

**INDIAN INSTITUTE  
OF  
TECHNOLOGY  
KHARAGPUR**

**NPTEL  
National Programme  
On  
Technology Enhanced learning**

**Applied Multivariate statistical Modeling**

**Prof. J. Maiti  
Department of Industrial instrumentation and management  
IIT Kharagpur**

**Lecture – 23**

**Topic**

**MLR – Model Adequacy Tests**

Good morning today we will discuss multiple linear regression Model Adequacy Tests.

(Refer Slide Time: 21:60)

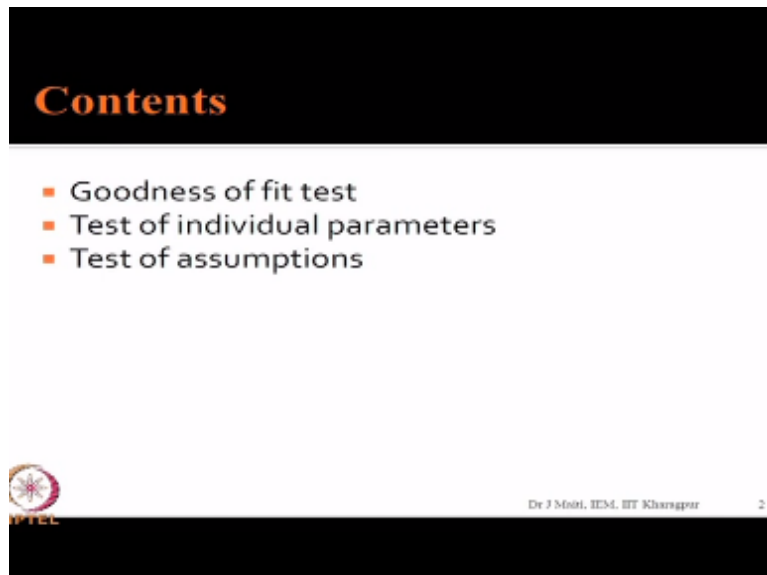
MLR - Model Adequacy Tests.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \epsilon$$
$$= X\beta + \epsilon$$
$$\hat{\beta} = (X^T X)^{-1} X^T y \quad E(\hat{\beta}) = \beta, \text{Cov}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$$
$$\hat{\epsilon} = y - \hat{y} = (I - H)y \quad E(\hat{\epsilon}) = 0, \text{Cov}(\hat{\epsilon}) = \sigma^2 (I - H)$$
$$H = X (X^T X)^{-1} X^T$$

So, what we have learned so far about multiple linear regression, we have seen that the general equation is like this  $y = \beta_0 + \beta_1 X_1 + \beta_2 x_2 + \dots + \beta_p x_p + \epsilon$  and which we can write in terms of  $x \beta + \epsilon$ . We have already estimated the value of  $\hat{\beta} = (x^T x)^{-1} x^T y$  and we also estimated the expected value of  $\hat{\beta}$  which is  $\beta$  and covariance of  $\hat{\beta}$  which is actually  $\sigma^2 X^T$  transpose  $X^{-1}$ . We also have seen  $\hat{\epsilon}$ , which is basically  $y - y$  predicted and this one we have seen in terms of  $(I - H) y$ .


We also have seen expected value of  $\hat{\epsilon}$  is zero and covariance of  $\hat{\epsilon}$  that is  $\sigma^2 (I - H)$ , where  $H$  is the hat matrix, which is  $X$  into  $x^T X^{-1} x^T$  up to this much we have covered and we want to test today that the model whatever we have we will develop under this equation or this equation. So, whether the model is adequate enough to explain the data given.

(Refer Slide Time: 03:00)



**Contents**

- Goodness of fit test
- Test of individual parameters
- Test of assumptions

  
IIT K

Dr J Mohd, IEM, IIT Kharagpur 2

So, in this slide today's contents are goodness of fit test, Test of individual regression parameters and Test of assumptions, these three things we will consider.

(Refer Slide Time: 03:17)

Goodness of fit test

$$R^2 = \frac{\text{Explained variance (of } y)}{\text{Total variance (of } y)}$$

Coefficient of determination

$$SST = SSR + SSE$$

$y \rightarrow \delta y^2$   
 $SST \rightarrow (n-1) \delta y^2$   
 $n = \text{no of obs.}$

Sum square regression  
Sum square error

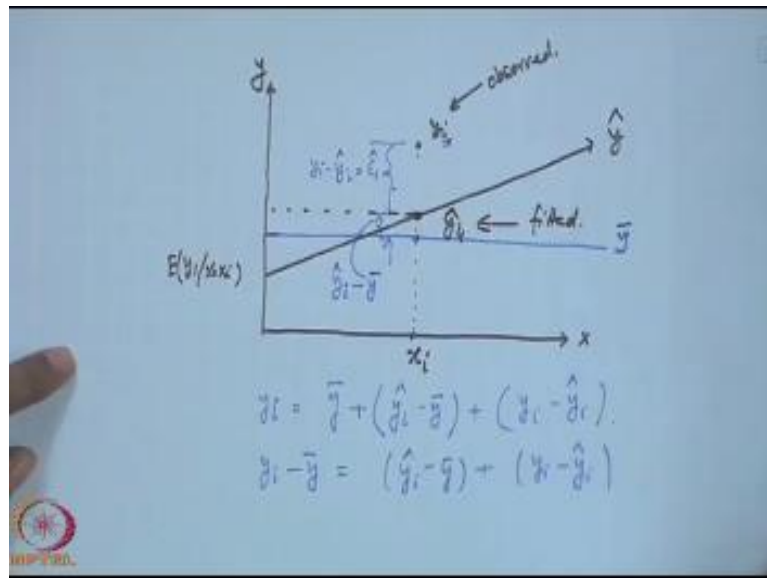
Let us see the first the goodness of fit test in goodness of fit test we will use one statistic called  $R^2$  square, which is known as coefficient of determination. This will give us a ratio of explained variance definitely variance of  $y$  by unexplained by total variance unexplained + explained by total variance of  $y$  Okay. So, that mean what you mean to say that  $y$  is having the  $y$  values they vary from one to another and there is variance of  $y$ , which is  $s^2$  stands for the variance and standard deviation and related to  $y$  to the square that is variance.

Then the total variability in  $y$  if we define like this SST sum square total that will be your  $(n - 1) s^2$ , where  $n$  is number of observations. So, this total variability will be decomposed into two parts model explained variability + unexplained variability. So, what is your will be doing now? That SST can be decomposed into SSR that R stands for regression. That is sum square regression plus, what is unexplained that is sum square, sum squares error, this one is the sum square, total sum squares total.

So, all related to the dependant variable we are talking about we are talking about SST sum square total for  $y$ . Then this variability is explained by regression equation the explained portion is coming

under SSR and unexplained portion is coming under SSE. It is similar to in ANOVA, what we have seen the same thing Okay.

(Refer Slide Time: 06:20)



So, in order to establish this relationship let us explain in with one variable, let it be that y versus X regression line let it be like this. So, this one is your  $\hat{y}$  and let there is one observation of  $X_i$  that is  $i^{\text{th}}$  observation on x. Then if you want to predict what will be the expected value of  $y_i$  expected value of  $y_i$  given X equal to  $X_i$ . This is nothing but this value, So and this one is the predicted value of  $y_i$ .

