Engineering Econometrics

Prof. Rudra P. Pradhan
Vinod Gupta School of Management
Indian Institute of Technology, Kharagpur

Lecture – 19
Linear Regression Modelling (Contd.)

Hello everybody. This is Rudra Pradhan here. Welcome to Engineering Econometrics. Today, we will continue with Linear Regression Modelling in that to the structure of simple regression modelling and in other words, it is a structure of bivariate econometric modeling. In the last couple of lectures, we have already discussed the issues of you know regression modelling that too with respect to two variables where, one is treated as a dependent variable and the other one is treated as independent variable. And, we have gone through the kind of you know structures, how to build the model, bivariate model and how to estimate the model and how to do you know empirically test a particular you know problem.

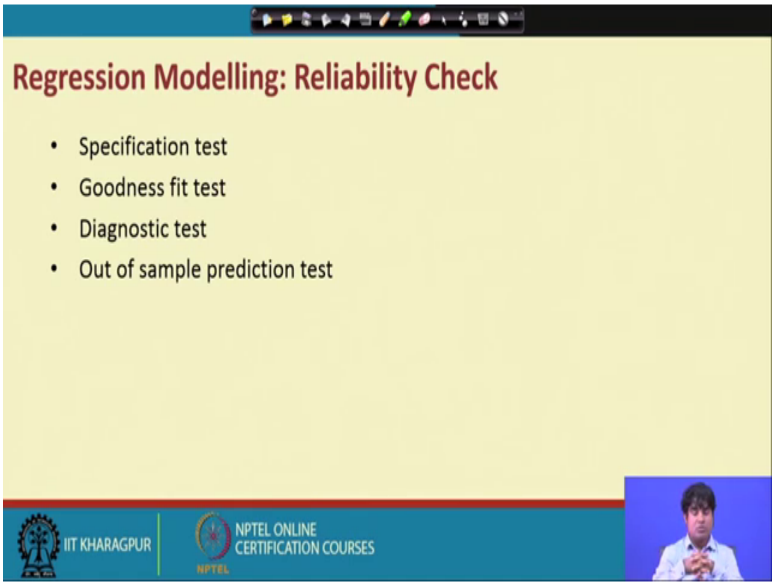
The way we have discuss his you know it is mostly the derivation of the parameters corresponding to the availability of dependent variable, independent variables and the kind of you known data. And for instance, you know the with two variables y and x and the regression modelling simply represented as a y equal to alpha plus beta x and against by various mechanism and with the help of you know ability of data we like to estimate the model and that to know the values of the parameters alpha and beta respectively. And after knowing the you know parameters, then will have a kind of you know regression line that is the, that is the estimated lines or it is called as a line of the best fit and on the basis of that estimated line or line of the best fit, we like to predict y with the availability of x or any value of x you can predict y.

Since, we have the information about alpha and beta, but the issue is that having the data and the you know identification of dependent variable and independent variable and with the application of values taken in, we can get the estimated values that is alpha hat and beta hats. And, for that we can go ahead with the prediction and forecasting but the structure of econometrics is that we have to test the model before you use the model for any kind of you know prediction and forecasting's whatever may be the engineering problem.

So, once we have the estimated model, even you know with the values of you know parameters that is alpha you know alpha hat and beta hat and we like to test the model and then we have to validate that the particular estimated model is good one or best one and free from all kinds of you know errors and after that we can go ahead with the prediction and forecasting's. Of course, we have gone through some standard checks by at obtaining the error term which is the difference between the actual and the predicted and with the help of this, you know error terms we write to verify certain things that sum of the error term should be equal to 0 and error variance should be homogeneous in nature. There should not be any kind of you know heteroscedasticity issue.

So, the model can be used for prediction and forecasting but only checking the sum of the error and the error variance it is not enough to justify or to say that the model is the best one or the line is the best one to go ahead with prediction in forecasting. So, what are the things we are supposed to check that we will discuss today and after knowing all this things, then we can extend this model in the context of multiple regression modelling and multivariate regression modelling. So, this standard structure is like this.

(Refer Slide Time: 04:39)



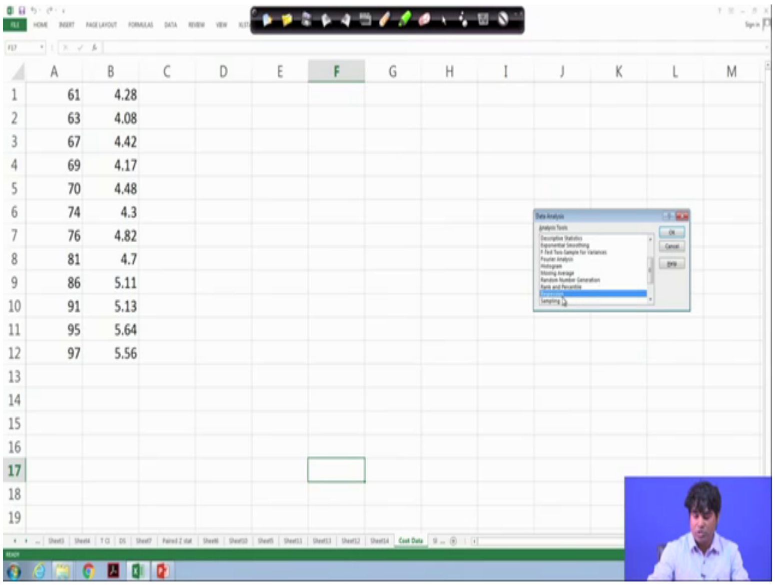
Regression Modelling: Reliability Check

- Specification test
- Goodness fit test
- Diagnostic test
- Out of sample prediction test

So, once we get the estimated models so, we have to go through the followings you know test procedure. So, first one is the Specification test, second one is the Goodness fit test, third one is the Diagnostic test and fourth one is the Out of sample prediction test. Of course, you know in the last lectures, we have we have use the software that is the

Excel data analysis package to obtain these parameters and including these parameters we have gone through so many other you know outputs.

