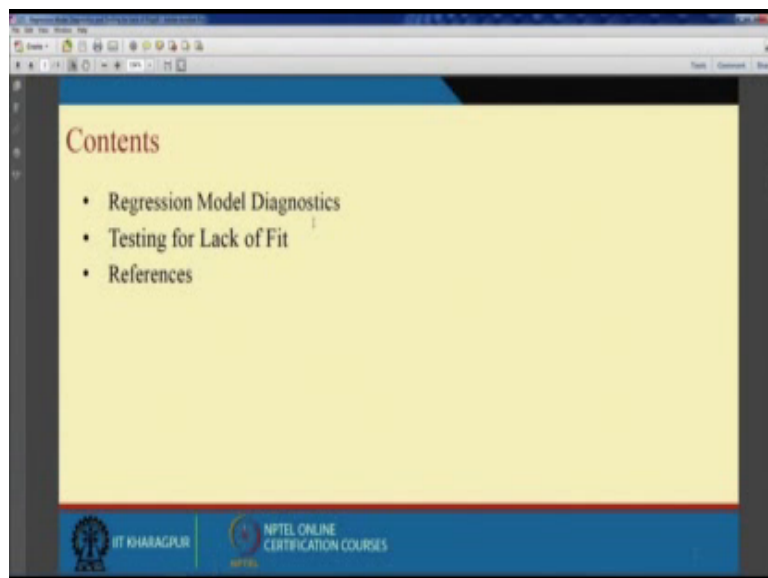


**Design and Analysis of Experiments**  
**Prof. Jhareswar Maiti**  
**Department of Industrial and Systems Engineering**  
**Indian Institute of Technology, Kharagpur**

**Lecture – 23**  
**Multiple Linear Regression: Model Diagnostics and Testing for Lack of Fit**

Welcome in this lecture we will see some of the diagnostics for multiple linear regression and one more concept we will introduce which is known as lack of fit and test for lack of fit.

(Refer Slide Time: 00:42)



Essentially the diagnostics measures, lack of fit and its tests will be discussed.

(Refer Slide Time: 00:49)

**Regression Model Diagnostics**

◆ **Standardized residual**

$$d_i = \frac{e_i}{\hat{\sigma}} \quad i = 1, 2, \dots, n \quad \text{where } \hat{\sigma} = \sqrt{MSE}$$


- Standardized residuals have mean=0 and variance is nearly 1.
- Most of them lie in the interval  $-3 \leq d_i \leq 3$

$\hat{y} = X\hat{\beta}$        $e = y - \hat{y}$   
 $= X(X'X)^{-1}X'y$        $Cov(e) = \sigma^2(I - H)$   
 $= Hy$

where  $H = X(X'X)^{-1}X'$  is called hat matrix.  
 It maps the vector of observed values into a vector of fitted values.

The variance of the  $i$ -th residual  $V(e_i) = \sigma^2(1 - h_{ii})$

Studentized residuals  $r_i = \frac{e_i}{\sqrt{\hat{\sigma}^2(1 - h_{ii})}} \quad i = 1, 2, \dots, n$  where  $\hat{\sigma}^2 = MSE, V(e_i) = 1$



And for model diagnostics we will be using standardized residuals.

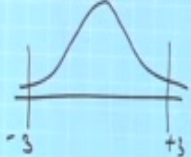
(Refer Slide Time: 00:59)

Multiple Linear Regression

- ✓ Standardized residuals
- ✓ Studentized residuals
- ✓ PRESS statistics.
- ✓ Leverage points / influential obs.

Lack of fit (LOF)

$i$	$e_i$	$d_i$	$E(e_i) = 0$
1	$e_1$		
2	$e_2$		
...	...		
$i$	$e_i$		
...	...		
$n$	$e_n$		

$$d_i = \frac{e_i}{\sqrt{MSE}} = \frac{e_i}{\hat{\sigma}} \sim N(0, 1)$$


Standardized residuals then studentised residual and press statistics, press statistics and also we introduce leverage points or influential observations and these are the things will be discussed for diagnostic and then I already told you we will go for lack of fit. LOF related to regression, everything will be related to regression linear regression, multiple linear regressions.

