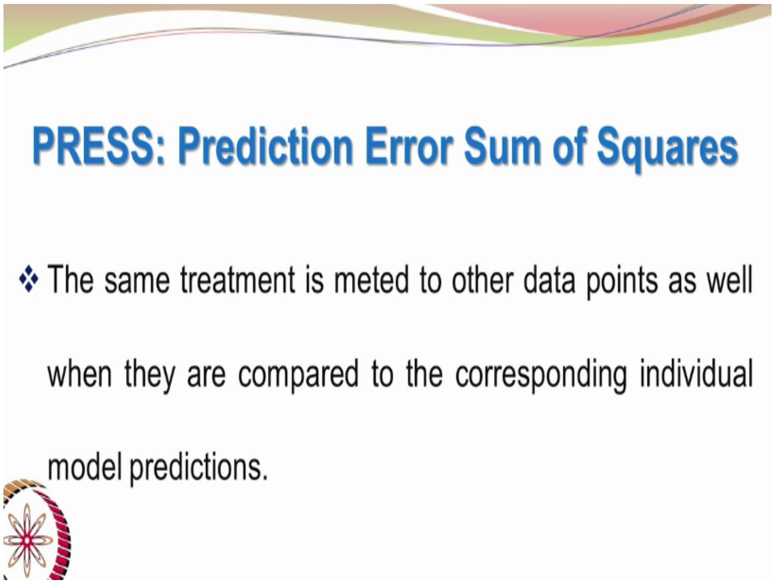
The main difference is when you are looking at the residual sum of squares for the ith data point, the prediction is based on a model that actually excluded the ith data point, for example if I am calculating the residual sum of squares for the first data point, I would develop a model with the remaining data points and I would have a model which did not use the first data point. Then I will use the model to predict the response for the first experimental data point.

Then the difference between the experimental value and the prediction value is squared to give the residual sum of squares or the error sum of squares for the first data point. Now when I am going to the next second data point, I will first develop a model without the second data point. So I will have a model equation. Then I will subtract the experimental response for the second data point with the model prediction based on the remaining data points except the second.

All other data points, then that model prediction is subtracted from the second experimental data point response and that is squared. Similarly, I do it for all the remaining data points in the set.

This may look to be a bit tedious, but there are ways in which this can be done much faster, but that is beyond the scope of this course. So we want the press value to be also quite small.

(Refer Slide Time: 37:25)

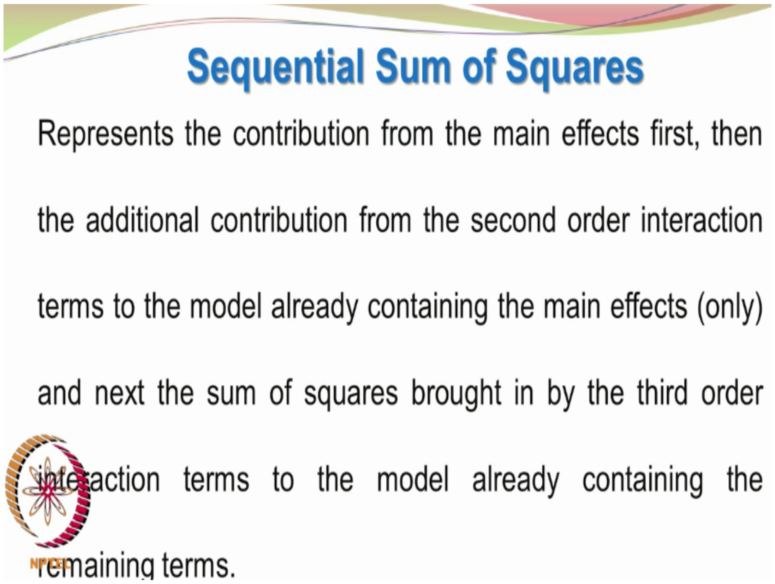


PRESS: Prediction Error Sum of Squares

- ❖ The same treatment is meted to other data points as well when they are compared to the corresponding individual model predictions.

So that is why let me sort of summarize the main difference to watch out for is that the prediction for the ith data is based on a regression model equation that excluded that particular data point, but use the remaining data points to develop the model equation. So the same treatment is meted out to other data points as well when they are compared to the corresponding individual model predictions.

(Refer Slide Time: 37:51)



Sequential Sum of Squares

Represents the contribution from the main effects first, then the additional contribution from the second order interaction terms to the model already containing the main effects (only) and next the sum of squares brought in by the third order interaction terms to the model already containing the remaining terms.

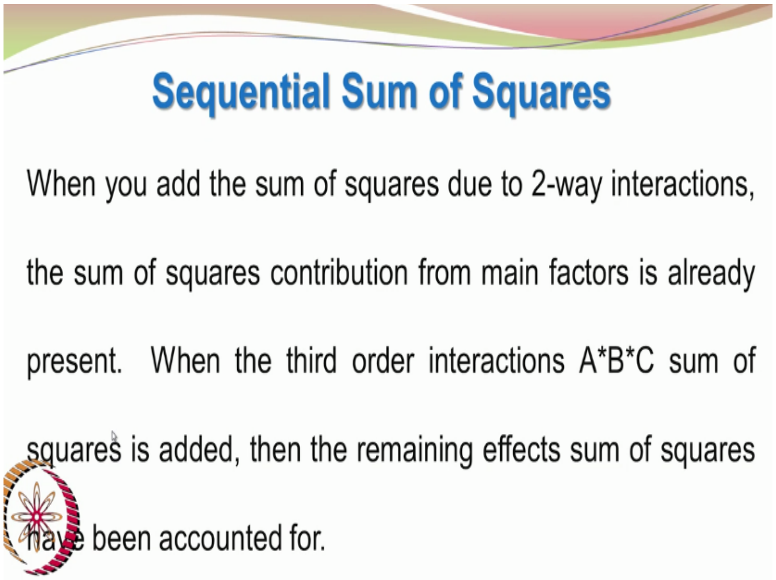
So earlier, we were looking at prediction error sum of squares or the press now you will be looking at sequential sum of squares. As the name implies it is the gradual model development focusing on first the main effects, then the second order effects or in other words the product of factors taken 2 at a time. So once we are done with the main effects, then we consider the effect of adding factors 2 at a time to a model already containing the main effects.