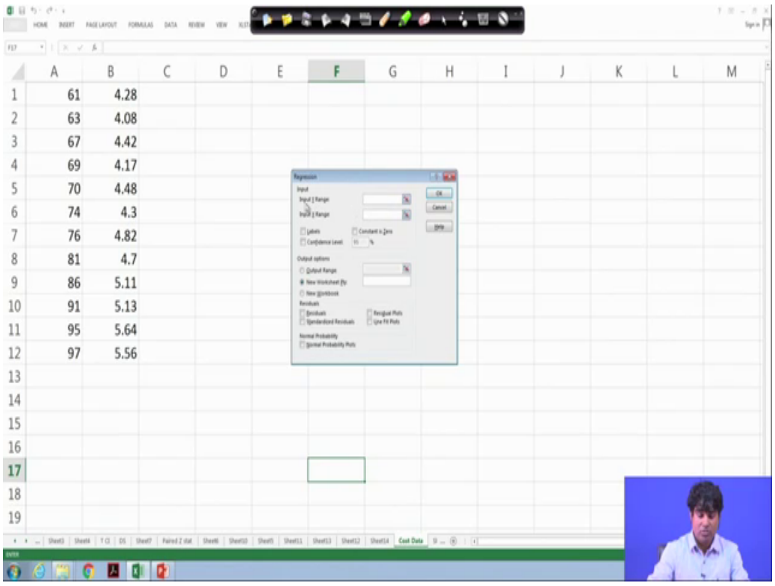
(Refer Slide Time: 05:24)



	A	B
1	61	4.28
2	63	4.08
3	67	4.42
4	69	4.17
5	70	4.48
6	74	4.3
7	76	4.82
8	81	4.7
9	86	5.11
10	91	5.13
11	95	5.64
12	97	5.56

For instance, if you go through these particular you know problem, so this is the problem which we have actually analyze yesterday, that is what actually we like to justify once again I am highlighting how you have obtain these regression results. So, put you know you know just click that analysis, then the package will appears. So, just put on regression.

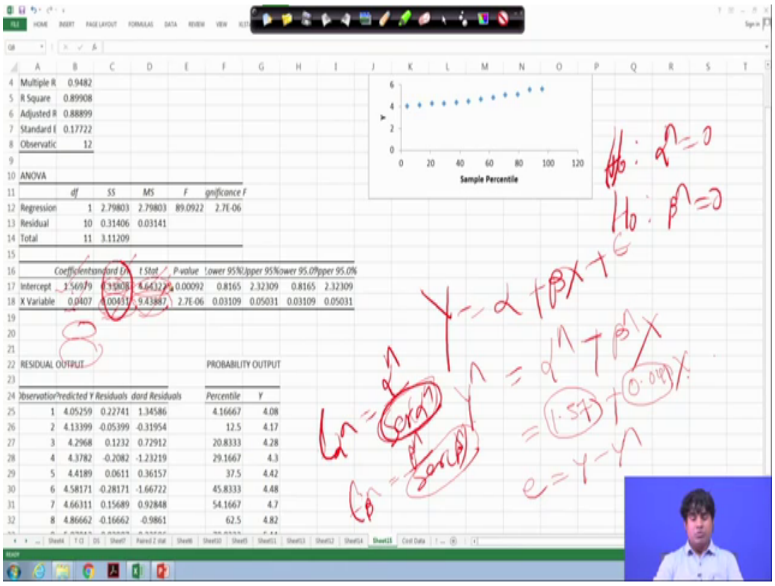
(Refer Slide Time: 05:48)



	A	B
1	61	4.28
2	63	4.08
3	67	4.42
4	69	4.17
5	70	4.48
6	74	4.3
7	76	4.82
8	81	4.7
9	86	5.11
10	91	5.13
11	95	5.64
12	97	5.56

Then, the box will appear like this and so, the box will you know the means the box should ask to put to y indication and x indication that is the range. So, here the second one is the y range and we have to highlight the particular cost factor and then, we have to put the passengers numbers and then allow the software to go ahead. So, we like to report all these items and then putting ok.

(Refer Slide Time: 06:22)



Multiple R	0.9482
R Square	0.89908
Adjusted R	0.88899
Standard E	0.17722
Observatic	12

ANOVA

	df	SS	MS	F	gnificance F
Regression	1	2.79803	2.79803	89.0922	2.7E-06
Residual	10	0.31406	0.03141		
Total	11	3.11209			

	Coefficients	tandard Er	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	1.56979	0.33808	4.64322	0.00092	0.8165	2.32309	0.8165	2.32309
X Variable	0.0407	0.00431	9.43887	2.7E-06	0.03109	0.05031	0.03109	0.05031

RESIDUAL OUTPUT

Observation	Predicted Y	Residuals	dard Residuals
1	4.05259	0.22741	1.34586
2	4.13399	-0.05399	-0.31954
3	4.2968	0.1232	0.72912
4	4.3782	-0.2082	-1.23219
5	4.4189	0.0611	0.36157
6	4.58171	-0.28171	-1.66722
7	4.66311	0.15689	0.92848
8	4.86662	-0.16662	-0.9861

PROBABILITY OUTPUT

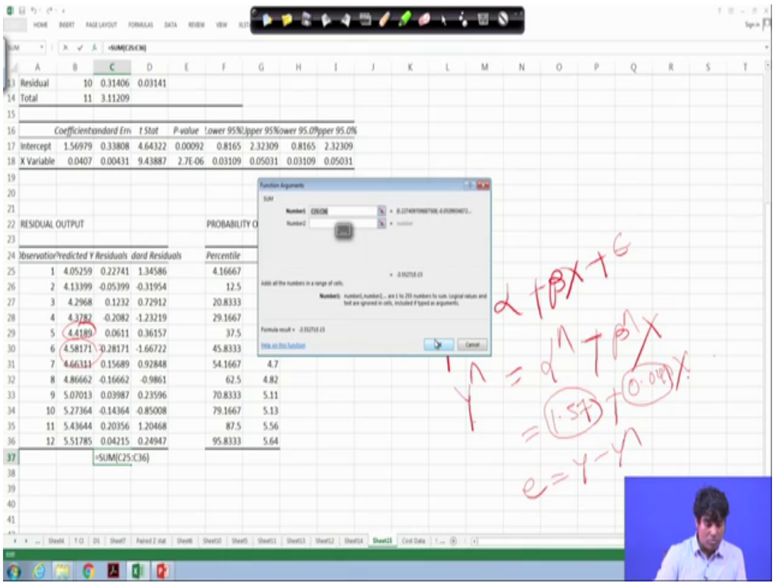
Percentile	Y
4.16667	4.08
12.5	4.17
20.8333	4.28
29.1667	4.3
37.5	4.42
45.8333	4.48
54.1667	4.7
62.5	4.82

FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW
6
4
2
0
0 20 40 60 80 100 120
Sample Percentile
$H_0: \hat{\alpha} = 0$
$H_0: \hat{\beta} = 0$
$Y = \alpha + \beta X + \epsilon$
$\hat{Y} = \hat{\alpha} + \hat{\beta} X$
$= (1.57) + (0.041) X$
$e = Y - \hat{Y}$
$t_{\hat{\alpha}} = \frac{\hat{\alpha}}{Se(\hat{\alpha})}$
$t_{\hat{\beta}} = \frac{\hat{\beta}}{Se(\hat{\beta})}$
READY

And obviously, then we will get the results like. So, this what the standard results which we have obtained in the in fact, a last class, we have obtained this and we have discussed. And now, we will go more about this you know these results. So, up to last lectures, we have discuss this much only. So, that is how to obtain the parameters that means, technically for a bivariate models, so we have y equal to alpha plus beta x and then we will introduce error terms epsilon and on the basis of a estimation, we will get estimated equation that is y cap equal to alpha cap plus beta cap x.