So, let us see the 2 standardized residuals by this time you understand what is the meaning of standardized. So, you have  $i$  equal to 1 to  $n$  number of observations. So, you have  $n$  number of residuals  $e_1, e_2, e_i, e_n$ . So, what is the expected value of  $e_i$ , it will be 0.

So, by standardized residual we mean we will create suppose  $d_i$ , where  $D_i$  equal to  $e_i$  by MSE you know the concept of MSB also. So, this is nothing, but  $e_i$  by sigma square cap the sigma square cap is estimated by MSE. So, it will follow normal distribution, we and it variants for all standardized. So, what do you have done, you have essentially this is nothing we can write  $e_i$  minus expected value of  $e_i$  divided by variance of  $e_i$  square root that is the standardization. So, this one is 0 and like this. So, it will it should be all most normal 0, 1.

You read normal and what happened easier, if suppose this is unit normal distribution this is minus 3 and this one is plus 3 it is seen that most 99.7 percent observation will fall under this, but in most of the observation for minus 3 to plus 3. So, if any of the standardized residual is more than 3 in absolute value then what we can say that it is not normal it is out layers kind of things or other way I can say that it is not normal. So, this is, this is what is the studentized, the standardized a residual will tell you standardized residual will tell you about 2 things; obviously, we can assume that as it is part that much plus minus 3 sigma level away.

So, we can say also say that their out layered also, sometimes we can say in this manner also and what other way you can say that this is not normal. Now, second one is your studentized residuals. So, before studentized residuals you just know what is hat metrics. So, all of you know that  $\hat{y}$  equal to  $X\hat{\beta}$  and if I put the  $\hat{\beta}$  is nothing, but  $X^T X^{-1} X^T y$  then the resultant quantity is  $X^T X^{-1} X^T y$ . Now, we would say that  $H$  equal to  $X^T X^{-1} X^T$  and this is  $H y$ , this  $H$  is a very interesting matrix which is known as hat matrix and what it says, it maps the vector of observed values into the vector of fitted values.

So, that mean the hat matrix gives you information about a every observation some statistic about every observations that is a beautiful thing, that beta coefficients talks about the influence of the variable on the response, where in the from the hat matrix you

will know that influence of every individual observations. So, on the response that is the interesting one.

So, if  $y$  cap is  $H y$  then  $e$  then residual will be  $y$  minus  $y$  cap which is  $y$  into  $i$  minus  $H$ .

(Refer Slide Time: 06:58)

Handwritten mathematical derivations on a grid background:

$$e = y - \hat{y} = y(I - H)$$

$$Cov(e) = (I - H) Cov(y) = \sigma^2 (I - H)$$

$$V(e_i) = \sigma^2 (1 - h_{ii})$$

Studentized residual =  $r_i = \frac{e_i}{\sqrt{\frac{\sigma^2}{n-2} (1 - h_{ii})}}$

$$= \frac{e_i}{\sqrt{MSE (1 - h_{ii})}}$$

$$V(r_i) = 1$$

To the right of the equations is a small diagram showing a regression line with data points and a normal distribution curve.

Now, what we require, suppose if we require that covariance of  $e$  then this will be  $I$  minus  $H$  covariance of  $y$  which is nothing, but variance of sigma square  $I$  minus  $H$ . Now, then what is the variance of  $i$  th observation, the variance of  $i$  th observation will be sigma square  $1$  minus the  $i$  th diagonal elements of the  $H$  matrix which is  $H_{ii}$ .

Means a, we can write like this that  $H_{11}$ ,  $H_{22}$ ,  $H_{ii}$  like  $H_{nn}$  we have  $n$  observations. So,  $n$  cross  $n$  this will be and definitely this will be  $H_{1,2}$  like this other values will be there, other values will be there the diagonal elements  $H_{ii}$  is very important element, this is the influence of the observed observations on the regression line or in regression plane. So, then your variance for  $i$  th variance will be this. So,  $i$  th residual variance will be sigma square  $1$  minus by this then the studentized residual is if we say this is  $r_i$  which is  $e_i$  minus variance of  $e_i$  which is sigma square  $1$  minus  $H_{ii}$ .

