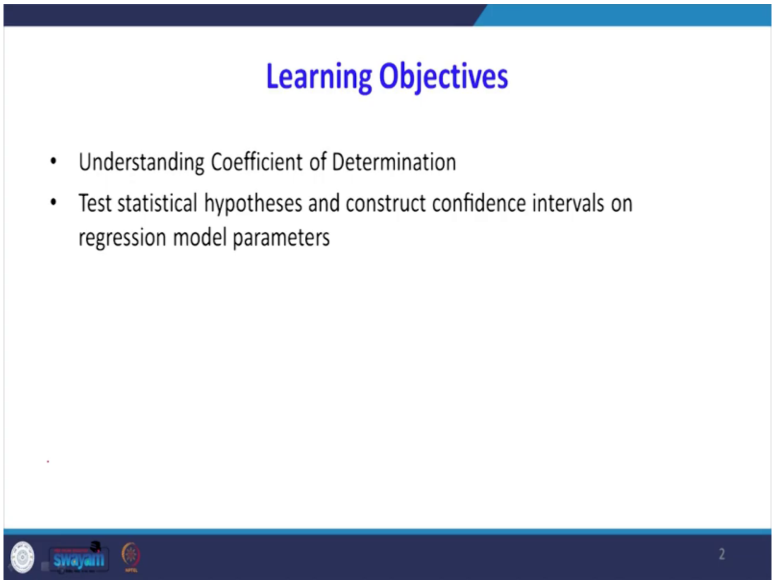
Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Management Studies
Indian Institute of Technology – Roorkee

Lecture – 30
Linear Regression – III

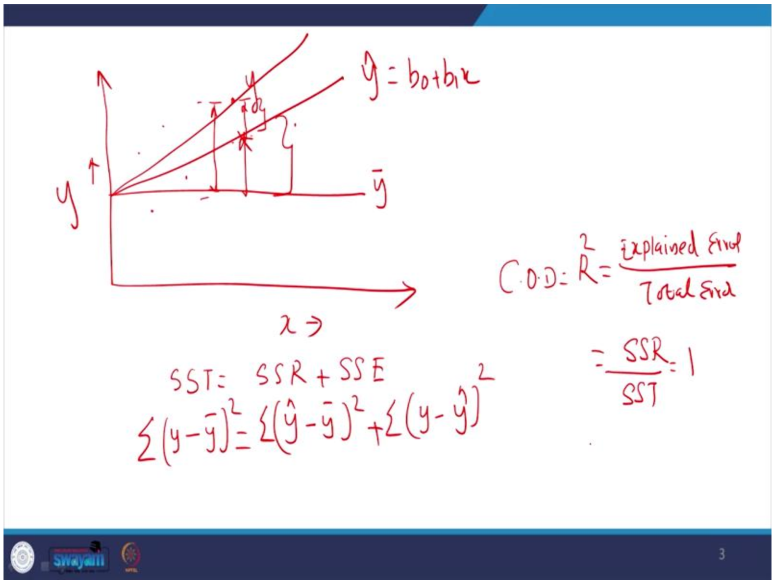
Dear students in the previous class I would take in a sample example I have explained to construct how to construct a regression equation.

(Refer Slide Time: 00:37)



In this class I will explain what is the meaning of coefficient of determination and test statistical hypothesis and construct confidence interval and regression model parameters there are one parameter is the be the coefficient of x is one parameter.

(Refer Slide Time: 01:01)



Now I will explain what is the goodness of it that is coefficient of determination r square assume that why value assume that there is no independent variable the easiest way for prediction is mean of the y what is the meaning of there is no independent variable. Suppose I have demand for say first eleven month I want to predict in 12th month what is the demand of a particular product, the easiest way is you find the mean of the previous data.

So that will be used as the mean for the next data so without considering any independent variable suppose if there is no independent variable so the actual point is say this is the y is the actual value. So, the one way to predict without any independent variable is say y bar but I know there is a one independent variable what is this much this is actual this is predicted, so this much is my error what is this error this point to this point this is error.

What is the error total error we can say total error actual - predictor. Suppose if I know one independent variable that I have as I am assuming that is affecting my dependent variable then that regression equation is like this, so this I am writing y hat equal to you can call it as b 0 + b 1 b x, now what has happened now this much distance because this point so this much distance is see total error is this much, so this much error is you could order it this way.