And, by the way error term will be removed in the process and where alpha hat a we have here alpha hat that is the intercept coefficient which is equal to 1.5 a round figure 5, 1.57 plus beta hat equal to 0.041 so, 0.041 and x. So, this is the alpha coefficient and this is the beta coefficient and this is x, x which is independent variable and y hat is the dependent variable predicted and y minus y hat will get the error component which is nothing but y minus y hat. And that is how in the extreme below, in the extreme below you will be find you will be find a, you will be find the kind of you know this is what the actually predicted.

(Refer Slide Time: 07:56)



=SUM(C25:C36)

	A	B	C	D	E	F	G	H	I
13	Residual	10	0.31406	0.03141					
14	Total	11	3.11209						
15									
16		Coefficient	tandard Err	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	pper 95.0%
17	Intercept	1.56979	0.33808	4.64322	0.00092	0.8165	2.32309	0.8165	2.32309
18	X Variable	0.0407	0.00431	9.43887	2.7E-06	0.03109	0.05031	0.03109	0.05031

RESIDUAL OUTPUT

PROBABILITY O

Observation	Predicted Y	Residuals	dard Residuals		Percentile	
1	4.05259	0.22741	1.34586		4.16667	
2	4.13399	-0.05399	-0.31954		12.5	
3	4.2968	0.1232	0.72912		20.8333	
4	4.3782	-0.2082	-1.23219		29.1667	
5	4.4189	0.0611	0.36157		37.5	
6	4.58171	-0.28171	-1.66722		45.8333	
7	4.66311	0.15689	0.92848		54.1667	4.7
8	4.86662	-0.16662	-0.9861		62.5	4.82
9	5.07013	0.03987	0.23596		70.8333	5.11
10	5.27364	-0.14364	-0.85008		79.1667	5.13
11	5.43644	0.20356	1.20468		87.5	5.56
12	5.51785	0.04215	0.24947		95.8333	5.64
		=SUM(C25:C36)				

$\alpha + \beta X + \epsilon$

$\hat{Y} = \hat{\alpha} + \hat{\beta} X$

$= \boxed{1.57} + \boxed{0.04} X$

$e = Y - \hat{Y}$

That means technically, the predicted equation equal to 1, 1.57 plus 0.04 x. So, now, for every x, if you put x value and then you will get y hat value. So, as a result, for every x we have a y predicted and then a for every x, we have we have actual y and the difference by default will give you the error terms. This is have the series we have 12 samples and for every samples, we have a predicted value that is we that is for y and against we have actual y. So, the difference with the give you the residual component this part and the so, the this is what is called as a standard residuals. And of course, so the, so the residuals will give you some kind of you know signals just you can put equal to sign and crosscheck whether the sum will be appearing you know 0 not.

So,, so this will give you some kind of you know indications how to obtain this particular you know component ok. So, this is a this is what the process but ultimately in fact, last class we have already discuss this sum is coming actually equals to 0 and what will you do here ah? We like to you know validate this models by using all these you know test structures whatever we have discussed right now that is a specification test, goodness fit test, diagnostic test and out of sample prediction test. In fact, out of sample prediction test is the kind of you know structure mostly used for time series kind of you know modelling and; that means, technically when your data is with respect to time series, then out of sample prediction test is somehow mandatory.

But in that case of you know cross sectional kind of you know estimation or cross sectional kind of you know modeling, we makes keep that particular out of sample prediction test. So, what is you know more important is to go through specification test, goodness fit test and diagnostic test. The first one is the specification test which is more or less you know you know connecting to these parameters, alpha hat and beta hats. Like you know for a bivariate model, alpha hat and beta hat are the two pillars. So, the first check that is the specification check or specification test is to you know is to you know use to justify that the value of alpha hat and beta hat are you know significant one to go ahead with the kind you know predictions. So, ultimately so, the value of alpha hat and the value of you beta hat need to be tested whether there you know statistically significant.

So, when we talk about specification test or something like you know goodness fit test or diagnostic test or out sample prediction test, so we have some you know we have step you now step by step you know process to you know proceed and then we will give a kind of you know signal that the model is free from all these you know obstacles. So, the standard procedure for specification test is nothing but you know you know going for hypothesis testing and to test that the particular parameters is a statistical significant; that means, alpha hat need to be statistical significant and beta hat need to be a structural significant. Since it is a two variable case, so we are supposed to check only the significance of alpha hat and beta hat and where there is no question of the you know any compromise.

So, it is question of you know yes no because there is only one parameter which can actually validate the model that is the beta and alpha is the intercept which may not have so important you know for the time being, but beta coefficient is more important because it indicates the impact of x and y and whatever the impact or you know and the kind you know cause will you go to y. So, that should be statistically significant, otherwise that particular variable cannot be considered as the you know primary variable or the most important variable to analyze the particular you know engineering problem.

So, ultimately the procedure of hypothesis testing is you know to start with the formulation of null and alternative hypothesis and then we have to you know select a test statistics. For instance, in the case of specification test, we our objective is to test the parameters alpha and beta. So obviously, the null hypothesis will be the null hypothesis will be simply H 0 where alpha hat equal to 0 and again for beta it is beta hat equal to 0, then to delegate this particular you know statements we like to use theta T statistics and then we like to calculate T of alpha hat and T of beta hat and then we have to you know check whether the particular value is statistically significant or not.

So that means, first you have to formulate the null a null hypothesis and then, the choice of the test statistic. In the case of specification test which specifically use T statistics and then we like to fix the probability level. So, we can fix at 1 percent level, 5 percent level and 10 percent level. So, on the one side, we have to report calculated T statistics and then on the other side, we have to report the tabulated statistic on the basis of the a probability level 1 percent, 5 percent and 10 percent. So, the that is called as a critical value or tabulated value; the critical value and tabulated value specifically depends upon three things that is the a probability levels at 1 percent, 5 percent or 10 percent and then the a degree of freedom which exclusively depends upon the value of n and k that is the sample size and the number of variables or number of parameters in the system.

So, we have to go to the T tables and then corresponding to the degree of freedom and at a particular probability level you have to find out the critical value. So, that is one side of the game and the other side of the game, we have to report the calculated T statistic which is nothing but actually T alpha it is called as a T alpha hats which is a nothing what you know alpha hat by standard error of alpha hat and then for T beta hat. So, it is beta hat divided by standard error of beta hat. So, ultimately we need to calculate standard error alpha hat and standard error of beta hat and that is nothing but you know square root of variance of alpha hat and square root of variance of beta hat. And at the value of standard error of alpha hat and beta hat depends upon the error variance that a that is sum square errors which is actually give you the kind of you know structure to report the T statistic of alpha hat and T statistic of beta hat.