Obviously you can write this one this divided by  $MSE (1 - H_{ii})$  and all those things you know and then what will be the variance of  $r_i$  variance of  $r_i$  will be  $1$  because it is this minus the min value will be  $0$  by this the variability this also  $1$  mean  $0$  mean this  $0$  variance is  $1$  well. This is known as studentized residual all right, it is similar to

standardized residual, but if some of the observations, in standardized residual what happened you have considered  $e_i$  by  $\sigma$  that  $m s e$  square root of a  $m s e$  you consider, but here you are multiplied this term, if everyone is contributing equally fantastic, but otherwise what happened if there is some of the observation contribute more  $H_i$  value will be more and then its contribution to the residual will be also more.

So, this is what is studentized otherwise in interpretation point of view they are similar to standardized residuals.

(Refer Slide Time: 10:46)

**Regression Model Diagnostics (Contd.)**

❖ **PRESS Residuals** (PRESS = Prediction Error Sum of Squares)

$$\text{PRESS} = \sum_{i=1}^n e_{(i)}^2 = \sum_{i=1}^n (y_i - \hat{y}_{(i)})^2 \quad e_{(i)} = \frac{e_i}{1 - h_{ii}} \quad \text{PRESS} = \sum_{i=1}^n \left( \frac{e_i}{1 - h_{ii}} \right)^2 \quad R_{\text{Prediction}}^2 = 1 - \frac{\text{PRESS}}{SS_T}$$

Data points having large  $h_{ii}$  are called high influence observations.

**Viscosity example**

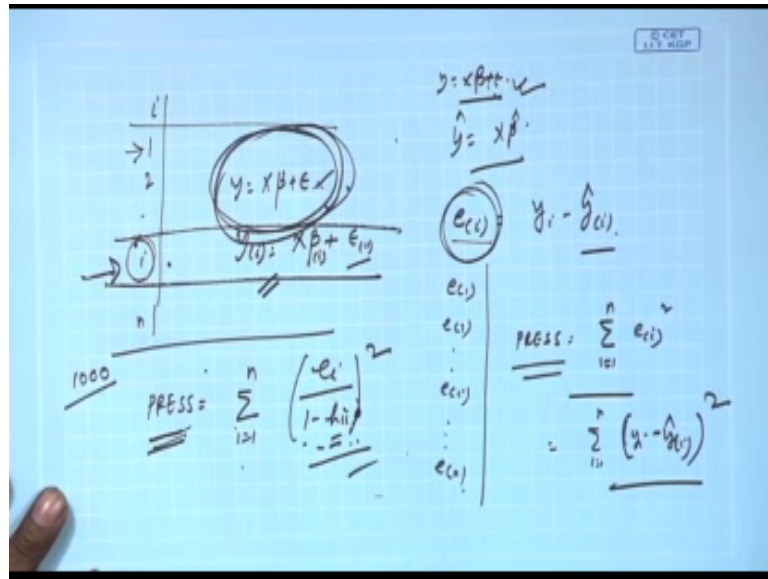
$$R_{\text{Prediction}}^2 = 1 - \frac{\text{PRESS}}{SS_T} = 1 - \frac{5207.7}{47,635.9} = 0.8907$$

Therefore, we could expect this model to "explain" about 89 percent of the variability in predicting new observations, as compared to the approximately 93 percent of the variability in the original data explained by the least squares fit. The overall predictive capability of the model based on this criterion seems very satisfactory.

IT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Now, we will come to another very important one is press statistics or press residual, press stands for prediction error sum of squares, What is the procedure here?

(Refer Slide Time: 11:02)



Procedure is you have  $i = 1, 2, \dots, n$  number of observations and you fitted a model  $y = X\beta + \epsilon$ , instead of these you do one thing you fit a model  $y_i = X\beta$  or if I say  $X\beta + \epsilon$  something like this plus epsilon or epsilon something like this. What I mean to say you remove the  $i$ th observation while fitting this model, then what happens if you remove the  $i$ th observation and then using the fitted model you predict the value of  $i$ th observations.