This is  $y_i$  predicted let the original value of that  $y_i$  is somewhere here this is the observed value. So, observed this is fitted and let us also define the average, let the average of y is here Okay. Actually it can be little up. Anyhow, we will just assume that it is here then this value is  $\bar{y}$ , this value is  $\hat{y}_i$  this value is  $y_i$ . So, what is this value then?  $y_i - \hat{y}_i$  this value is nothing but error value correct.  $\bar{x}$

Then what is this value this one, So if you have no information about X that mean you have only information about y, then the best fit value is y. Suppose, you want to say what is the next value

you will consider  $\hat{y}$  as  $X$  is available as a result what is happening you are able to find out some fitted or predicted value  $\hat{y}_i$ . Definitely, it is which what I can say that closer to  $y_i$ , which is the observed one.

So, this is the amount for  $y_i$ , which is explained by the regression line if there is no regression line that mean no  $X$  information is with you. You will go by  $\bar{y}$ . whether this is the if you have only the  $y$  information this is the best fit. So, then this amount what is this  $y_i - \bar{y}$ ? This is what is explained by the regression line then  $y_i - \hat{y}_i$ , which is error cannot be explained by the regression line mm.

When you are talking about one variable case  $X = 1$   $y = 1$  one always one in multiple regression. So, then if I want to write this, why I cannot write like this  $\bar{y} + ?$  When this  $1 +$  this  $1 +$  this one. So,  $y_i - \hat{y}_i + y_i - \hat{y}_i$  you can write like this  $\bar{y} - \bar{y}$  cancelled out  $\hat{y}_i$  and  $\hat{y}_i$  will be cancelled out  $y_i$  will be there. So, if you recall the ANOVA part, then you can very well understand that what we are going to do  $\hat{y}_i - \bar{y} + y_i - \hat{y}_i$  you require to square the terms first. Then take summation over all observations you will be getting this  $(\hat{y} - \bar{y}) + (y - \hat{y})$  definitely sum of this  $(\hat{y} - \bar{y})^2 + (y - \hat{y})^2$  some co that co cross product that portion will become 0.

(Refer Slide Time: 11:19)

Handwritten mathematical derivation on a whiteboard:

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$SST = SSR + SSE$$

$$\frac{SST}{n-1} = \frac{SSR}{p} + \frac{SSE}{n-p-1}$$

$$R^2 = \frac{SSR}{SST} >> 0.90$$

in explaining applied + experience

$< 0.90$  Social/Political Management

NPTEL logo is visible in the bottom left corner of the whiteboard image.

So, ultimately it is similar to your this one that ANOVA the total variation of the y variable, but you are able to explain 52 percent. So, that means some vari part. So, what will happen? Ultimately, you will get something like this. This  $\sum_{i=1}^n (Y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$  What is this One Y, Y1, Y2 like yn 12 n, this square plus i equal to one ton y i minus y i cap whole square. So, what is this one? This is our SST, because i fi 1, I have y with y one to like y ni equal to 1 to n. I am asking you, what will be the variance component here you will say that  $1/n - 1$  into  $y_i - \bar{y}^2$ .

That is sum  $\sum_{i=1}^n (y_i - \bar{y})^2$  this one is nothing but this quantity  $y_i - \bar{y}^2$ . This is that is S S T / y n - 1, so this is S S T. So, similarly this one will be S S R, because this is the amount, which is more, which is basically explained by the regression. Then this is another amount, which is quantitative, which is not explained by the regression. So, similarly, what will be your degrees of freedom here n - 1? What will be the degrees of freedom here?

How many parameters you are estimating p + 1 p + 1. So, p plus 1 - 1 that will be p, then what will be the degrees of freedom here n - s p minus one getting? So, for total y the SST part sum square total as you haven observations n - 1 is the degrees of freedom, you are estimating p + one parameter including inter shift. So, that is why p plus one minus one this is the degrees of freedom and SSE this n - p - 1 this is the degrees of freedom. So then what is  $r^2$  R<sup>2</sup>

Is explained variance what is explained variance SSR divided by unexplained total variance that is S S T. So, this  $r^2$  value should be greater than equal to 0.90 for engineering application or if you do experiments laboratory experiments, but if you collect data from field maybe less than 0.90 is also acceptable. Because, you are collecting data from field, which is more what I can more variable in nature more dynamic more volatile in nature.

So, there what will happen ultimately it will be very difficult to get r square equal to 0.9 you may get less, but you can then you can go for suppose for social sciences, social administrative management sciences. It can be less than 0.90 but it all depends on the accuracy required Okay. Now, in  $r^2$  there is one problem the problem is if I write down R<sup>2</sup> little bit different way. That this is R / S S T, which is S S T - S S E / S S T, because this is quite obvious. So, 1 - SSE / SST is correct.

What is your SSE? In last class I think, yes last class only I told you that  $S E^2 = SSE / n - p - 1$ . We have shown this so instead of SSE i will write  $S E^2 n - p - 1 / S S T$ . I can write  $s y^2$  into  $n - 1$  you sees  $s y^2 = SST / n - 1$ . So,  $SST = s s y^2$  into  $n - 1$ ,

(Refer Slide Time: 17:00)

$$R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST}$$

$$= 1 - \frac{SSE}{SST}$$

Now  $n = \dots$

$$R^2 = 1 - \frac{\sigma^2 (n-p)}{\sigma^2 (n-1)} = 1 - \frac{n-p}{n-1}$$

Cond:  $n \geq p+1$

$R^2$  is sensitive to  $n$  and  $p$ .

$n > 5p$	$< 30$
$n \geq 10p$	
$n = 10p$	

So you can write in this manner. Now, condition is like this that  $n = p + 1$  what will be the value of  $R^2$   $1 - S E^2$  into  $n - n$ . Because  $n = p + 1$   $p + 1 / s y^2 n - 1$  this is 0. So, it is one this is the limiting condition what will happen if you number of observation is equal to number of parameter estimated your  $R^2$  will become one saturated. So, that means  $R^2$  is sensitive to  $n$  and  $p$  and you require  $n$  much greater than  $p$  to rely on our square.

So, in multiple regression what it is said, that it is expected that. That  $n$  is will be like this less than equal to  $5 p$  not less than it should be what mean to say it should atleast be greater than equal to  $5 p$  desirable is  $10 p$ . If it is more than  $10 p$  no problem, but again very large  $n$  also not desirable in many cases suppose if you go for  $k_i^2$  test sometimes in later part of the lectures, we will see that  $k_i^2$  will be using many some places.



Then there if  $k_i^2$  is also sensitive to sample size more sample size also not good. So, for regression purpose what I mean to say here that if your sample size is atleast five times of  $p$  that is the starting point you have to have this amount of data and  $10p$  is the desirable one and in no case. Ultimately, it should be less than 30, suppose there is only one  $p$ , so that may you may think that it will be 5 into one 5, that is not the case.