And once we have done that so we have now a model with main effects and then the second order interactions or product of 2 factors. Suppose you have model A with main effects A, B, and C. First you develop a model with only main effects A, B, and C then you will look out the second order effects AB, BC, AC, and after having developed this model then consider the effect of third order interactions which would be ABC.

So you are developing the model as main effects second order interactions and then the third order interaction. So what we do is represents the contribution to the total sum of squares and the main effects, then the additional contribution from second order interactions to the model, already containing the main effects and next the sum of squares brought in by the third order interactions to the model already containing the remaining terms.

So we can gradually see that there would be less and less impact to total sum of squares by higher order terms. In some cases, the interactions may be contributing to the total sum of squares more than the main effects, but beyond second order interactions may be third order interactions the higher and higher order interaction would be contributing negligibly to the total sum of squares and their value would really not be there. So it is another way of telling that do not develop a model beyond the third order interaction.

(Refer Slide Time: 40:12)

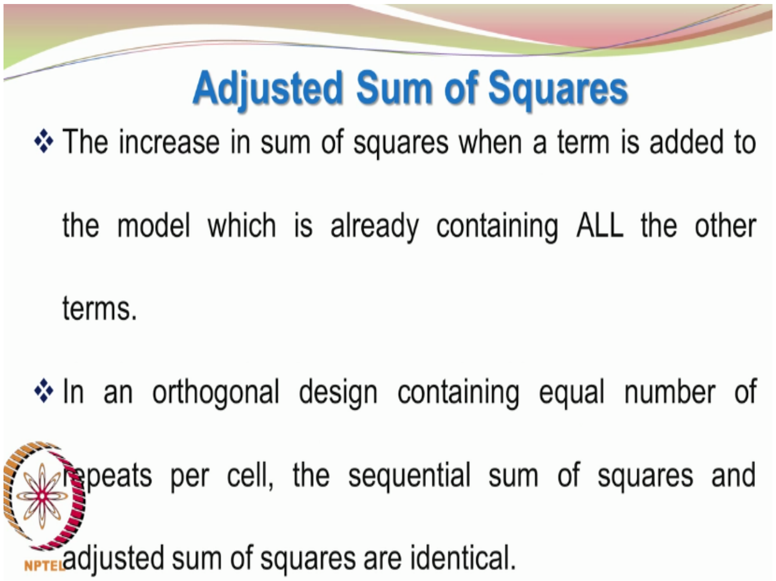


Sequential Sum of Squares

When you add the sum of squares due to 2-way interactions, the sum of squares contribution from main factors is already present. When the third order interactions A*B*C sum of squares is added, then the remaining effects sum of squares have been accounted for.

So repeating what I said when you add the sum of squares due to 2-way interactions the sum of squares contribution from main factor is already present when the third order interaction A, B, C sum of squares is added then the remaining effects sum of squares have been accounted for and the remaining effects meaning the main effects and the second order interactions.

(Refer Slide Time: 40:33)



Adjusted Sum of Squares

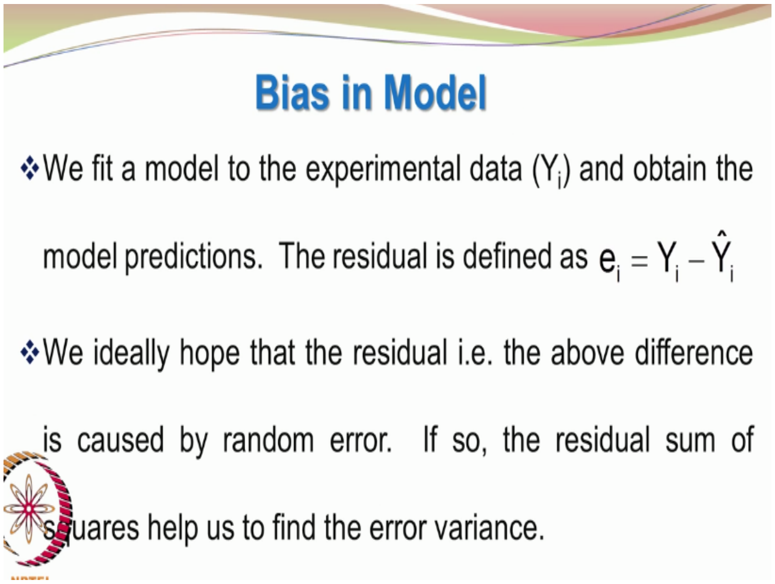
- The increase in sum of squares when a term is added to the model which is already containing ALL the other terms.
- In an orthogonal design containing equal number of repeats per cell, the sequential sum of squares and adjusted sum of squares are identical.

Now we come to another term, called as the adjusted sum of squares, this represents the increase in sum of squares when the term is added to the model which is already containing all the other terms. The adjacent sum of squares is different from sequential sum of squares. The sequential sum of squares as the name implies we are doing it sequentially, systematically in an organized fashion. So what we do is we develop a model without a main effect A let us say.

So we develop a model with B, C then we even do AB, BC, AC, then we also do ABC, then we finally add the factor or regressor variable A at the very end and see the regression sum of squares brought in by it. So this is the increase in sum of squares when a term is added to the model which is already containing all the other terms. In an orthogonal design containing equal number of repeats per cell the sequential sum of squares and adjusted sum of squares are identical.

This is another beauty of the orthogonal designs. The statistically designed experiments, the factorial design of experiments are usually orthogonal and so you have the advantage of the adjusted sum of squares being = the sequential sum of squares. So it does not really matter whether the factor is added in the beginning or in the end it is contributing to the sum of squares in an identical fashion. But in nonorthogonal designs you can even note that the sequential sum of squares and adjusted sum of squares are not the same.