So, in the slides, show we will discuss how we will obtain all these you know standard errors for alpha hat and beta hat. So, once you get the standard error of alpha hat and standard error of beta hat, so we can get the t statistics and for alpha hat and beta hat then you can compare with the critical value and this is the manual procedure or the kind of you know clarification how you have to obtain all these things and how you can validate the particular you know parameters. So, far as a statistical significance is concerned then, all these jobs you know software can do for you but it is mandatory that you should know how these are all coming and how this validation is happening.

For instance in this case, so these are all actually standard error for alpha hat and beta hat; so, where I have reported here standard error of alpha hat and standard error of beta hat. So, ultimate with the standard error of alpha hat is reported hereby software and this is also for beta hat. So that means, technically software will be directly calculate standard error of alpha hat and standard error beta hat, then just you divide a alpha coefficient by standard error you will get it the T statistic that is t alpha and then again standard error of beta hat divide by standard error beta hats you will get actually T of beta hats. So that means, software already reported the alpha coefficient the beta coefficient, the standard error of alpha, standard error of beta and T of alpha hat and T of beta hat.

So ultimately, when you compare actually calculated with T tabulated, it is a kind of you know systematic process. Of course, manually there is a standard structures. So, usually we check at three levels; 1 percent, 5 percent and 10 percent and the is the signal is a if the calculated value will overtake the critical value, then we can reject the null hypothesis and in that contest, we can say that the particular parameters is statistically significant. So, first we start with 1 percent because at 1 percent, 1 percent level, the critical value is higher one then if you could not reject the null hypothesis there.

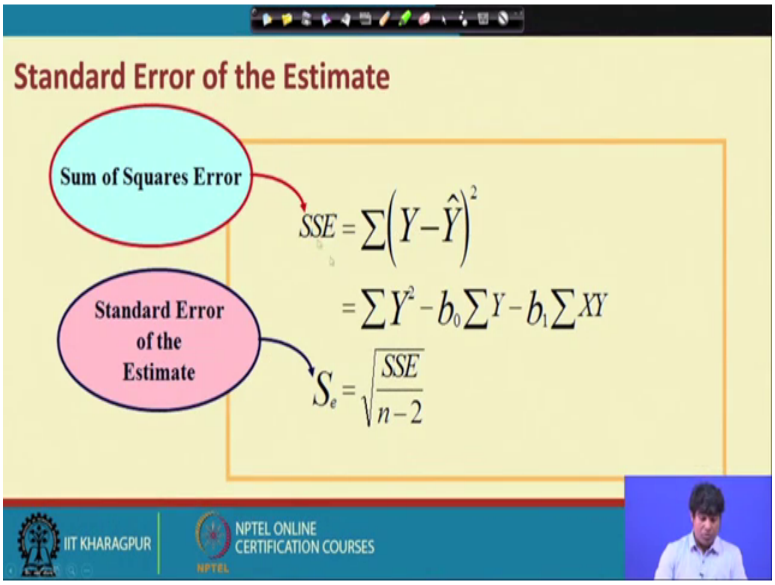
So, you can come back to 5 percent where the critical value which is slightly lower compare to at compare to 1 percent level then again you check the calculated with the tabulated, then you then you have to see whether the calculated will overtake the tabulated and if that is the case, then you can reject the null hypothesis there and if you could not, then again you come down to 10 percent where the critical value will be again lower compare to 1 percent and 5 percent against you have to compare the calculated with tabulated at 10 percent.

If a that will across the tabulated value, then you can reject the null; if not then finally, we have to conclude that you know the particular parameter is not statistically significant. But if you connect with the software, software will give you exact probability level where the particular parameter will statistically significant because these critical values are actually derive on the basis of simulation and that say when you put all these you know parameters and proxies, then automatically it will generate the a probability values you know start with you know 1 percent to 100 percent. Since, probability is between you know 0 to 1, so obviously the kind of you know significance level, so it move from 0 to 1 only.

So accordingly, so the fourth columns it gives you know p values. So, here the alpha coefficient is also statistically significant and beta coefficient is also statistically significant in that too at 1 percent reliable and that too at the highest level. So that means, the first check of this particular process is this specification test where we have to see whether the particular parameter is coming statistically significant or not. So, for that, we have to take the help of the statistics and follow the hypothesis testing of you know T test and then come with the kind of you know conclusion whether the particular parameter is statistically significant or not.

And in this case, we have to check alpha hat and hat and for this problem and with this particular you know result, will estimated output so, we can come to the conclusion that alpha is statistically significant and beta is statistically significant; that means, technically a so, this model passes through the specification test and after that we have to go for you know goodness of fit test because it must be satisfied with you know all test then finally, you can declare that this model is a good ones and go ahead with you know predictions. So, then ultimately, so with this you know statistical output we can actually proceed for the kind of you know discussions.

(Refer Slide Time: 21:25)



Standard Error of the Estimate

Sum of Squares Error

$$SSE = \sum\left(Y-\hat{Y}\right)^2$$

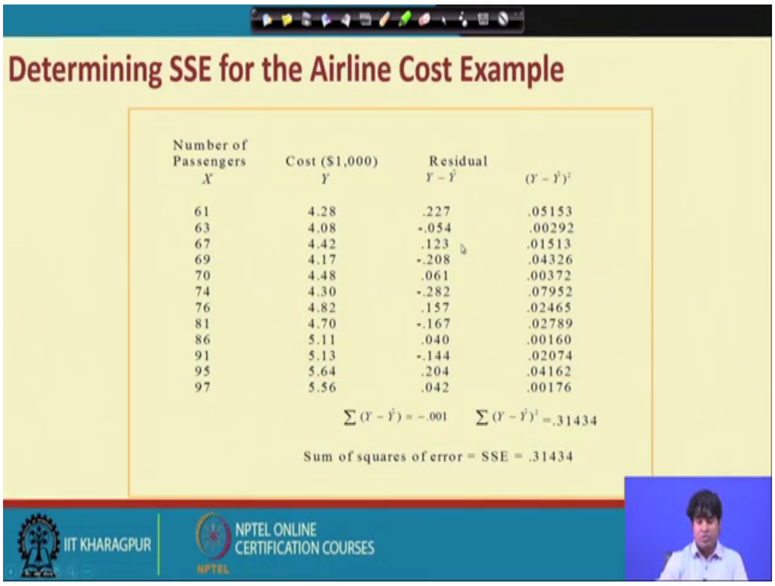
$$= \sum Y^2 - b_0\sum Y - b_1\sum XY$$

Standard Error of the Estimate

$$S_e = \sqrt{\frac{SSE}{n-2}}$$

So, ultimately so, so, here the specification test structure which we have already discuss that depends upon the standard errors which actually depends upon you know sum of the squares that is nothing but the difference between Y minus Y hat squares that is the error component and this is what the formula through which actually we can get the sum of the error, sum of squares of errors.