What you are doing? You are fitting the same regression model not considering all the data points  $i$ th observation you are removing or omitting then using these fitted model  $y = \hat{y}$  fitted equal to  $X\beta$  cap you are predicting the  $i$ th observations, then the error or what you are getting that  $e_i$  within bracket  $i$ , this one will be nothing, but that  $y_i$  minus  $y_i$  cap. You got this within bracket  $i$  we are using, we are saying that when you have fitted this model you have not considered the  $i$ th observations. Now, you can repeat the same thing, suppose you start with first observation find out this, what is this  $e_1$ , second observation find out this like this  $e_i$  like this  $e_n$ . So, you have total  $n$  number of residual values.

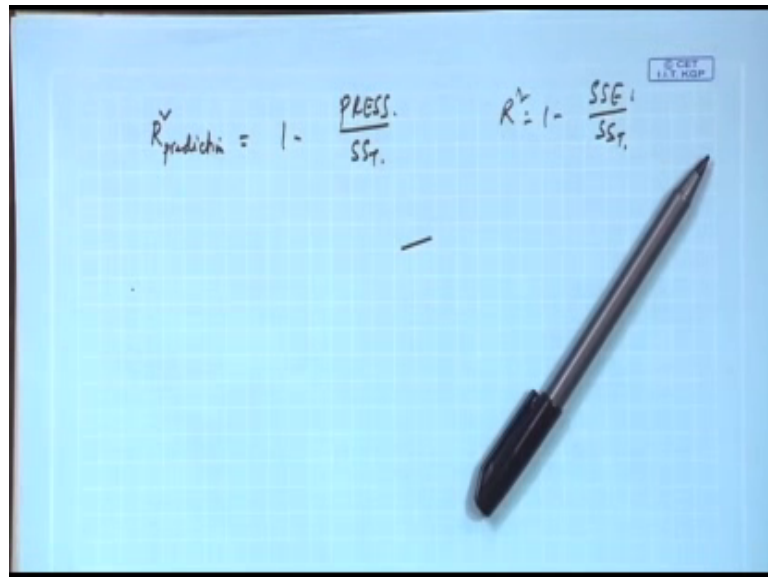
These are all basically predicted residuals basically and this value you are getting using this kind of approach and then you find out the press, what is the press will be? Pressure will be sum total of all those residual square  $i = 1$  to  $n$  this is nothing, but sum total of  $i = 1$  to  $n$ ,  $y_i$  minus  $y_i$  within this cap square. So, this is what is known as

press. So, by our listening this much you may be thinking that you have to go for n number of regression, means n times you have to run the regression first format the observation 1 then omit the observation 1 like this way you continuing up to n if n equal to 1000 that mean you have to run the program 1000 times ok.

But it is not so because even in from the one regression model also one time fitting reason you will get and that is interestingly that press is sum total  $\sum_{i=1}^n e_i^2$  by  $1 - \sum_{i=1}^n h_{ii}$  sorry. So, you do not require to go for n times during the regression omitting 1 observation at a time, you repeat the regression using all the values and then find out  $h_{ii}$  and also find out  $e_i$  then press  $\sum_{i=1}^n e_i^2$  by  $1 - \sum_{i=1}^n h_{ii}$  that is what is the what is the our e within bracket i.

So, that mean if you see the studentized residual that type that time,  $1 - \sum_{i=1}^n h_{ii}$  we have subtracted (Refer Time: 15:15) consider this place is this, then we will go for, suppose that here what happened every observation which is omitted is predicted using the regression model fitted through rest of the observations.