So here what is yes this much portion this much portion I am able to explain with the help of regression independent variable. So, total error is this point to this point so this much error with the help of independent variable x I am able to explain. So, the remaining error this one's this much distance is unexplained the error. So, what I am saying this is only one point there is no linear relationship like that there are different y values may be why here y may be here y may be here maybe here maybe here.

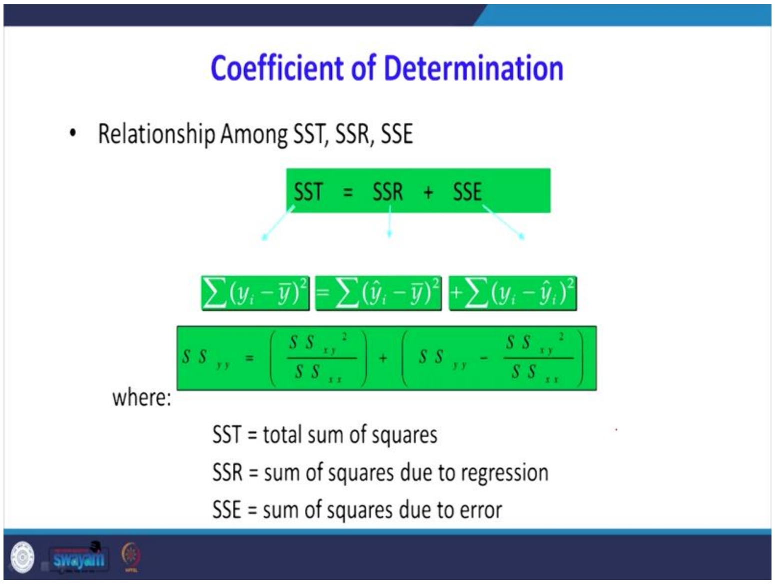
So if I find the total sum of square there is a total error then nothing but y SST . So, SST equal to SST + SSE what is the SST total sum of square what is the SSR a regression sum of square what is SSE error sum of square. So, now what is the logic behind this is the total error is this point to this point total error I am splitting that error due to how much error we are able to explain with the help of this independent variable x that is SSR, so the remaining portions that is the which is in the bracket that is unexplained error.

You when you look at this there will be a connection with on over in ANOVA what do I retain the same thing you know what you written SST equal to SS treatment + SSE your treatment is nothing but your independent variable . Now we will find out what is the formula this so what is SST so this point is y the total error y - this point is y bar whole square Sigma that is SST equal to so this much portion what is the error is y hat - y bar whole square Sigma + SSE unexplained error.

What is unexplained error y - y hat whole square Sigma so what has happened the total error is the regression sum of square + error sum of square. Here I want to predict the coefficient of determination, the coefficient of determination referred as r square is nothing but explain the error dude by total error what is explained error explained is nothing but SSR that means that much error we are able to explain because that much error is due to this independent variable x what is a total error this is SST total sum of square.

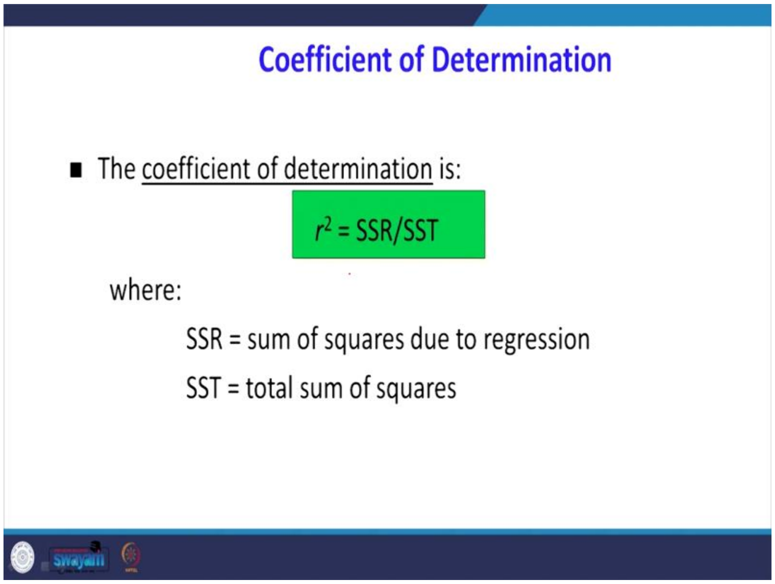
There is a two possibility of this r square is it cannot be more than 1 if it is 1 what is the meaning the total SST the numerator also SST denominator also SST so what we are saying this point this line pass through that point. If it is less than 1 so what is happening SSE error is smaller SST is bigger equal to 1 both are same. So, the upper limit of the r square is 1 the lower limit is 0 to 1 so 0 to 1 is the interval for r square.