So, I can say that the minimum value is 30 minimum  $n$  equal to minimum of 30 and  $5p$  whichever is minimum maximum of this you write should not be minimum maximum of this 30 and  $5p$  okay. So, in order to avoid this that dependency on  $n$  and  $p$ , if we modify  $R^2$  in this manner, what will happen?

(Refer Slide Time: 20:34)

The image shows handwritten mathematical derivations on a blue background. The equations are as follows:

$$R^2 = 1 - \frac{S_e^2 (n-p)}{S_y^2 (n-1)}$$

$$R_a^2 = 1 - \frac{SSE / n-p-1}{SST / n-1}$$

$$\text{Adjusted } R^2 = 1 - \frac{S_e^2 (n-p) / n-p-1}{S_y^2 (n-1) / n-1}$$

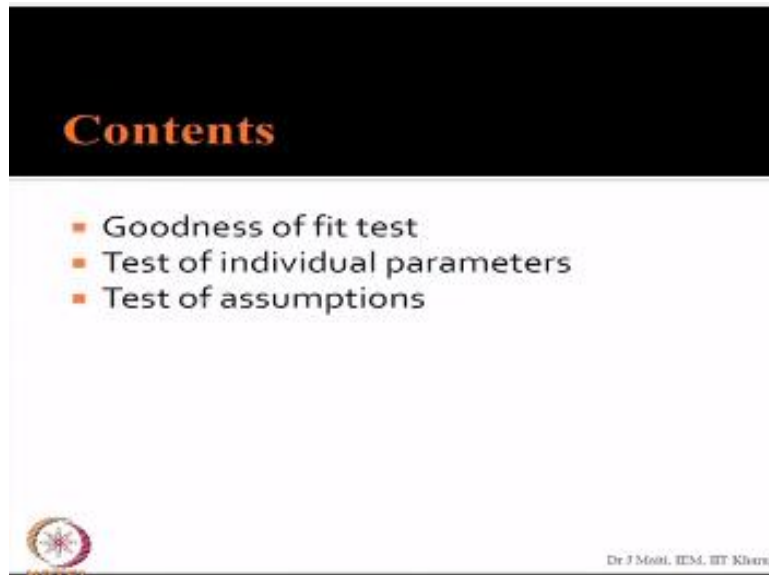
$$= 1 - \frac{S_e^2}{S_y^2}$$

Arrows point from the text 'Adjusted R^2' and '= R\_a^2' to their respective equations. A small logo is visible in the bottom left corner of the slide.

We have written this  $R^2 = 1 - S_e^2 / S_y^2$  into  $n - p - 1 / s y^2 n - 1$ . Let us create one more statistics  $R^2$ , which is known as adjusted  $R^2$  in this manner this /  $SSE / n - p - 1$  and  $SST / n - 1$ . That means the sum square error divided by its degrees of freedom sum square total divided by its degrees of freedom. Then what will happen? You can write  $S E^2 - p - 1 / n - p - 1 / s y^2 n - 1 / n - 1$ . So,  $n - p - 1 / n - p - 1$  will cancelled this also will be cancelled. So, it will be  $1 - S E^2 / s y^2$  it is irrespective of sample size it will give you a value and this is  $R^2$ .


So, adjusted  $R^2$ , which is known as  $R^2$  this is parsimonious. This is known as parsimonious fit measure in case of multiple regressions Okay. So, with this we want to see that the example we are discussing the same example and what are the values?

(Refer Slide Time: 22: 28)



**Contents**

- Goodness of fit test
- Test of individual parameters
- Test of assumptions

 Dr. J. Mohi. IEM, IIT Kharagpur

(Refer Slide Time: 22:31)

## An example

Sl. No.	Months	Profit in Rs million	Sales volume in 1000	Absentees m in %	Machine breakdown in hours	M-Ratio
1	April	10	100	9	62	1
2	May	12	110	8	58	1.3
3	June	11	105	7	64	1.2
4	July	9	94	14	60	0.8
5	Aug	9	95	12	63	0.8
6	Sep	10	99	10	57	0.9
7	Oct	11	104	7	55	1
8	Nov	12	108	4	56	1.2
9	Dec	11	105	6	59	1.1
10	Jan	10	98	5	61	1.0
11	Feb	11	103	7	57	1.2
12	March	12	110	6	60	1.2




This is the data, we will be using and here sales volume is our y absente is m and machine breakdown. These two is our independent variables for explanation of multiple regressions.

(Refer Slide Time: 22:49)

## Regression parameter estimates

$Y = 130.22 - 1.24X_1 - 0.30X_2 + e$

Observed	Fitted	Residuals	$C = (XTX)^{-1}$		
100	100.44	-0.44	40.9	0.114	-0.703
110	102.88	7.12	0.11	0.012	-0.004
105	102.32	2.68	-0.7	-0.004	0.012
94	94.82	-0.82			
95	96.41	-1.41			
99	100.69	-1.69			
104	105.02	-1.02			
108	108.45	-0.45			
105	105.07	-0.07			
98	105.71	-7.71			
103	104.42	-1.42			
110	104.77	5.23			
			SSE	$Y'(I-H)Y$	155
			$se^2$	$SSE/(n-p-1)$	17.22

 Dr. J. Maiti, IEM, IIT Kharagpur 4

Then last class we have fitted this one, we found that  $y$  equal to  $130.22 - 1.24 X_1 - 0.30 X_2 + \text{error}$ . This is my regression line using this regression line you have found out the fitted values getting. Basically, how we got this fitted values?

(Refer Slide Time: 23:17)

The image shows a person's hands writing on a whiteboard. The equation  $\hat{y} = X\hat{\beta}$  is written at the top. Below it, a horizontal line is drawn. Underneath the line, the letter 'X' is written, followed by a vertical line with a '1' at the top and a '1' at the bottom, representing a column of ones. In the bottom left corner, there is a circular logo with the text 'NPTEL' below it. In the top right corner, there is a small rectangular box containing the text '2021 117.60M'.

We are saying  $y = X\hat{\beta}$ , what are your X values X values one, like 12 values then X 1 values are there X 1 values are X 1 values.

(Refer Slide Time: 23:35)

## An example

Sl. No.	Months	Profit in Rs million	Sales volume in 1000	Absentees in %	Machine breakdown in hours	M-Ratio
1	April	10	100	9	62	1
2	May	12	110	8	58	1.3
3	June	11	105	7	64	1.2
4	July	9	94	14	60	0.8
5	Aug	9	95	12	63	0.8
6	Sep	10	99	10	57	0.9
7	Oct	11	104	7	55	1
8	Nov	12	108	4	56	1.2
9	Dec	11	105	6	59	1.1
10	Jan	10	98	5	61	1.0
11	Feb	11	103	7	57	1.2
12	March	12	110	6	60	1.2



(Refer Slide Time: 23: 43)

$$\hat{y} = X\hat{\beta}$$
$$= \begin{bmatrix} 1 & 9 & 62 \\ 1 & 8 & 58 \\ \vdots & \vdots & \vdots \\ 1 & 6 & 60 \end{bmatrix} \begin{bmatrix} 30.22 \\ -1.24 \\ -0.30 \\ \vdots \\ \vdots \end{bmatrix} = \begin{bmatrix} 100.44 \\ 102.88 \\ \vdots \\ 104.42 \\ 107.77 \\ \vdots \end{bmatrix}$$

The matrix  $X$  is labeled as  $12 \times 5$ . The vector  $\hat{\beta}$  is labeled as  $5 \times 1$ . The resulting vector  $\hat{y}$  is labeled as  $12 \times 1$ . There is a small box in the top right corner of the slide that says "Lecture 11: Regression".