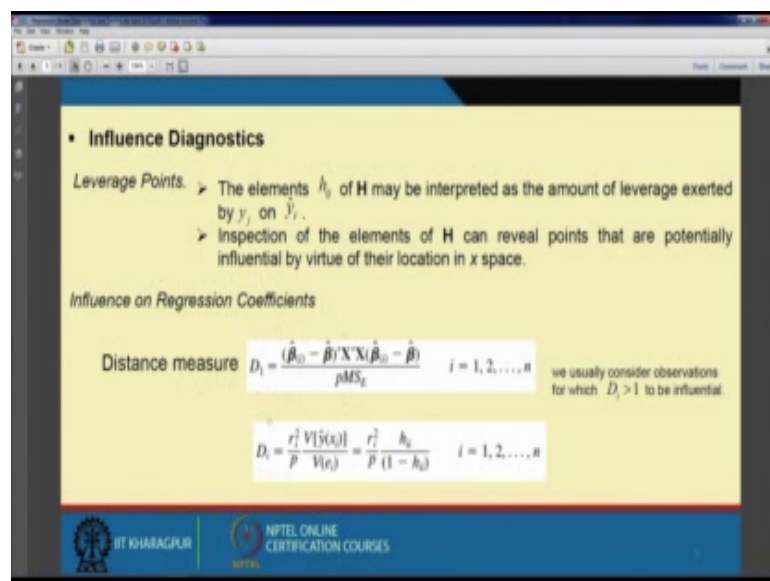
(Refer Slide Time: 15:38)



Then we can get also R square prediction, this will be 1 minus press by SS T usually what is our R square? R square is 1 minus SSE by SS T here instead of SS T we are using press. Now, let us see the example results.

So, the viscosity example what we have discussed I have discussed earlier. So, here when we have computed  $r^2$  using  $SSR$  square equal to  $1 - SSR/ SST$  we got that value is 0.93. Now, when we are using  $r^2$  square prediction using the press statistics this  $r^2$  square prediction value is coming 0.89. So, what is the conclusion here? We could expect the model can explain about 89 percent of the variability in predicting new observations as compared approximately 93 percent of the variability in the original data explained by the least square fit the overall predictive capability of the model based on this criteria seems very satisfactory.

(Refer Slide Time: 17:04)



Now, come to the influential diagonals influential diagnostics or influential observation here 1 is leverage point, what is the leverage point the elements  $H_{ii}$  this should be  $H_{ii}$  the elements  $H_{ij}$  of  $H$  may be interpreted as amount of leverage exerted by  $y_j$  on  $y_i$ . Inspection of the element  $H$  can rebuild points that are potentially influential by virtue of their location in the  $x$  space. So, as I told you the leverage points when if you see the diagonal elements of  $H$ ,  $H_{11}$  to  $H_{nn}$ .

So, these are the basically the influences are exerted by each of the observations. So, and also up diagonal elements are also there. So, there what happened the actually, if when we compare the  $y_i$  object, actual observation with the predicted observations. So, that sends you have to think of. So, another major, one major used for this influence



generation arm a quantification of influence that is known as that is known as D d beta measures distance measure D 1 or even D i.

So, it is basically D i. So, this is nothing, but what happened the same way that you eliminate the i th observations and feed the regression line and then the beta whatever the estimated value you are getting and when this estimated value and the full model. Means in full means the including all the observation you got beta value then using this 2 and the x design matrix this kind of this kind of distance is measured here it is D i, this distance is well if any of the D i values for is greater than 1 then that is influential. So, here also you may be thinking that you require to you require to go for n number of regression it is not because we have this a hat matrix.

From this hat matrix this D i can be computed like this R i studentized residual that is also you know p is the number of parameters to be estimated known and H i i is coming from hat matrix. So, if any D value is more than 1 that is influential.

(Refer Slide Time: 20:24)

**Testing for Lack of Fit**

$SS_E = SS_{PE} + SS_{LOF}$   $SS_{PE}$  is the sum of squares due to pure error and  $SS_{LOF}$  is the sum of squares due to lack of fit.