(Refer Slide Time: 42:25)



Bias in Model

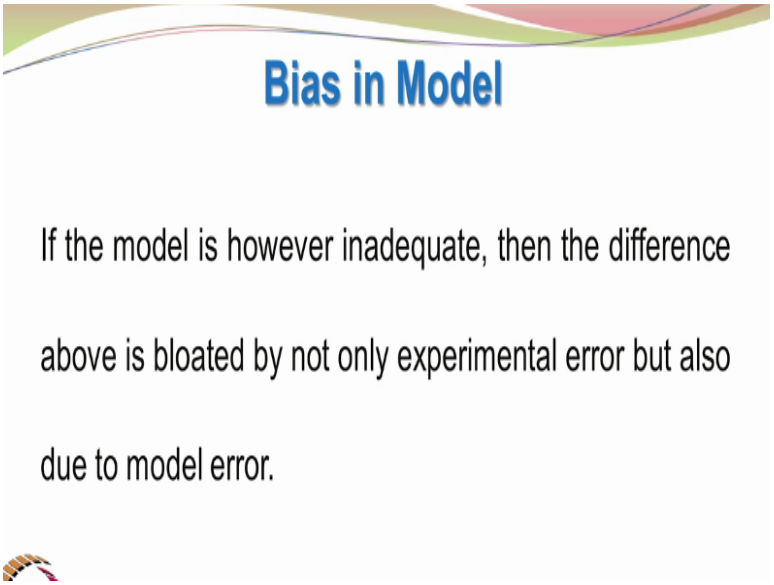
- ❖We fit a model to the experimental data (Y_i) and obtain the model predictions. The residual is defined as $e_i = Y_i - \hat{Y}_i$
- ❖We ideally hope that the residual i.e. the above difference is caused by random error. If so, the residual sum of squares help us to find the error variance.

Now we look at the term Bias in the model. So what we do is we fit a model to the experimental data yi and obtain the model predictions. The residual we know by now is Yi - Y hat and we hope that the residual which is defined above is only caused by random error. If so, the residual sum of squares helps us to find the error variance. This is a very important concept. Whatever we are unable to explain by the model we really hope that it is because of random effects only.

But if you think a bit deeper the difference between the experiment value and the model prediction may not always be due to experimental error alone may be you have not developed a sufficiently acceptable model, may be the person who was doing the model development was very lazy and when there are 2 factors or 2 factors and interactions affecting the modular influencing the model.

He might have taken the easy way out and developed a regression equation with only one regressor variable even though 2 regressor variables and the interactions are influencing the process physically. In such a situation you cannot argue that the discrepancy between the experimental value and the model prediction is only because of random error. It can be also because of an inadequate model. This is what we are going to discuss from the slides on.

(Refer Slide Time: 44:04)

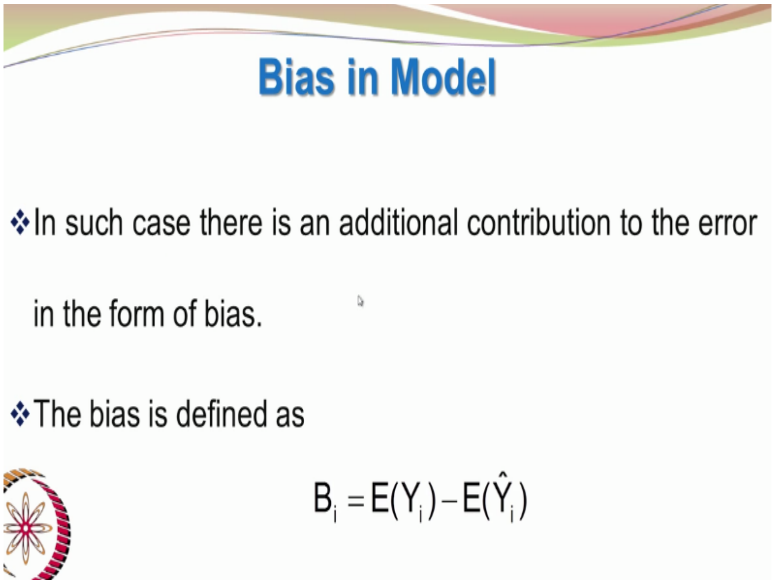


Bias in Model

If the model is however inadequate, then the difference above is bloated by not only experimental error but also due to model error.

So if the model is however inadequate that is very important. The model is however inadequate, then the difference above is bloated or increased by not only experimental error, but also due to model error. This is very important. So you are having experimental error, random error and then you also have the model error so how do you split or dealing it the 2 errors. The residual sum of squares is containing both the modeling and also the random error, how do you want to split them.

(Refer Slide Time: 44:44)



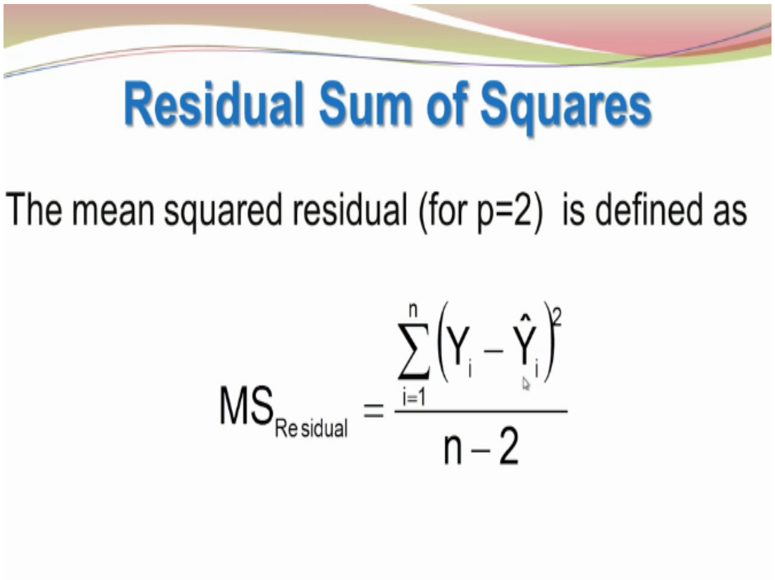
Bias in Model

- ❖In such case there is an additional contribution to the error in the form of bias.
- ❖The bias is defined as

$$B_i = E(Y_i) - E(\hat{Y}_i)$$

For that purpose, we define a bias and call it as the expected value of the experimental response and the expected value of the model prediction for the ith experimental condition. Expected value of the experimental response is matching with the expected value of the model prediction then the bias will be = 0. On the other hand, if the expected value of the experiment response is different from the expected value of the model prediction then you have a nonzero bias.