Are absenteeism and machine breakdown show absenteeism like 9, 8 like this up to 6 and breakdown hours 62, 58 like this 60. This is your X you have computed beta what is the beta cap values one 30.22 then -1.24 - 0.30. So, this one is 12 cross 1, 2, 3 then this is three cross one. So, what you will get you will get a 12 cross 1 vector of  $\hat{y}$ . This is what is fitted values you say in the fitted values are 100.44, 102.88. So, like this you will be getting up to at the end 104.42, 107.77 okay.

(Refer Slide Time: 25:10)

## Regression parameter estimates

$Y = 130.22 - 1.24X_1 - 0.30X_2 + e$

Observed	Fitted	Residuals	$C = (X^T X)^{-1}$		
100	100.44	-0.44	40.9	0.114	-0.703
110	102.88	7.12	0.11	0.012	-0.004
105	102.32	2.68	-0.7	-0.004	0.012
94	94.82	-0.82			
95	96.41	-1.41			
99	100.69	-1.69			
104	105.02	-1.02			
108	108.45	-0.45			
105	105.07	-0.07			
98	105.71	-7.71			
103	104.42	-1.42			
110	104.77	5.23			
			SSE	$Y'(I-H)Y$	155
			$se^2$	$SSE/(n-p-1)$	17.22

Dr. J. Maiti, IEM, IIT Kharagpur

So, what we require now? That fitted values are there are.



(Refer Slide Time: 25:14)

The image shows handwritten mathematical work on a whiteboard. At the top, the equation  $\hat{y} = XB$  is written. Below it, a matrix equation is shown: 
$$\begin{bmatrix} 1 & 9 & 62 \\ 1 & 8 & 58 \\ \vdots & \vdots & \vdots \\ 1 & 6 & 60 \end{bmatrix} \begin{bmatrix} B_0 \\ B_1 \\ B_2 \end{bmatrix} = \begin{bmatrix} 100.44 \\ 102.87 \\ \vdots \\ 104.44 \\ 104.77 \end{bmatrix}$$
 The matrix is labeled  $12 \times 3$  and the vector is labeled  $12 \times 1$ . Below this, the terms  $SSR, SSE, SST$  are listed. The formula  $SST = (n-1) s_y^2$  is written, with an arrow pointing to a calculation of  $\hat{e} = y - \hat{y}$ . This leads to a matrix subtraction: 
$$\begin{bmatrix} 100 \\ 110 \\ \vdots \\ 103 \\ 110 \end{bmatrix} - \begin{bmatrix} 100.44 \\ 102.87 \\ \vdots \\ 104.44 \\ 104.77 \end{bmatrix} = \begin{bmatrix} -0.44 \\ 7.13 \\ \vdots \\ -1.14 \\ 6.23 \end{bmatrix}$$
 The resulting vector is labeled  $\hat{e}$ .

y is also available you require to calculate SSR, SSE, SST that is our work. Now, SST you will be you will be calculating very well like this using this  $n - 1$  into  $s_y^2$  that is the formula. Now, if I want to calculate SSE you are required to find out what is the residuals this residuals will be this is  $y - \hat{y}$  so if you will be getting already y values are there y values are 100, 110. So, in the same manner 12 values 103 and 110, then fitted values you got 100.44, 102.88 in the same manner you will be getting here 100.4 and 204.77. That means finally, the error values will be minus 0.447 point one two same manner you come -1.42, 5.23. So, you have your error part also from here y value you calculated this one from this you can calculate  $SE^2$  getting me, how to calculate?

(Refer Slide Time: 27:08)

The image shows handwritten mathematical work on a light blue background. On the left, there is a vertical vector of ones,  $\hat{\epsilon}_{n \times 1}$ . To its right, the Sum of Squared Errors (SSE) is calculated as  $SSE = \hat{\epsilon}^T \hat{\epsilon}$ . Below this, the Mean Squared Error (MSE) is calculated as  $MS = \frac{\hat{\epsilon}^T \hat{\epsilon}}{n-p-1} = \frac{153.68}{12-9}$ . This is further simplified to  $= \frac{153.68}{9} \approx 17.05$ . At the bottom, the text 'SSR · SST =' is written.

So, our  $E^{\wedge}$  is  $n$  cross one if I make a one. How to calculate. SSE, which is  $\hat{\epsilon}_{n \times 1}^T \hat{\epsilon}^T S E^2$  is / degree of freedom  $n - p - 1$ . So, this value this value will be  $\epsilon$  t this value will be around 155 some think  $153.68 / n$  is  $12 - p - 1$  is 3. So,  $153.68$  by  $9$ , so it will be  $17.05$ , let it be similar this value will be there okay, so then SSR[FL]. Now, we have to find out SST, SST value we have to check SSEI calculated.

(Refer Slide Time: 28:31)

The image shows handwritten mathematical work on a blue background. On the left, there is a note:  $E(\hat{\beta}) = 0$ . To the right, the calculation for SSE is shown:  $SSE = \hat{E}^T \hat{E}$ , followed by  $s_0^2 = \frac{\hat{E}^T \hat{E}}{n-p-1} = \frac{153.68}{12-3}$ , which simplifies to  $= \frac{153.68}{9} \approx 17.05$ . Below this, the formula for SST is given as  $SST = \frac{1}{n-1} \sum (y_i - \bar{y})^2$ .

SST value will be 324.92, how do I calculate same way your  $y$  is  $y$  is  $n$  cross  $1$ . So,  $1 / n - 1$  if you do  $y^T y$  you will be getting I think  $y^T y - \bar{y}^T y - \bar{y}^T y$  this one you have to do, but here we have not done this manner, but we assume that this will be 0, but actually when you calculate that expected value  $\epsilon$  of that may not be 0.

(Refer Slide Time: 29:26)

SSR, SSE, SST.

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2$$

$$\hat{\epsilon} = y - \hat{y}$$

$$= \begin{bmatrix} 100 \\ 110 \\ 103 \\ 110 \end{bmatrix} - \begin{bmatrix} 100.44 \\ 102.88 \\ 106.02 \\ 104.55 \end{bmatrix} = \begin{bmatrix} -0.44 \\ 7.12 \\ -3.02 \\ 5.45 \end{bmatrix}$$

$$\sum \epsilon = \frac{\sum \epsilon_i}{n}$$

Handwritten notes and matrices are visible on the whiteboard, including a 3x1 matrix with values -1.24, -0.50, and 102.87, and a 12x3 matrix with values 1, 6, 60 in the first row.

Because, your data is this one data is this, so what will be the expected value that is mean value that sum of this  $\epsilon / n$  it is theoretically it should be 0, but it may not be 0. If you calculate like this I think this value will be 155.