(Refer Slide Time: 21:46)



Determining SSE for the Airline Cost Example

Number of Passengers X	Cost ($1,000) Y	Residual $Y-\hat{Y}$	$(Y-\hat{Y})^2$
61	4.28	.227	.05153
63	4.08	-.054	.00292
67	4.42	.123	.01513
69	4.17	-.208	.04326
70	4.48	.061	.00372
74	4.30	-.282	.07952
76	4.82	.157	.02465
81	4.70	-.167	.02789
86	5.11	.040	.00160
91	5.13	-.144	.02074
95	5.64	.204	.04162
97	5.56	.042	.00176
		$\sum(Y-\hat{Y}) = -.001$	$\sum(Y-\hat{Y})^2 = .31434$

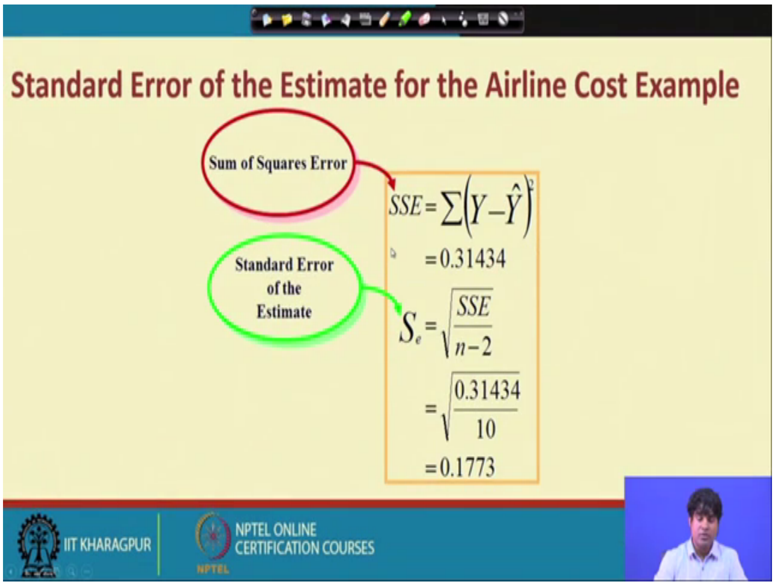
Sum of squares of error = SSE = .31434

And ultimately, this will give you the kind of you know signal whether you know particular you know parameter will be statistically significant or not. So, usually lower the s standard error, higher is the T alpha and T beta and higher the standard error you know standard error, so the if the T value will be lower. So, ultimately it depends upon again sum of the squares of the error because the standard error component, standard of component with respect to alpha hat and beta hat ultimately depends upon sum of squares of the you know error.

So, technically if for two variables case, you know sum this square errors standard error of the estimates will be a square root of sum of squares error and divide by n minus 2. For three variable case, it will be n minus 3 you know then for generalization case, it will be divide by n minus k. So, ultimately, so this is what the kind of you know structure and then we have to see how we can obtain this and this is what the actual data and we have already discussed this is X information and this is y information and this is the residuals which is nothing but y minus y hat where Y hat not reported here.

But, we can obtained by putting a you know or by using the estimated equations y hat equal to alpha hat and alpha hat plus beta hat x and ultimately we will get the residuals and after getting the residuals, so you can find out square of the residual and that is the sum of the squares of the errors. So, in the last column, so this is the indication about the sum of the squares of the error and. So, this will be give you some kind of you know structure to obtain the you know standard error of alpha hat and beta hat.

(Refer Slide Time: 23:56)



Standard Error of the Estimate for the Airline Cost Example

Sum of Squares Error

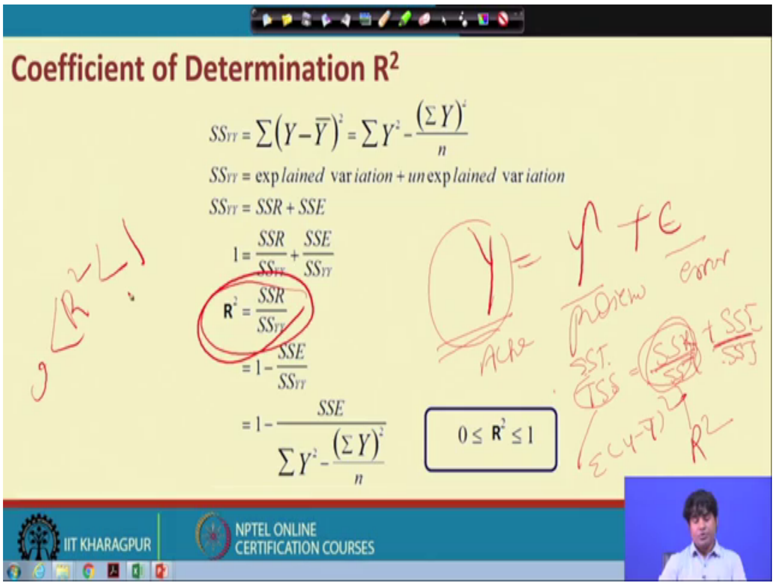
$$SSE = \sum (Y - \hat{Y})^2 = 0.31434$$

Standard Error of the Estimate

$$S_e = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{0.31434}{10}} = 0.1773$$

So, ultimately the sum of squares error for this problem is coming 0.31 and ultimately standard error of the estimate will be coming square root of this particular component and divide by n minus 2. So, it is coming ultimately 0.1773.

(Refer Slide Time: 24:13)



Coefficient of Determination R²

$$SS_{YY} = \sum\left(Y-\overline{Y}\right)^2 = \sum Y^2 - \frac{\left(\sum Y\right)^2}{n}$$

$$SS_{YY} = \text{explained variation} + \text{unexplained variation}$$

$$SS_{YY} = SSR + SSE$$

$$1 = \frac{SSR}{SS_{YY}} + \frac{SSE}{SS_{YY}}$$

$$\mathbf{R}^2 = \frac{SSR}{SS_{YY}}$$

$$= 1 - \frac{SSE}{SS_{YY}}$$

$$= 1 - \frac{SSE}{\sum Y^2 - \frac{\left(\sum Y\right)^2}{n}}$$

$$\boxed{0 \le \mathbf{R}^2 \le 1}$$

And the other part of this particular you know process is the Goodness of fit test. So, heres we have three items all together and that too y, y hat and error component. So that means, this is the actual and this is the predicted and the difference will give you the error component. So, as a result, if you actually a checking through variance factors then, you can find out the mean and then standardize it and squaring both the sides that will be give you the kind of you know structure here called as you know TSS equal to SSR cross SSE. So, SSR is called as a you know explained variations or sum square regressions and then, SSE represents sum square you know errors. So, sometimes we can use ESS plus RSS explained sum of square and residual sum of squares.