(Refer Slide Time: 45:28)



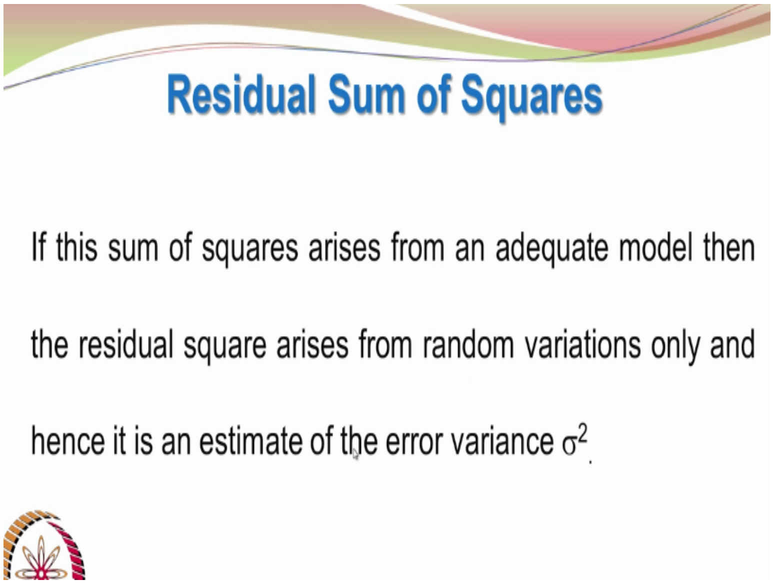
Residual Sum of Squares

The mean squared residual (for p=2) is defined as

$$MS_{Residual} = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n-2}$$

So when you have mean square residual let us say P = 2 for convenience you are having only 2 parameters then the mean square residual is given by I = 1 to n yi - y predicted y, y hat I whole square/n - p or n - 2. This is the mean square residual.

(Refer Slide Time: 45:53)



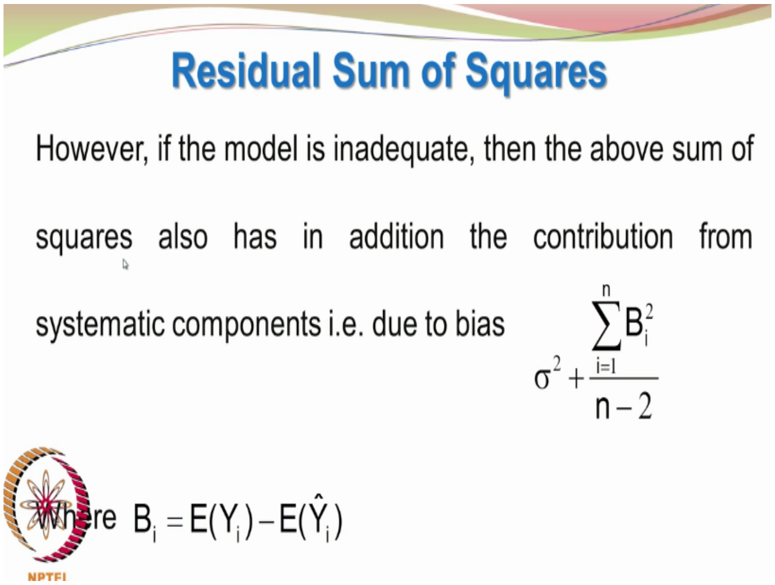
Residual Sum of Squares

If this sum of squares arises from an adequate model then the residual square arises from random variations only and hence it is an estimate of the error variance σ^2.

If this sum of squares arises from an adequate model, then the residual squares arises from random variations only and hence it is an estimate of the error variance sigma square.

We do not know the error variance sigma square. So we are hoping to the residual sum of squares will give us an idea or an estimate about the error variance, but if the residual sum of squares is also having the variation due to an inadequate model, then we cannot use the residual sum of squares to get a good idea or a good estimate on the experimental error. The mean sum of squares will be higher than the experimental error contribution. So we have to be careful.

(Refer Slide Time: 46:37)



Residual Sum of Squares

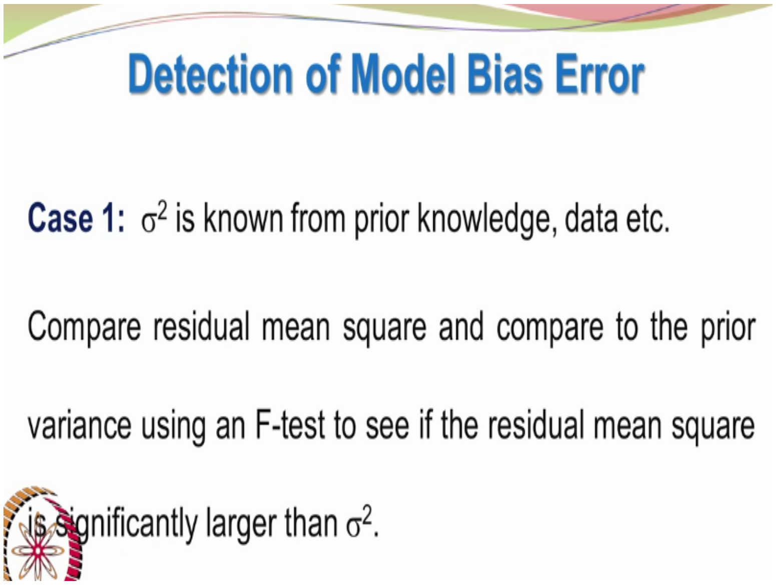
However, if the model is inadequate, then the above sum of squares also has in addition the contribution from systematic components i.e. due to bias

$$\sigma^2 + \frac{\sum_{i=1}^{n} B_i^2}{n-2}$$

Where $B_i = E(Y_i) - E(\hat{Y}_i)$

However, if the model is inadequate, then the above sum of squares, the residual sum of squares also has an addition the contribution from systematic components i.e. due to bias. So we have sigma square which is the error variance and + sigma Bi square/n - 2. So the residual sum of squares has contribution from sigma square and the bias contribution.