(Refer Slide Time: 29: 56)

The image shows handwritten mathematical derivations on a whiteboard. On the left, there is a note:  $E(\hat{\beta}) = \beta$ . To the right, the Sum of Squared Errors (SSE) is calculated as follows:  $SSE = \hat{e}^T \hat{e}$ , then  $\hat{\sigma}^2 = \frac{\hat{e}^T \hat{e}}{n-p-1} = \frac{153.68}{12-3}$ , which simplifies to  $= \frac{153.68}{9} \approx 17.05$ . Below this, the Sum of Squares Total (SST) is given as  $SSR \cdot SST = \frac{1}{n-1} \sum (y_i - \bar{y})^2$ .

So, 0 that mean value is 0 then this will become 155 by calculated, but there is definitely that slight difference will be there. Because, of small data set other way it should follow even though that expected value of epsilon cap is not 0, but it should not be high value Okay. So, using this SST you will be you are able to calculate SST. SST is this one I was talking about  $\epsilon^2 = \text{by } n-1 =$  to this correct so then SST is known. So, SST is  $y - \bar{y} \text{ t } - \bar{y}$  this value we have found out that 324.92. So, my SSR is SST - SSE now, so 324.92 - 153.68, which is 171.23 Okay.

(Refer Slide Time: 31:30)

$$R^2 = \frac{SSR}{SST} = \frac{171.23}{324.92} = 0.527$$

$$R_a^2 = 1 - \frac{SSE/(n-1)}{SST/(n-1)} = 1 - \frac{(153.68)/9}{324.92/11} = 0.422$$

$n = 12$   
 $R^2 \geq 0.90$   
 $R_a^2 \geq 0.90$   
 $R_a^2 < R^2$

So SSR is known, SST is known. So,  $R^2$  value will be  $SSR / SST$ , which is  $171.23 / 324.92$ . It is 52, 0.527 less very less Okay. Then what is your  $R_a^2$  Value? This is  $1 - SSE$  by  $n-p - 1$  by  $SST$  by  $n - 1$ . So, that is  $1 - SSE$  is how much  $153.68$  divided by  $9$  that is the degree of freedom by  $SST$  is  $324.92 / 11$   $n - 1$  and this quantity is  $0.422$ . So, from  $0.52$  from  $0.527$  to  $0.422$ , it is a I think reasonable change this is, because your  $n$  is only  $12$ .

This is because of this value change is there see  $R_a^2$  drastically change to  $0.42$  from  $0.527$ . Now, what can you conclude from this straight away you can say that as  $R^2$  is  $0.527$ . This model is not fit what I said  $R^2$  should be greater than equal to  $0.90$ . Similarly,  $R_a^2$  should be greater than equal to  $0.90$ . So, both the things definitely whenever you will get  $R$  square greater than  $R^2$  greater than  $R$  square.

That should not that you get, or not you will not get it may be  $R^2$  equal to less than equal to  $R_a^2$  less than equal to  $R^2$  and that is what is happening here, okay. Whatever in this particular example it is not fit but some example it will you will fit that you get a good fit. So here using  $R^2$  you will basically or this  $SST$ ,  $SSR$  you are what you are creating here you are creating a ANOVA table.


(Refer Slide Time: 34:29)

## Goodness of fit test

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	171.23	85.62	5.01	0.034
Residual Error	9	153.68	17.08		
Total	11	324.92			

**R-Sq = 52.7% R-Sq(adj) = 42.2%**



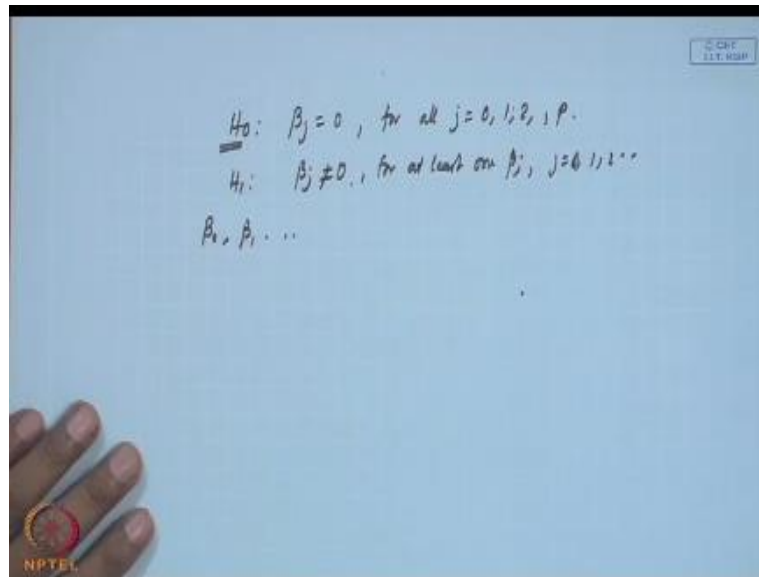
Dr. J. Maiti, IEM, IIT Kharagpur

5

You see the ANOVA table, analysis of variance source regression residual error total degrees of freedom there are three parameters to be estimated, so 3-1, 2 residual errors is 9 because total degrees of freedom is 11, 11- 2 that will be 9 and the sum squares 171, 153 is so we can create MS, what is MS? SSR by degree of freedom and your MSE will be SSE by degrees of freedom and the quantity MSR by MSE will follow F distribution with the respective numerator and denominator degree of freedom it is similar to ANOVA case, okay.

Here what you are finding out your  $R^2$  is saying that your model is not good, but your f statistics is saying that the p value is how much here. 0.034. It is it is less than 0.5 then what is the difference between this, whichever is this what we are testing here okay. I think you can understand this thing also in reference to ANOVA.

(Refer Slide Time: 36:09)



So when you make this type of ANOVA table our null hypothesis is that  $\beta_j = 0$  for all  $j$ . That mean there is no regression coefficient which is significant and our alternative hypothesis is that there is atleast 1  $\beta_j$  for at least 1  $\beta_j \neq 0$ , 1 like this that mean if  $\beta_0, \beta_1$  or anyone of the  $\beta$  parameter are significantly contributing in explaining the variability of  $y$  then this  $H_0$  will be rejected,  $R^2$  is a consolidated measure.

That how much variability of why you are able to explain. If it is one 52% you are not happy you want 90% or more, but 52 even if 52% is explained but that means some parameters some of the independent variables they have importance they are explaining that 52 % it is not 0%.