So that means, a total sum of squares which is actually a summations Y minus Y bar whole square, this is called as a total sum of squares and which is equal to sum square you know explained that is depends upon you know Y hat minus Y hat bar and then sum the error component which is the difference between the actual y and the predicted Y. So, ultimately, for any kind of you know problems when you use regression modelling that too bivariate econometrics set up, so TSS always equal to you know ESS plus RSS or

SSR plus SSE that is sum of sum of squares you know regressions and sum of squares you know error.

So, that is otherwise called as a explained variations and unexplained variations. Explained variations will take care this part and unexplained variations will take care this part and the total will exactly equal to you know exactly equal to TSS that is it sum squares totals. In other words, you can you can write this one also SST that is sum of square total, then sum of squares regressions, sum of the squares error.

So, now if you divide sum of square total in the right hand side, then it becomes 1. So, sum of square divide by sums sum of square total and this is sum of squares again total. So, ultimately, so sum of square you know this is actually regression divide by sum of square total is called as you know percentage variation of you know independent variable to dependent variable and this part actually called as a component called as you know r squares that is called as a coefficient of determination which is actually a of you know derive here. So, r square is the coefficient of determination, capital R square which is the ratio between sum of square regressions divided by sum of square totals. So, then ultimately this will give you the signal where the fitness of the particular model.

Usually, R square range, usually R square range is 0 to 1 s so. So, the it is a kind of you know, it is a kind of you know indication that you know what is the what is the particular you know fit of a model. For instance, if R square equal to 1, then the model will be perfectly fit and if R square equal to 0 the model is completely unfit and if R square equal to close to once it is a highly fit and if R square is close to let us say 0.5 this is moderately fit and if it is close to 0 it is very low fit.

So, ultimately, the structure of regression modelling depends upon goodness of fit test. So, goodness of fit test will give you the indication that you know what is the level of you know best fit. So, is it if perfectly fit the so, 100 percent correct or if it is lack of 100 percent, so maybe 90 percent, 80 percent 60 percent or it is very low fit equals to say 0 or 1 percent, 2 percent or 10 percent like that. So, it depends upon you know variations and sometimes a R square you know you know is consider as a good indications for the kind of you know prediction.

The usual understanding is that higher the R square, higher is the predictions where is lower the R square, lower is the prediction or where is the fitness, but this is not always correct, but in the first instance if you are model gives higher square, then the goodness of it will be you know consider as a high and if the R square is low, the goodness of fit to that particular model will be a low. So, ultimately we have to see what is the level of you know R squares. So, then the strength of the model depends upon these value of you know R squares high or low or moderate and then you have to go ahead with the kind of you know predictions.

So, like in the specification test, we have alpha coefficient beta coefficient then you have to test alpha coefficient and beta coefficient to validate the model. So, the goodness fit structure is also like that. So, in the goodness fit test, the indicator is the R square that is the coefficient of determination which is the ratio between sum of square regression divide by sum of square total and which a which will be lie between 0 to 1s and close to 1 is a high fit, close to 0 is low fit.

And then, then we have to check whether the particular r square is statistically significant or not. Having high R square or having low r square is a is not so important, but what is important whatever value of R square you will be obtain from the process that should be statistically significant. Compared to specification test where the job is to you know validate the parameters, so test this statistical significance. So, parameters here the job is to validate the R square, so whatever values of the R square in the models, so you have to check whether they are statistically significant or not.

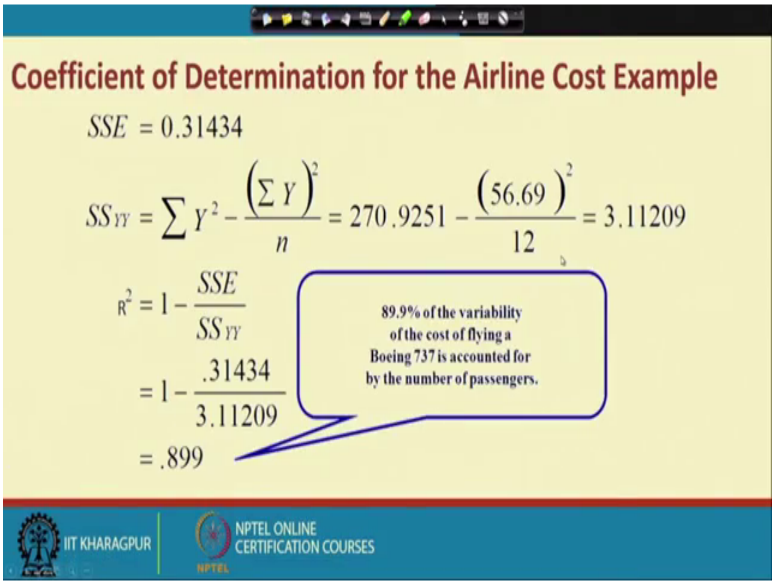
In the case of specification test, the test statistic which you frequently which is called as a T statistics, but in the case of goodness of fit test, so the test statistic we which you frequently which is called as a F statistic which is actually a the you know ratio between explained sum of square and unexplained sum of square and exclusive depends upon you know R square value. So, when R square is high and sample size is will be very high, then by default higher square will lead to high F statistics and if ask it is lower sample size sample size is low, then the value of f statistic will also comparatively low.

So, ultimately to validate this kind of you know you know parameters or this kind of you know scenario, so every times you must have high sample size then there is a high chance that the particular parameters or the particular parameter with respect to alpha beta or you know a R square will be statistically significant. So, now, corresponding to specification test where we have use T statistic here we have to use F statistics and the procedure of statistic is more or less same unlike you know T statistics. In the F statistics, the first step is to set the norm that where you know a let us say R square equal to 0 and then we have to chose this test statistic to validate R squares and for that we have we have to consider every times F and because this is the variance statistics and then and R square is the variance indicator and after using F statistic, you have to fix the probability level of significance at 1 percent, 5 percent and 10 percent.

Then, the decision making process will be the comparison between calculated F, calculated F and then the tabulated F. The calculated F which will be derived though this particular formula which you which will it again disperse in the case of a particular table called as a analysis of variance and in the tabulated statistics, the value of F exclusively depends upon the probability level of significance that is at 1 percent, 5 percent and 10 percent and the degree of freedom. Compared to T statistic, in the case of statistic, we have two degree of freedom corresponding to sum square explained and sum square error term.