(Refer Slide Time: 47:12)



Detection of Model Bias Error

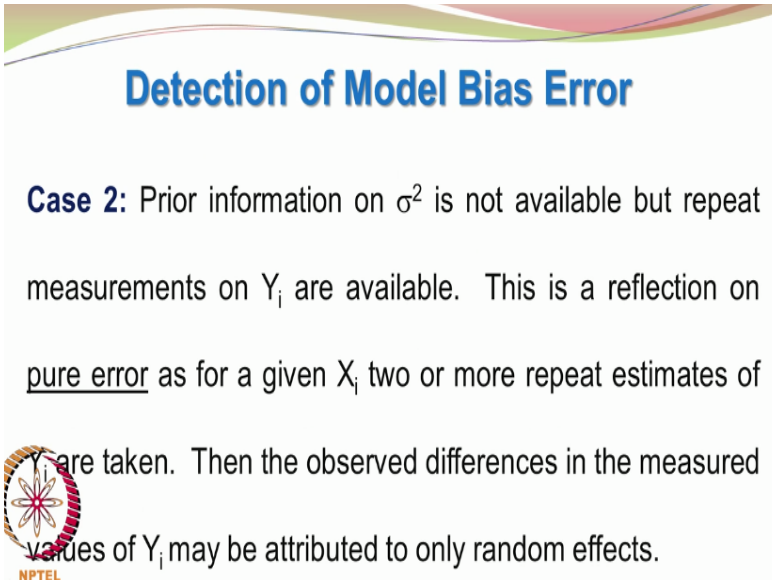
Case 1: σ^2 is known from prior knowledge, data etc.

Compare residual mean square and compare to the prior variance using an F-test to see if the residual mean square is significantly larger than σ^2.

How to find out whether we are having an adequate model or an inadequate model? So what we do here is let us say that we know sigma square from prior knowledge or from experience or previous data sets and so on, so you have a fair idea about sigma square. So what we do is compare the residual mean square that is sigma yi - y hat I whole square/n - 2 or n - p and compared to the prior variance using an F test to see if the residual mean square is significantly larger than sigma square.

So you compare the residual sum of squares/degrees of freedom with sigma square and then see whether the residual sum of squares/n - p is comparable to sigma square and for this case if it is statistically significant, the residual mean square cannot be statistically = sigma square and then the model is said to have a lack of fit. So we should reconsider the model as it is inadequate in the present form.

(Refer Slide Time: 48:31)



Detection of Model Bias Error

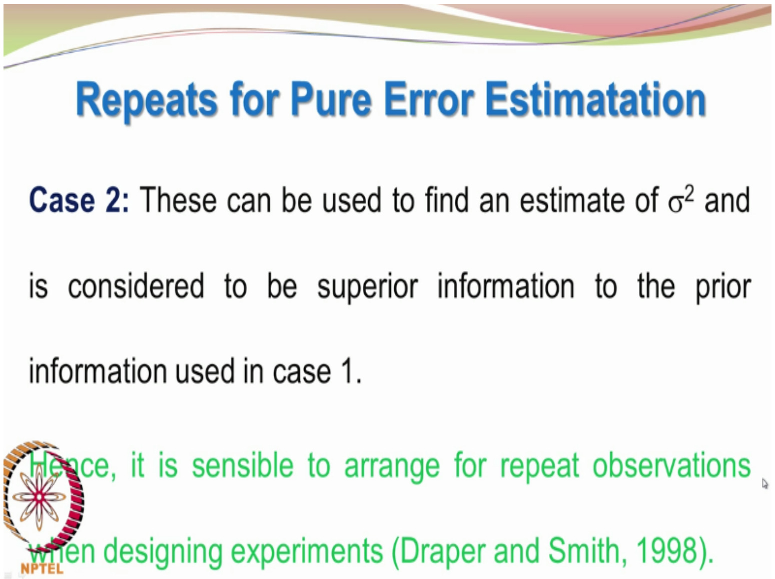
Case 2: Prior information on σ^2 is not available but repeat measurements on Y_i are available. This is a reflection on pure error as for a given X_i two or more repeat estimates of Y_i are taken. Then the observed differences in the measured values of Y_i may be attributed to only random effects.

On the other hand, if you do not have information on sigma square which is usually the case, but repeat measurements on yi are available. This is another reason why you should perform repeats in your experiment. So when you have repeat measurements, this is a reflection on the pure error, because when you repeat experiments you are not going to get the identical response. You will be having different values of the response for repeated experiments.

So this you can use to obtain an idea about the random fluctuations or the random variations the sum of squares caused by true random variations. So we can even call it as sum of squares due to pure error, because the repeats represent pure error and we are hoping that when you repeat the experiments you are making sure that all the variables are kept at their assigned values in all the runs. Even if 1 value of the variable changes slightly then it cannot be called as a genuine repeat.

So what do you do is repeat measurements on yi are available and this is the reflection of pure error or unadulterated error as for a given xi 2 or more repeat estimates of yi are taken. Then the observed differences in the measured values of yi may be attributed only to random effects.