(Refer Slide Time: 37:39)

## Goodness of fit test

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	171.23	85.62	5.01	0.034
Residual Error	9	153.68	17.08		
Total	11	324.92			

**R-Sq = 52.7% R-Sq(adj) = 42.2%**

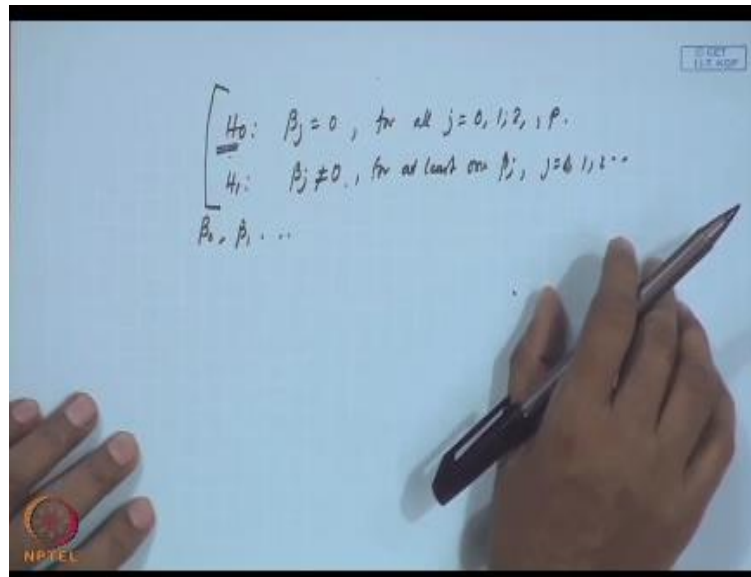


Dr J Malvi, ICM, IIT Kharagpur

5

So as a result you are doing another test.

(Refer Slide Time: 37:40)



Collectively you are basically checking that there is that influence of all the regression parameters are 0 and alternative hypothesis is at least one of the regression parameters has effect now this is the case and under this situation our test says that my model is fit.


(Refer Slide Time: 38:09)

## Goodness of fit test

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	171.23	85.62	5.01	0.034
Residual Error	9	153.68	17.08		
Total	11	324.92			

**R-Sq = 52.7% R-Sq(adj) = 42.2%**

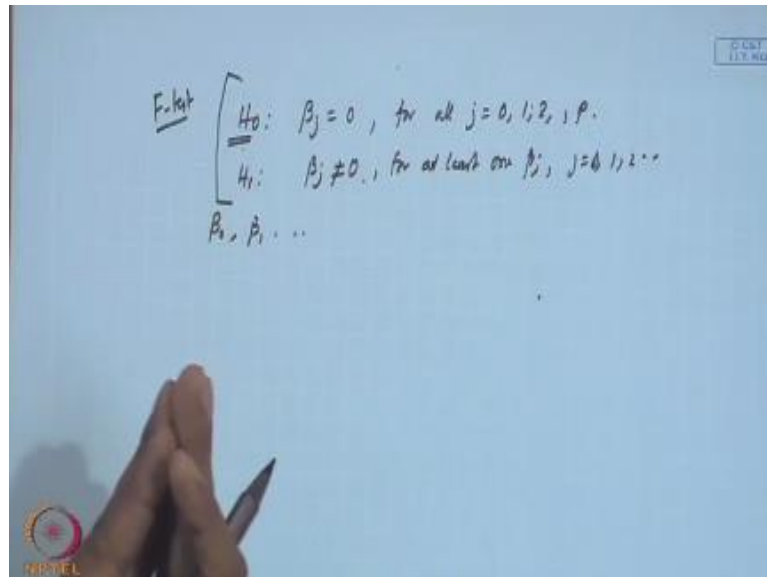


Dr / Maiti, IEM, IIT Kharegpur

5

Okay, so it is not contradictory it is, it is R square it is the overall measure, it is saying that, yes overall I you may know to be happy with the explanation, because you are not able to explain the total variation of the y variable, but you are able to explain 52%. So, that means some variables are contributing and that is being tested here. If this is the case then definitely we want to test the individual parameters.

(Refer Slide Time: 38:50)



Now, collectively this f test is saying that there are some parameters, which are contributing, f test is saying, R square is saying that overall variability explanation is poor. Now, of R square you may disconnect the model absolutely no problem, we may not go further, but as 52% variability is explained 52.7% variability is explained. So, then you may be interested to know also that it is something, which you know now. So, I want to know, which variable is contributing towards this direction?

(Refer Slide Time: 39:34)

$$F\text{-test} \begin{cases} H_0: \beta_j = 0, \text{ for all } j = 0, 1, 2, \dots, p. \\ H_1: \beta_j \neq 0, \text{ for at least one } \beta_j, j = 0, 1, 2, \dots \end{cases}$$

$$\beta_0, \beta_1, \dots$$

Test of individual parameters.  

$$\beta_j, j = 0, 1, 2, \dots, p$$

$$E(\hat{\beta}_j) = \beta_j$$

$$C = (X^T X)^{-1} = \begin{bmatrix} c_{00} & & \\ & c_{11} & \\ & & \ddots \end{bmatrix}$$

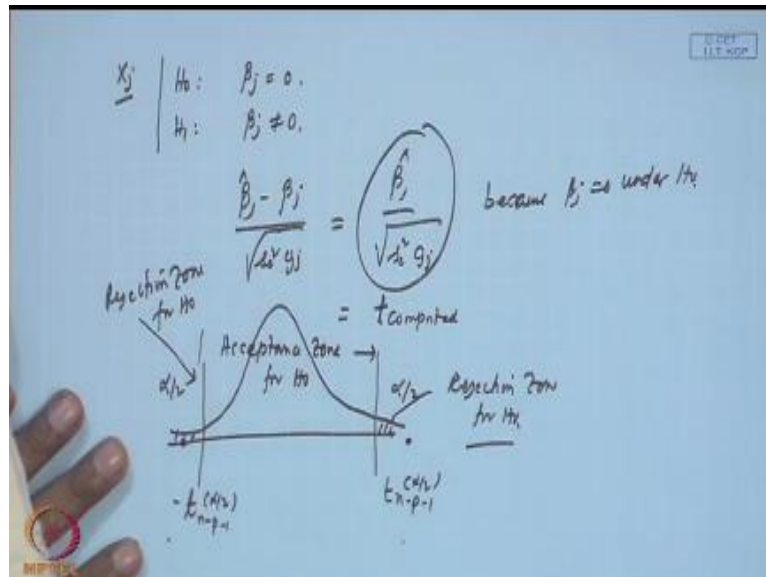
$$\frac{\hat{\beta}_j - E(\hat{\beta}_j)}{SE(\hat{\beta}_j)} \sim t_{n-p-1}$$

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{c_{jj} \sigma^2}} \sim t_{n-p-1}$$

So, then next topic is test of individual parameters. So, individual parameter mean  $\beta_j$  0, 1, 2, p you have seen earlier that the  $\hat{\beta}$  the estimate of  $\beta$ . So, that means expected value of  $\hat{\beta}_j$  will be  $\beta_j$ , okay. Now we can also create one statistics like  $\hat{\beta}_j$  minus expected value of  $\hat{\beta}_j$  by standard error of  $\hat{\beta}_j$ . This will follow t distribution n-p-1 will be the degrees of freedom, okay. So, that means this quantity this quantity becomes  $\hat{\beta}_j - \beta_j$  by now this value SE  $\hat{\beta}_j$  can you remember, what will be this value last class  $SE^2 C_{jj}$  and last class we have seen this one,  $C_{jj}$  where C is nothingbut  $(X^T X)^{-1}$  this will be a matrix of p x p.