So, then then you have to get the critical value for F statistics. In the other side, you we have actually calculated F statistic then we can compare calculated F with the tabulated F and check whether calculated F will overtake the tabulated at to at what level 1 percent, 5 percent and 10 percent and accordingly, we can justify the significance of this particular you know indicator that is the coefficient or determinations or the indication of you know goodness of fit. So, this is this is the standard procedure of you know goodness of fit test. So, in order to know more about this particular you know process, so let us move further and this is what the coefficient of determination with the calculation.

(Refer Slide Time: 34:28)



Coefficient of Determination for the Airline Cost Example

$$SSE = 0.31434$$

$$SS_{YY} = \sum Y^2 - \frac{\left(\sum Y\right)^2}{n} = 270.9251 - \frac{\left(56.69\right)^2}{12} = 3.11209$$

$$R^2 = 1 - \frac{SSE}{SS_{YY}}$$

$$= 1 - \frac{.31434}{3.11209}$$

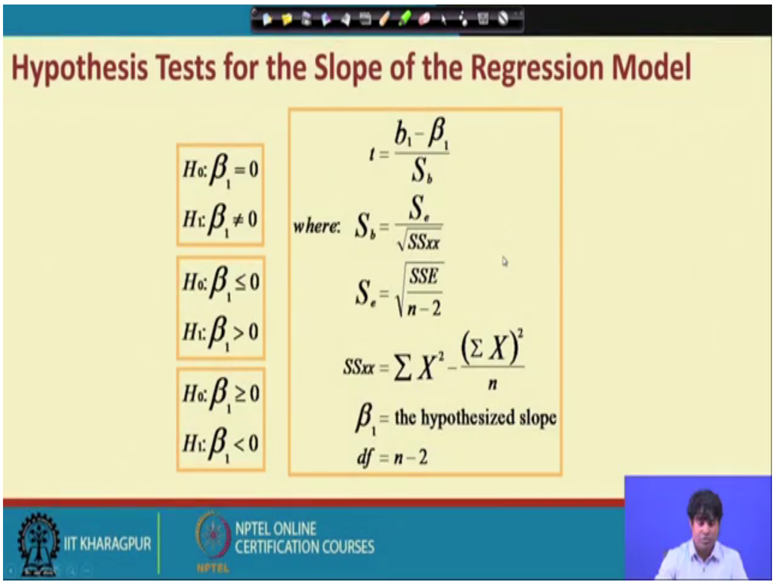
$$= .899$$

89.9% of the variability of the cost of flying a Boeing 737 is accounted for by the number of passengers.

So, ultimately R square equal to 1 minus sum of squares you know errors. So, divided by total sum of squares in sum; sum of square error we have already derived in the last slides and just putting this once and then you will calculate total sum of square which is the difference between sum of Y minus Y bar whole squares and ultimately, it is coming in this particular you know problem in this particular problem. This is coming exactly as you know 0.899; that means, around 90 percent.

So, the interpretation is that you know 90 percent of the variability of the cost of flying F you know you know airline is accounted for by the number of passengers. So that means, this is actually say most important and significant variable which can you know lead to the determination of you know airline cost. So, as per the theoretical kind of you know understanding and kind of technological hint, so the empirical result results are also supporting and validate the particular you know hypothesis.

(Refer Slide Time: 35:54)



Hypothesis Tests for the Slope of the Regression Model

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

$H_0: \beta_1 \leq 0$

$H_1: \beta_1 > 0$

$H_0: \beta_1 \geq 0$

$H_1: \beta_1 < 0$

$$t = \frac{b_1 - \beta_1}{S_b}$$

where: $S_b = \frac{S_e}{\sqrt{SSxx}}$

$$S_e = \sqrt{\frac{SSE}{n-2}}$$

$$SSxx = \sum X^2 - \frac{(\Sigma X)^2}{n}$$

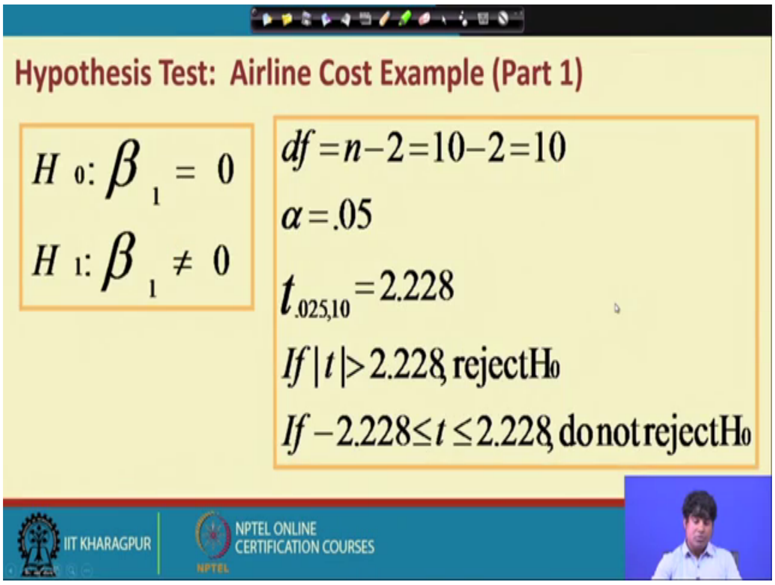
β_1 = the hypothesized slope

$df = n - 2$

So, ultimately, so this is the kind of you know structure. So, this is what actually the R square indications and this is what the a kind of you know parameter specifications and that too we have already discussed with respect to the a excel you know regression output. And ultimately, so this is this is for actually this is for you know beta coefficient then this is accordingly means standard error for you know beta coefficient and then ultimately the degree of freedom is calculated size is 12 and n minus 2 that will be coming actually 10.

Ultimately, these are all actually manual understanding the software is already you know obtain these results and then compared and also declared the statistical significance of these parameters and that is the probability at to what level of probability this parameters are statistically significant.