$y_i - \hat{y}_i = (y_i - \bar{y}_i) + (\bar{y}_i - \hat{y}_i)$

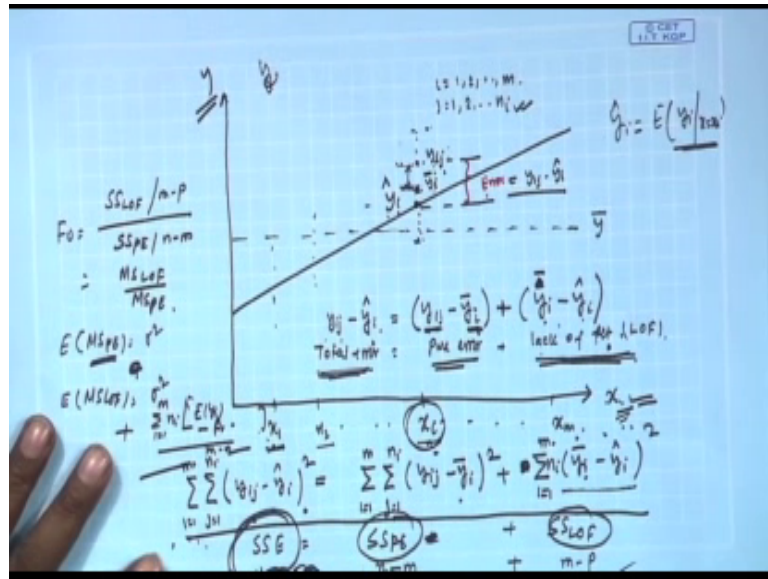
$\sum_{i=1}^n \sum_{j=1}^{n_i} (y_{ij} - \hat{y}_{ij})^2 = \sum_{i=1}^n \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^n n_i (\bar{y}_i - \hat{y}_i)^2$

$SS_{PE} = \sum_{i=1}^n \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$      $\sum_{i=1}^n (n_i - 1) = n - m$      $SS_{LOF} = \sum_{i=1}^n n_i (\bar{y}_i - \hat{y}_i)^2$

IFT KHARAGPUR    NPTEL ONLINE CERTIFICATION COURSES

Now, I will come to an another interesting concept which is known as lack of fit and how to test lack of fit.

(Refer Slide Time: 20:34)



First let us see this diagram, let us assume that this is x the regression and y is the dependent variable it. Here I am showing with one regression variable, but it can be for many regression variable. Suppose x 1 is the first observation x 2 second observation x i sorry first fixed value for x these are the m number of fixed values for the regression.

Now, let the i this 1 i th 1 is the x i 1 and here suppose you have conducted experiment keeping x at x i suppose n number of times or n i number of times. So, your observations will be y observation will be n i here, n i number of observations you will get and similarly here also n 2 here n 1 and like this and you have seen in one way analysis of variance also suppose keeping the power factor 165 number of observations experiment were done like this so in the same manner. Now, when you fit the regression equation for one factor experiment case you will get a regression line like this. So, any point suppose for so; that means, what is that this point on the regression line.

This point on regression line is y i cap. So, y i cap is the value of y when x equal to x i. So, this is basically I can say the expected value of y given x equal to x i kind of things. So, this is y i cap; now you have n 1, n i number of observations here. So, this n i number of observations if you compute the mean value that is y i bar maybe falling here, now you constant what is the what is the error here, total error here if i consider a general suppose you think of a particular observation y i j, i from 1 to m the mp regression fixed values and j is basically the replication at a particular fixed value of x.

So, now what have been the total error for a particular observation  $y_{ij}$  will be that  $y_{ij}$  minus  $y_i$ . So, these can be so  $y_{ij}$  minus  $y_i$  can be partitioned, like this  $y_{ij}$  minus  $y_i$ . That means, the of the mean value or average value of all the  $y$  when  $x$  equal to  $x_i$  which is coming here, plus we can write that  $y_i$  minus  $y_i$ . So, then each of the observation here is subtracted by their observations means. So, this is giving you pure error and then remaining one because total error is what, every observation minus the predicted value. So, the pure error, pure error is this basically every observation minus those observations mean value or average value then the rest is lack of fit rest is lack of fit.

So, actually what do you mean to say there is lack of fit if the mean of the observed  $y$  for  $x$  equal to  $x_i$  will not coincide with the predicted value, if it is not does not coincide with the predicted value. So, that is your lack of fit. So, now this is basically the deviation part you take square it, you square this side take the sum for across all  $j$   $j$  equal to 1 to  $n_i$  and across all  $i$   $i$  equal to 1 to  $m$  and you do algebraic manipulation you will be getting this equation.