And somewhere, in the diagonal line that  $c_{jj}^{\text{th}}$  variable is there this you are taking as  $C_{jj}$  fine. Now, this we are saying it will follow this distribution. Now, this will follow t distribution we are we are saying this follow under  $H_0$ , so we will test some.

((Refer Slide Time: 42:02))



Hypothesis here what is this  $H_0$ ,  $H_0$  is  $\beta_j$  equal to 0 then  $H_1$  is  $\beta_j$  not equal to 0 what is the difference this versus f test here we are saying  $\beta_j$  equal to zero for all and alternatively at least 1 is not 0 here you are not testing collectively you are testing for a particular variable  $X_j$  its contribution he contribution is  $\beta_j$  so you are saying  $H_0$   $\beta_j$  equal to 0 and  $\beta_j$  not equal to 0 we are not considering the other variables getting me so under this case what will happen to this that  $\beta_j \text{ cap} - \beta_j$  by  $Se^2 C_{jj}$  will become  $\beta_j$  by  $Se^2 C_{jj}$  because  $\beta_j$  equal to 0 under  $H_0$  so we are putting  $\beta$  value 0.

So this is what is t value you find out in any software if you run you get what I can say that output in that output you will find out one table individual parameters then the standard error then t values this t value is this  $\beta$  this one now this quantity follows t distribution what does it mean it is a two tailed distribution so this side let it be  $\alpha$  by 2 this side also  $\alpha$  by 2 then this is  $t_{n-p-1, \alpha/2}$   $t_{n-p-1, \alpha/2}$  this is - this is + so if your computed t this is the t computed  $\beta_j \text{ cap}$  by its standard deviation.


If this one falls here or here either the left extreme or right extreme that the rejection zone then  $H_0$  will be rejected so this is our acceptance zone for  $H_0$  then this side is rejection zone for  $H_0$  and this is also rejection zone. Correct so let us see what is I think you will be able to do this one with

all the information are with you all information i we have already seen  $\beta_j$   $C_{jj}$  also given to you  $Se^2$  is known to you everything is known to you so you will be able to compute this.

(Refer Slide Time: 45:24)

### Parameter test

Predictor	Coef	SE	T	P	VIF
Constant	130.22	26.43	4.93	0.001	
Absenteeism%	-1.2432	0.4480	-2.78	0.022	1.092
Machine BH	-0.2999	0.4586	-0.65	0.529	1.092



Dr. J. Mallick, IIT Kharagpur

6

Now this is the parameter test table in this table you see this is the predictor or explanatory variables or independent variable one is  $\beta_0$  which is the constant term absentee is m machine breakdown hours now we have seen that  $\beta_0$  is one 30.22 absentee is m - 1.24 and machine breakdown hours - 0.30 since that we have given now these are the standard error the standard error just let me see this you see this.

(Refer Slide Time: 46:07)

## Regression parameter estimates

$Y = 130.22 - 1.24X_1 - 0.30X_2 + e$

Observed	Fitted	Residuals	$C = (XTX)^{-1}$		
100	109.44	-0.44	40.9	0.114	-0.703
110	102.88	7.12	0.31	0.012	-0.004
105	102.32	2.68	10.7	-0.004	0.012
94	94.82	-0.82			
95	96.41	-1.41			
99	100.69	-1.69			
104	105.02	-1.02			
108	108.45	-0.45			
105	105.07	-0.07			
98	105.71	-7.71			
103	104.42	-1.42			
110	104.77	5.23			
			SSE	$Y'(I-H)Y$	155
			$se^2$	$SSE/(n-p-1)$	17.22

Dr. J. HAN, IISc BT KTH/IGIPU 4

This is c matrix. In this C matrix, how these things are coming?



(Refer Slide Time: 46:17)

Handwritten notes on a screen showing the calculation of the variance-covariance matrix  $C$  and the standard errors of the regression coefficients  $\beta_0$ ,  $\beta_1$ , and  $\beta_2$ .

The matrix  $C$  is given as:

$$C = (X^T X)^{-1} = \begin{bmatrix} 40.9 & & \\ & 0.012 & \\ & & 0.012 \end{bmatrix}$$

The standard error of the error term is given as  $Se^2 = 17.08$ .

The standard errors are calculated as:

$$SE(\beta_0) = \sqrt{17.08 \times 40.9} = 26.43$$

$$SE(\beta_1) = \sqrt{17.08 \times 0.012} = 0.448$$

$$SE(\beta_2) = \sqrt{17.08 \times 0.012} = 0.458$$

The corresponding t-values are calculated as:

$$t = \frac{\text{estimate}}{\text{SE}}$$

$$t_0 = \frac{130.22}{26.43} = 4.93$$

$$t_1 = \frac{-1.24}{0.448} = -2.78$$

$$t_2 = \frac{-0.30}{0.458} = -0.65$$

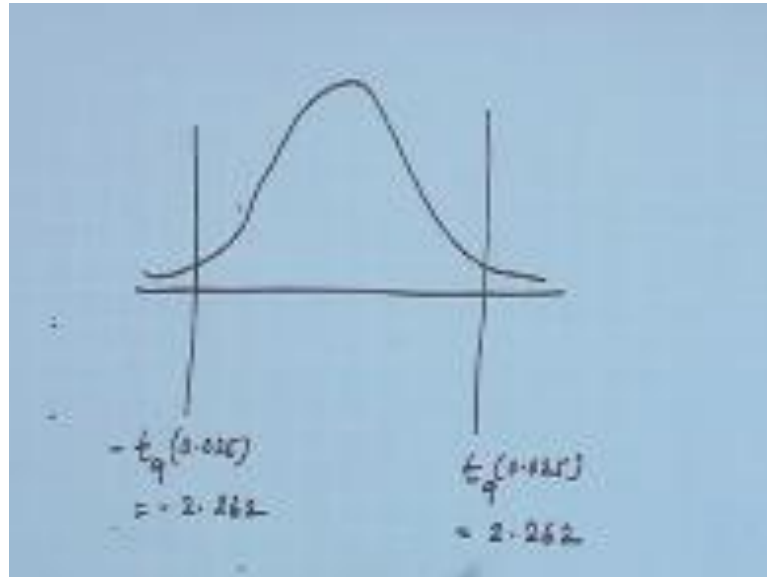
You can write that  $C$  is  $X$  transpose  $X$  inverse. This is inverse, this one I have written in  $X_1$  this is  $X$  transpose  $X$  inverse. Now, this value is 49, I am writing only the diagonal 1.2 0.12 off diagonal elements are also there. So, then this one is  $C$  and your standard error of  $\beta_j$  this is  $\sqrt{Se^2 C_{jj}}$  what is  $Se^2$  value we got?  $Se^2$  I think I told you that 17.08 so  $Se^2$  is 17.08 correct so if I want to find out standard error of  $\beta_0$  then you will be writing 17.08 into what you multiplied 40.9.

So this value you will get some value this value will be basically, 26.43. So, similarly standard error of  $\beta_0$  it will be 17.08 x 0.012. This value will be your 0.448 and standard error of  $\beta_0$  this is also same 0.8 into this one is also 0.02012 and some values are there 0.125 or 4 something there, this one is 0.058 here it will be 0 like this so rounding errors effects are there. Once you know this values that  $Se$  are known, what will be your  $t$  values corresponding  $t$  values can you not find out  $t$  values are estimate by standard error first one estimate is 130.2  $\beta$ .