(Refer Slide Time: 36:54)



Hypothesis Test: Airline Cost Example (Part 1)

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

$df = n - 2 = 10 - 2 = 10$

$\alpha = .05$

$t_{.025,10} = 2.228$

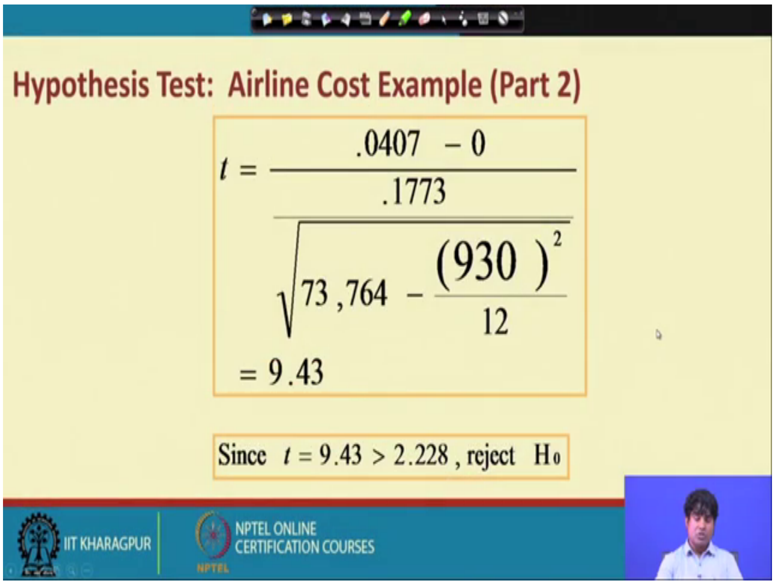
$If\ |t| > 2.228,\ \text{reject} H_0$

$If\ -2.228 \leq t \leq 2.228,\ \text{do not reject} H_0$

And ultimately, this is the procedure through which we actually you justify the you know the importance of these parameters alpha hat and beta hat and r square. Till now, we have gone through the Specification test and Goodness fit test that too to validate the regression you know equations or regression line with respect to all these parameters alpha hat, beta hat and the coefficient determination r square.

And the test procedure also we have discussed that is with respect to t test and for these specification test and the opt test for the goodness of fit test and by using all these kind of you know structure and connecting to this particular you know airline problem, so both the parameters are statistically significant.

(Refer Slide Time: 37:47)



Hypothesis Test: Airline Cost Example (Part 2)

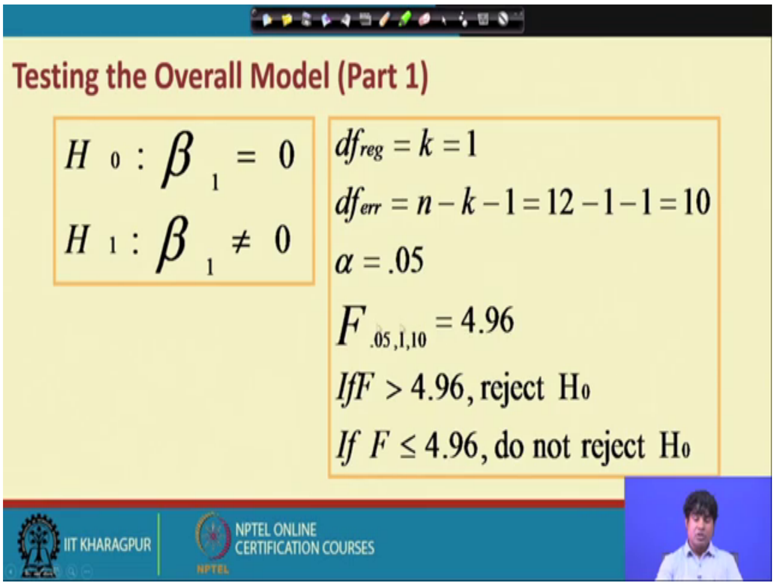
$$t = \frac{.0407 - 0}{\dfrac{.1773}{\sqrt{73,764 - \dfrac{(930)^2}{12}}}}$$

$$= 9.43$$

Since $t = 9.43 > 2.228$, reject H_0

And goodness of fit also showing statistically significant, these are all we have already discussed.

(Refer Slide Time: 37:49)



Testing the Overall Model (Part 1)

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

$df_{reg} = k = 1$

$df_{err} = n - k - 1 = 12 - 1 - 1 = 10$

$\alpha = .05$

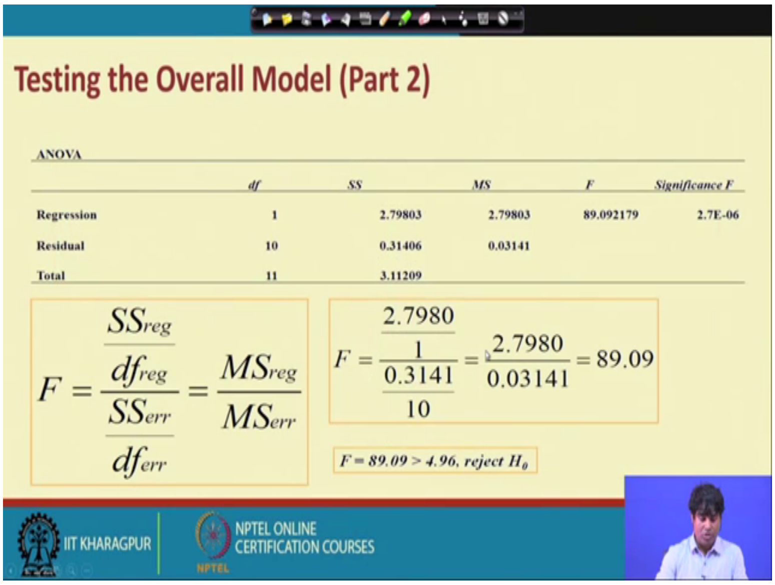
$F_{.05,1,10} = 4.96$

$If F > 4.96$, reject H_0

$If\ F \leq 4.96$, do not reject H_0

And this is what the F statistics and it is also coming actually statistical significant ok. So, as a result, so this is this is ok. So, this is 4.96 and ultimately the F statistic is coming you know this is calculate this is tabulated actually. So, 5 percent and degree of freedom 1 and you know 1 and 10. So, this is how the indication through which actually you derive the F tabulated. So, the n is the sample size k is the number of parameters.

(Refer Slide Time: 38:33)



Testing the Overall Model (Part 2)

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	2.79803	2.79803	89.092179	2.7E-06
Residual	10	0.31406	0.03141		
Total	11	3.11209			

$$F = \frac{\frac{SS_{reg}}{df_{reg}}}{\frac{SS_{err}}{df_{err}}} = \frac{MS_{reg}}{MS_{err}}$$

$$F = \frac{\frac{2.7980}{1}}{\frac{0.3141}{10}} = \frac{2.7980}{0.03141} = 89.09$$

$F = 89.09 > 4.96$, reject H_0

And then finally, we can obtain this particular you know structure. So, here is the final results how you have to obtain and ultimately the F value is actually coming 89.09 that is the calculated and the tabulated which we have here is 4.96 and if the calculated F is greater than to this tabulated value we can reject if the calculated F is less than to tabulated then, we cannot the reject and accordingly if you go to this particular final output, then the F is coming actually 89.1.

So, which is actually very high compared to the tabulated value 4.96. So, as a result, we will reject the null hypothesis and you will come to the conclusion that you know, so this model is you know with the goodness of fit test and the specification test. So, with this will stop here and we will continue this you know discussion in the next lecture.

Thank you very much.