(Refer Slide Time: 49:54)



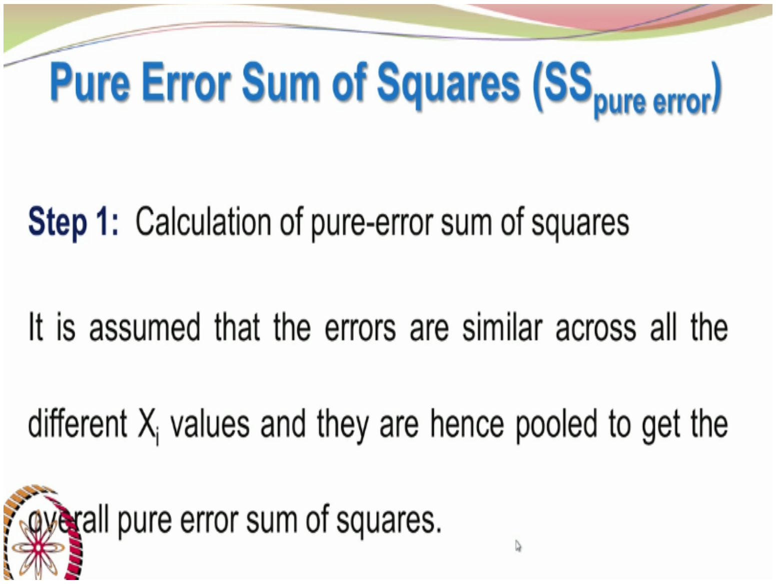
Repeats for Pure Error Estimatation

Case 2: These can be used to find an estimate of σ^2 and is considered to be superior information to the prior information used in case 1.

Hence, it is sensible to arrange for repeat observations when designing experiments (Draper and Smith, 1998).

So continuing with the case to where we do not know sigma square and this is usually more often the case, it is very essential to have repeat experiments in our program or plan and this is brought out very nicely by Draper and Smith in year 1998 book.

(Refer Slide Time: 50:20)



Pure Error Sum of Squares ($SS_{pure\ error}$)

Step 1: Calculation of pure-error sum of squares

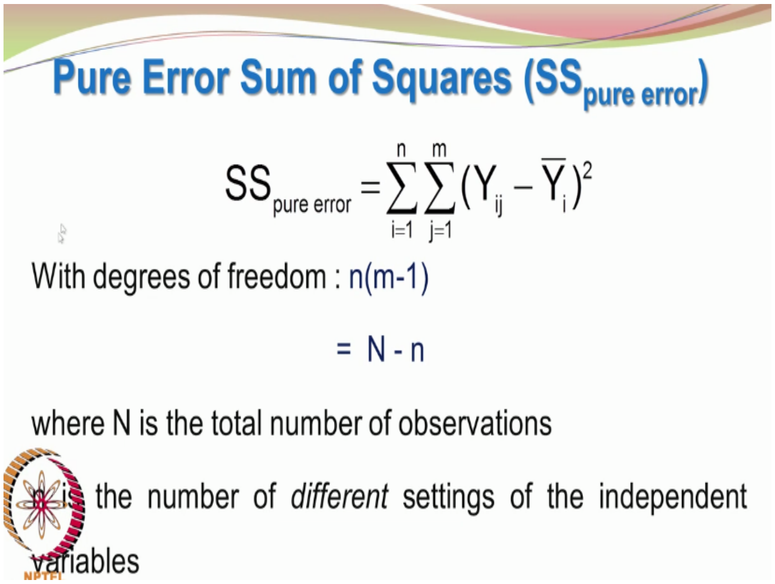
It is assumed that the errors are similar across all the different X_i values and they are hence pooled to get the overall pure error sum of squares.

Now we are now looking at pure-error sum of squares. So we call the sum of squares pure error and what we do is, we have repeated experiments across different experimental settings. First, experimental setting combination we do repeats. Then the next experimental setting combination we do repeats. So we will assume that the errors in the first repeat is for the first experimental setting the repeats are analogous to the repeats in the second experimental setting in the sense that the errors which influence the second settings are random.

And they also are identical with the errors that represent the first experimental setting that means all the errors which are influencing the first experimental setting and the errors influencing the second experimental setting come from the same family. You know that the errors are distributed normally with 0 mean and variance sigma square. So the error variance in the first experimental setting is identical with the error variance in the second experimental setting.

So that is what I mean. So with this assumption, we are going to pool the sum of the pure error so that we get better overall estimate.

(Refer Slide Time: 51:51)



Pure Error Sum of Squares ($SS_{pure\ error}$)

$$SS_{pure\ error} = \sum_{i=1}^{n}\sum_{j=1}^{m}(Y_{ij} - \overline{Y}_i)^2$$

With degrees of freedom : n(m-1)

= N - n

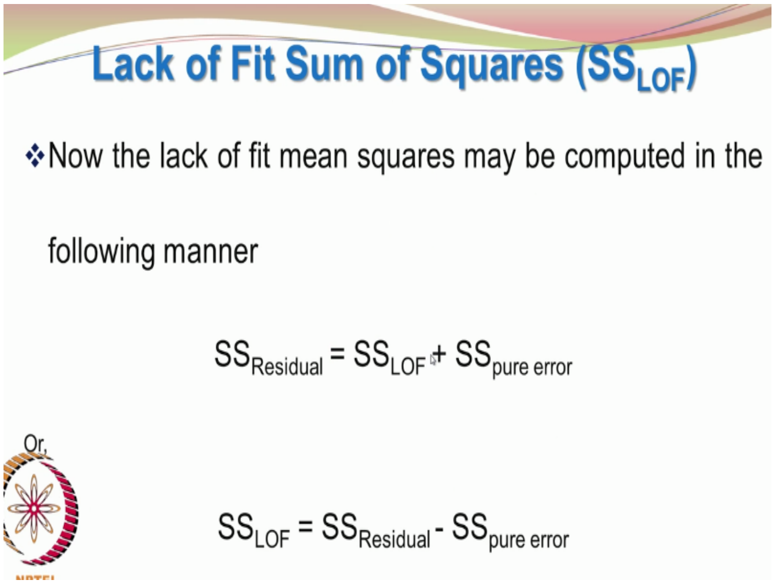
where N is the total number of observations

n is the number of *different* settings of the independent variables

So we have sum of squares of pure error is given by I = 1 to n, j = 1 to n yij - y bar I whole square. So you are having m repeats that means the repeats may be different for every experimental setting, but let us assume that normally we conduct same number of repeats for every experimental setting. Let us call that value as m. So j represents the repeat and I represents the experimental setting.