Where you will be getting the sum square error which is sum square pure error plus sum square lack of fit, now see this slide. So, I told you this is this. So, sum square pure error is  $y_{ij}$  minus  $y_i$  square this one and you have  $n_i$  observation for at the  $x_i$  level. So, as there are  $m$  such levels. So, total number of observe that degrees of freedom available with the pure this error is  $n$  minus  $m$  and lack of fit you are getting from this formula that  $n_i e a$  is this. So, so; that means, what happened; now lack of fit is computed.

(Refer Slide Time: 26:05)

**Testing for Lack of Fit (Contd.)**

The test statistic for lack of fit is 
$$F_0 = \frac{SS_{LoF}(m-p)}{SS_{LoF}(n-m)} = \frac{MS_{LoF}}{MS_{PE}}$$

The expected value of  $MS_{PE}$  is  $\sigma^2$ , and the expected value of  $MS_{LoF}$  is:

$$E(MS_{LoF}) = \sigma^2 + \frac{\sum_{i=1}^m \left[ E(y_i) - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right]^2}{m-p}$$

If the true regression function is linear,  $E(y_i) = \beta_0 + \sum_{j=1}^p \beta_j x_{ij}$ , and  $E(MS_{LoF}) = \sigma^2$

If the true regression function is not linear,  $\nexists E(y_i) = \beta_0 + \sum_{j=1}^p \beta_j x_{ij}$ ,  $E(MS_{LoF}) > \sigma^2$

If the true regression function is linear,  $F_0$  follows  $F_{m-p, n-m}$  distribution

**The regression function is not linear:  $F_0 > F_{\alpha, m-p, n-m}$**

IT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, let us compute a statistics to say that whether lack of fit is significant or not. So, these statistics is  $F_0$  SS lack of fit, what is SS lack of fit? This one by its degree of freedom by its degree of freedom means what happened how many x ray levels are there. So,  $i$  equal to 1 to  $m$  every level you are computing and you have also 1 degree lost that is why what is happening for  $p$  number of parameters you have lost to that  $p$  value degrees of freedom. So,  $m$  minus  $p$  by SS pure error by its degrees of freedom because pure error degrees of freedom the remaining degrees of freedom from SSE so that is  $n$  minus  $m$ . So, you are getting  $m$  s lack of fit by  $m$  s pure error.

Now, expected value of  $m$  s pure error is  $\sigma^2$  and the expected value of  $m$  s lack of fit is  $\sigma^2$  plus this quantity. So, what is  $\beta_0$  minus this or  $\beta_0$  plus this? This is nothing, but your the predicted value. So, an expected value of  $y_i$  that also that is the value which is basically falling on the regression line if they coincide this value will become 0.

So, that will lack of fit and this will be same so, but if they do not consider this quantity becomes more than 0. So, expected value of  $m$  s lack of fit will be more than  $\sigma^2$ , if the true regression function is linear then what happened expected value of  $y_i$  will become this  $\beta_0$  plus this in that case this becomes  $\beta_0$  plus this means this quantity becomes 0 and as a result I told you expected result lack of result will become  $\sigma^2$ .

If the true regression function is not linear this will not become 0, this quantity not become 0 because expected value  $y_i$  is not this in that case this quantity become positive and there is more than this, their lack of expected or a lack of fit is more than sigma square. If the true regression line is linear now the first case then this quantity  $f_0$  follows  $f_{m-1}$  distribution and the and what happen the regression function is not linear then your  $f_0$  the computed value will be greater than the theoretical  $f$  value.

So, so thank you very much for interesting hearing and I hope that you got enough inputs for regression and also enough inputs for anova, apart from the basic statistics part what we have given you earlier. So, I can tell you that the statistical poor statistics part for understanding the design understanding and analyzing that the desire experimental data. So, that is covered now and one important thing is to be covered further from the theoretical point of view is the sample size.

How to calculate the adequate sample size under different situations? So, that will also be discussed in. In fact, when we go for other lectures where we will, we bring those all those concept again all those sample size part we have discussed, but again we will bring all those things together ok.

Thank you very much.