Okay I think you can still remember this one 30.22 - 1.24 - 0.30, that I have given you earlier. So, this is the  $(\beta_j \beta_0 \beta_1 \beta_2)$  estimate value. So, this divided by standard error standard error is 26.43 second one is how much -1.24 divided by standard error 0.448. Then third one is your minus 0.30

divided by 0.5458 It is zero  $H_0$  is saying that  $\beta_j = 0$ . Then this values, all these values, this value will become 4.93. This is -2.78, this is -0.65. Correct? If this is the case, then you see one thing

(Refer Slide Time: 50:11)



Then you see one thing that this is my t distribution. So, I am creating my acceptance zone, this is  $t_{n-p-1}$  mean  $9\alpha$  value. Let it be 0.025 that is  $0.0205 / 2$  this is  $-t_{9, 0.025}$ , this value will be I think 2.62 so.

(Refer Slide Time: 51:01)

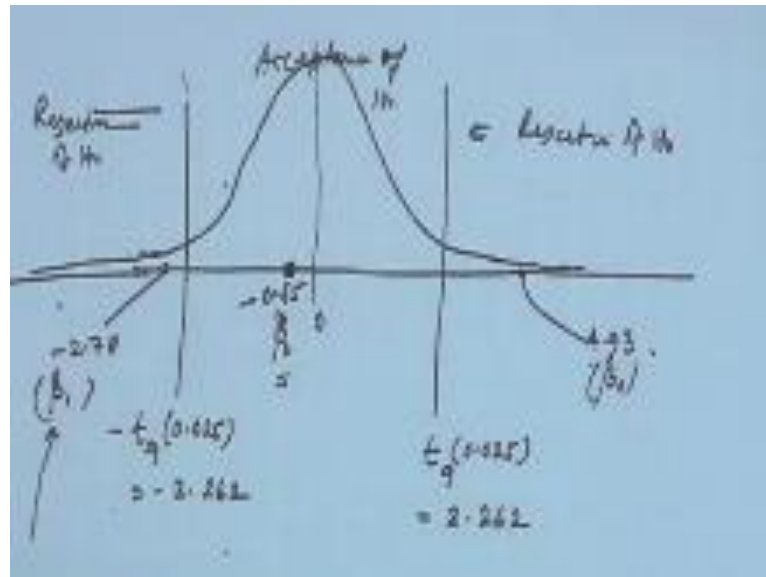
Handwritten mathematical notes on a blue background. The notes include the following:

- At the top left, a vector  $\begin{bmatrix} 130.22 \\ 24 \\ 30 \end{bmatrix}$  is written.
- Below it, a matrix  $\begin{bmatrix} 40.9 & & \\ & 0.012 & \\ & & 0.012 \end{bmatrix}$  is shown, with a circled  $0.012$  below it.
- To the right, the formula  $SE(\beta_j) = \sqrt{A_c^{-1} S_j}$  is written.
- Below that,  $A_c^{-1} = 17.08$  is written.
- Three standard error calculations are shown:
  - $SE(\beta_0) = \sqrt{17.08 \times 40.9} = 26.63$
  - $SE(\beta_1) = \sqrt{17.08 \times 0.012} = 1.448$
  - $SE(\beta_2) = \sqrt{17.08 \times 0.012} = 1.448$
- On the right side, a table is written:

	$\hat{\beta}$
$130.22 / 26.63$	$= 4.93$
$-124 / 1.448$	$= -85.7$
$-23 / 1.448$	$= -15.9$

What is your  $\beta_0$  value? 4.93, so

(Refer Slide Time: 51:05)

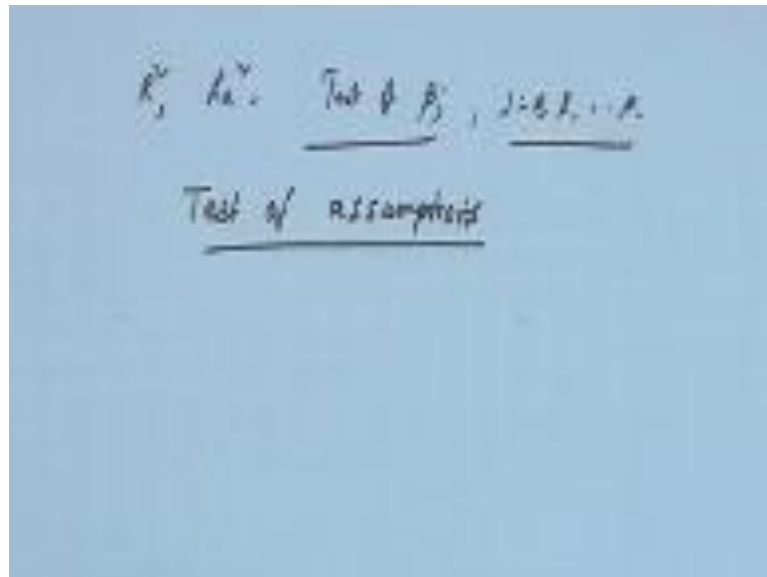


4.93 mean somewhere here. So, 4.93 second one is - 2.78, somewhere here third one is - this is my 0 -0.65. So, you have seen this is your rejection zone rejection of  $H_0$  rejection of  $H_0$  acceptance. Now, this one is related to  $\beta$  zero cap, this value this is  $\beta$  one cap related to this is  $\beta$  two cap.  $B_0$  rejected  $\beta_1$  also rejected only  $\beta_2$  is accepted.

So, that is why, what happen at least one of the variables. There are two variables  $X_1$  and  $X_2$  constant is the intercept term we are interested in this two concept is also there it is also significant. So,  $\beta_1$  is significantly contributing, if it is contributing, what is its contribution that also you may be interested to know? So, you also be interested to know, what will be the contribution of  $X_2$ ? This is the parameter that if one unit change in  $X_1$  is there  $\beta_1$  unit change in  $y$  will be there.

But from regression SSR point of view SSR can be your  $s_s \beta$ . That is  $\beta$  one also or  $X_1$  I can say  $S_s$   $X_1$  and  $X_2$  all those variables they are also contributing. So, you think if you can find out actually in regression, we only require this thing. We require this  $\beta$  value okay,

(Refer Slide Time: 53:36)



So this is what our parameter test is? So, essentially I told you that  $R^2, Ra^2$  and test of parameters test of  $\beta_j, j = 0$  to  $p$  this is the test after that you have to do the test of assumptions. Your model is fit and you are considering the model, then only the individual parameters and test of assumptions are required. Otherwise, if you think that the model fit is not good you are disconnecting the model no need of fitting the test of assumptions further, but when you accept a model you must do the test of assumptions. I think test of assumptions we will consider in the next class.

**NPTEL Video Recording Team**

**NPTEL Web Editing Team**

**Technical Superintendents**

**Computer Technicians**

**A IIT Karagpur Production**

**[www.nptel.iitm.ac.in](http://www.nptel.iitm.ac.in)**

**copyrights reserved**