So we have sum of squares of pure error is yij - y bar I whole square and then we do it for all the experimental settings. The degrees of freedom by now you should be able to show that it is n times m - 1. Every set of repeated experiments would have a degrees of freedom of m - 1 assuming me to be same for all experimental settings. So n * m - 1 will become N - n where n * m is the total number of runs and N is the number of independent experimental settings.

(Refer Slide Time: 53:03)



Lack of Fit Sum of Squares (SS_{LOF})

- Now the lack of fit mean squares may be computed in the following manner

$$SS_{Residual} = SS_{LOF} + SS_{pure\ error}$$

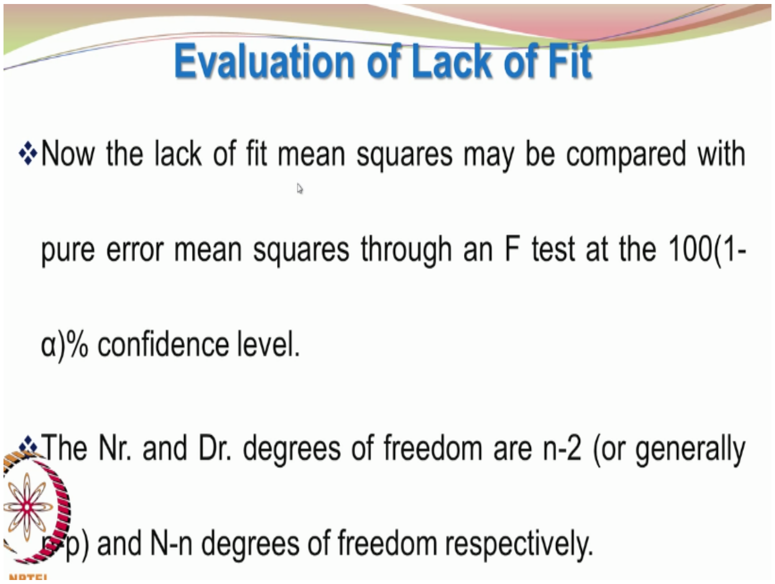
Or,

$$SS_{LOF} = SS_{Residual} - SS_{pure\ error}$$

Now the lack of fit sum of square is a very important quantity can be defined as follow. Sum of squares of residuals = sum of squares of lack of fit + sum of squares of pure error. The residual sum of squares is now split into lack of fit sum of squares and pure error sum of squares. Just now you have found the pure error sum of squares. The pure error sum of squares was found by using this relationship.

Now you can get the sum of squares of lack of it by simply subtracting the residual sum of squares with the sum of squares of pure error so that will give you the sum of squares for lack of fit.

(Refer Slide Time: 53:38)



Evaluation of Lack of Fit

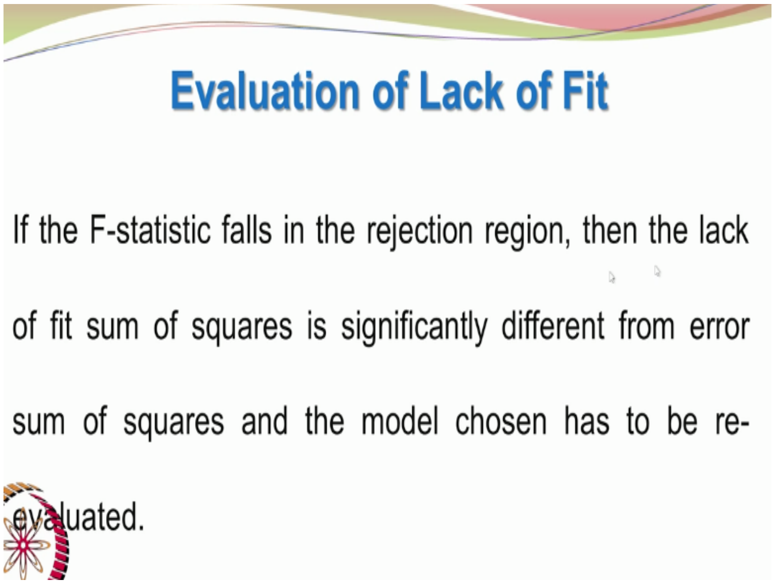
- ❖Now the lack of fit mean squares may be compared with pure error mean squares through an F test at the 100(1-α)% confidence level.
- ❖The Nr. and Dr. degrees of freedom are n-2 (or generally n-p) and N-n degrees of freedom respectively.

Now you can take the lack of fit mean squares with the pure error mean squares through an F test at the 100 * 1 - alpha percent confidence level. So you can easily find the degrees of freedom for the lack of fit sum of squares. You know that the sum of squares of pure error have a degrees of freedom of N - n please not capital N - small n. the residual sum of squares have the degrees of freedom of n - p where n is the total number of runs and p is the number of parameters.

So you have n - p and you have capital N - p and you have capital N - small n. The difference between the degrees of freedom for the sum of squares of residual and the sum of squares of pure error will give you degrees of freedom for the lack of fit. I request you to work it out yourself. So the lack of fit mean squares may be compared with the pure error mean square through a F test at the 100 * 1 - alpha% confidence level.

The numerator and denominator degrees of freedom are n - 2 or generally n - p and N - n degrees of freedom respectively. So the lack of fit sum of squares will have a degrees of freedom of N - p were small n is the number of independent settings and p is the number of parameters and the pure error sum of squares will have the degrees of freedom of capital N - small n. Capital N is the total number of runs and small n is the number of independent settings.

(Refer Slide Time: 55:28)

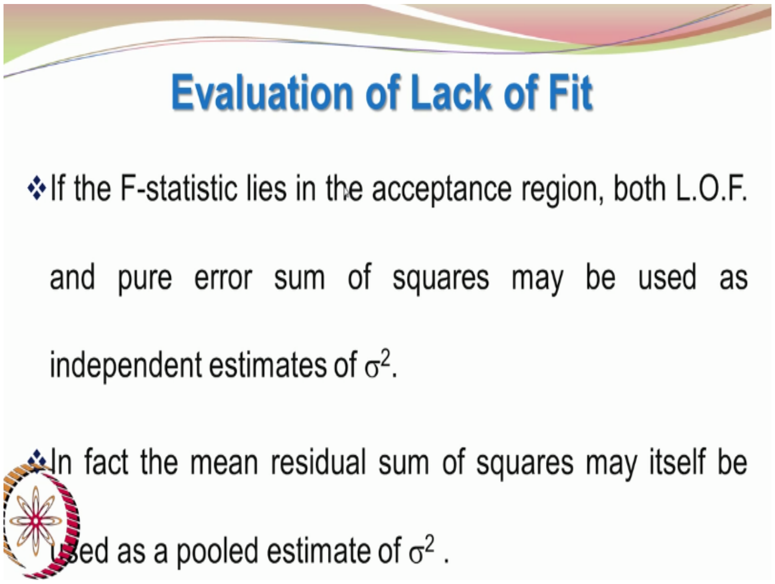


Evaluation of Lack of Fit

If the F-statistic falls in the rejection region, then the lack of fit sum of squares is significantly different from error sum of squares and the model chosen has to be re-evaluated.

If the F statistics falls in the rejection region, then the lack of fit sum of squares is significantly different from the error sum of squares and the model chosen has to be re-evaluated.

(Refer Slide Time: 55:38)



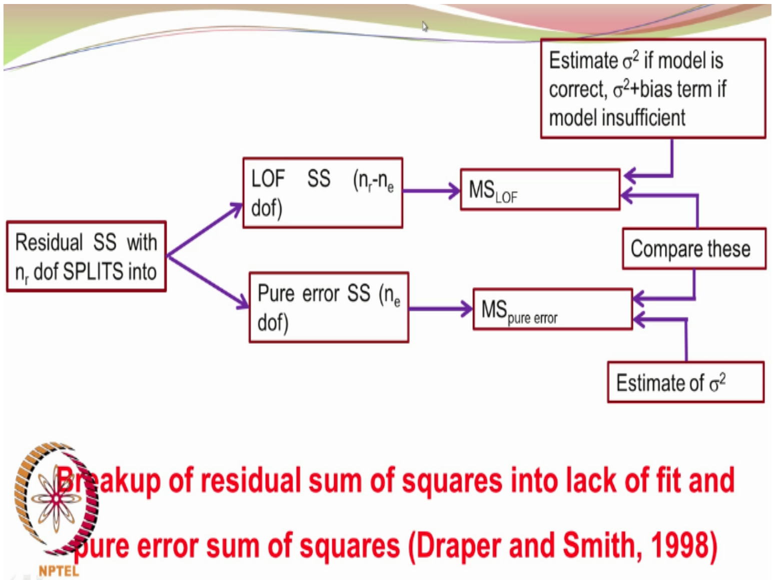
Evaluation of Lack of Fit

- If the F-statistic lies in the acceptance region, both L.O.F. and pure error sum of squares may be used as independent estimates of σ^2.
- In fact the mean residual sum of squares may itself be used as a pooled estimate of σ^2 .

If the F statistic lies in the acceptance region, both the lack of fit and pure error sum of squares may be used as independent estimates of sigma squares. In fact, the mean residual sum of squares itself may be then used as a pooled estimate of sigma square. If the sigma squared is based upon more data points, then that is better. In fact, if the degrees of freedom associated with the error estimation is higher than that error estimation is more valuable.

So the sigma squared based on the residual sum of squares if the model is adequate is recommended as a surrogate for finding sigma square. What I am trying to say is the mean square residual can be used as an estimate of sigma squared because it is having higher degrees of freedom. This is only to be done when the model which is fitted is adequate. If the model is not adequate, then you cannot use the residual mean square as an estimate for sigma square.

(Refer Slide Time: 56:49)



Residual SS with n_r dof SPLITS into

LOF SS (n_r-n_e dof)

Pure error SS (n_e dof)

MS_{LOF}

$MS_{pure\ error}$

Estimate σ^2 if model is correct, σ^2+bias term if model insufficient

Compare these

Estimate of σ^2

Breakup of residual sum of squares into lack of fit and pure error sum of squares (Draper and Smith, 1998)

So this may be nicely represented by flow diagram again given by Draper and Smith. So the residual sum of squares is split into nr residual degrees of freedom, it splits into lack of fit sum of squares and pure error sum of squares. Let us call the degrees of freedom associated with pure error sum of squares as ne and lack of fit sum of squares is nr - ne. So when you divide the sum of squares with the respective degrees of freedom you get the mean square lack of it.

And mean square pure error and then you compare the mean square lack of fit and the mean square pure error. The mean square lack of fit is an estimate of sigma square if the model is correct and sigma square + bias term if the model is insufficient and the mean square pure error is obviously going to be a reliable and true estimate of sigma square. Right this completes our discussion on the various aspects of linear regression, the various terms you may often encounter in a regression output hopefully after this lecture.

You will be able to appreciate the value of the different terms in the regression output rather than basing your judgment solely on the R square value. In fact, when you are explaining your results to your thesis supervisor or to your boss in the company or to your R and D manager it will make a good impression if you are able to provide more insight into the developed regression model equation.

Having said that please note that these regression model equations have not been really based on first principals and it is only an empirical equation, but still if it can represent the effect of various variables and the interaction between the variables in a reliable manner, the developed regression model is very useful, because many times in real life we cannot always model the processes from first principals. So this completes our discussion and thanks for your attention.