(Refer Slide Time: 07:32)



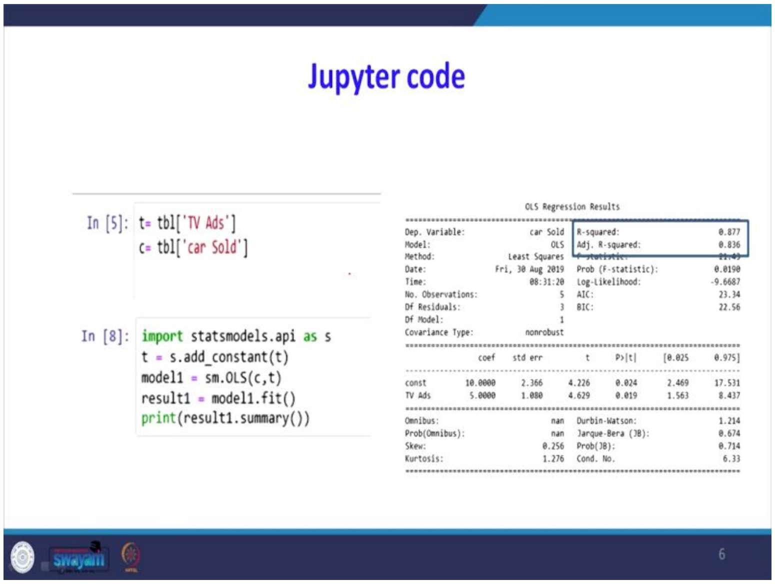
We will see this one yes see the relationship among SST, SSR and SSE. So, SST you see y - y bar whole square Sigma equal to SS our y hat - y bar whole square Sigma for SSE y - y hat whole square remember the previous also I was showing this SS yy that that SS y is nothing but Sigma of y - y bar whole square so, this is the a very handy formula if you are using calculator and a very short cut very quickly you can get the answer for what is SST SSR and SSE from that you can easily find out r square nothing but SSR divided by SST.

(Refer Slide Time: 07:58)



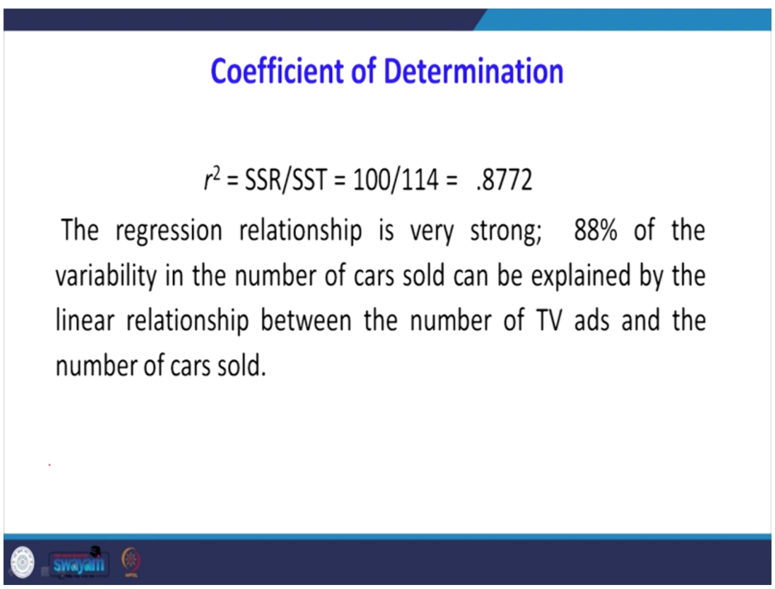
R square is SST / SST SSR is sum of square due to regression SST is total sum of square.

(Refer Slide Time: 08:07)



You see that the previous also I was saying what is the meaning of this r square. So, we are getting in our problem we are getting 0.87 I will show you how it has come yes the coefficient of determination formula is r square equal to SSR divided by SST, so SSR in our problem is 100 SST is 114 so I can show you this how this SSR 100 you see that SSR.

(Refer Slide Time: 08:42)



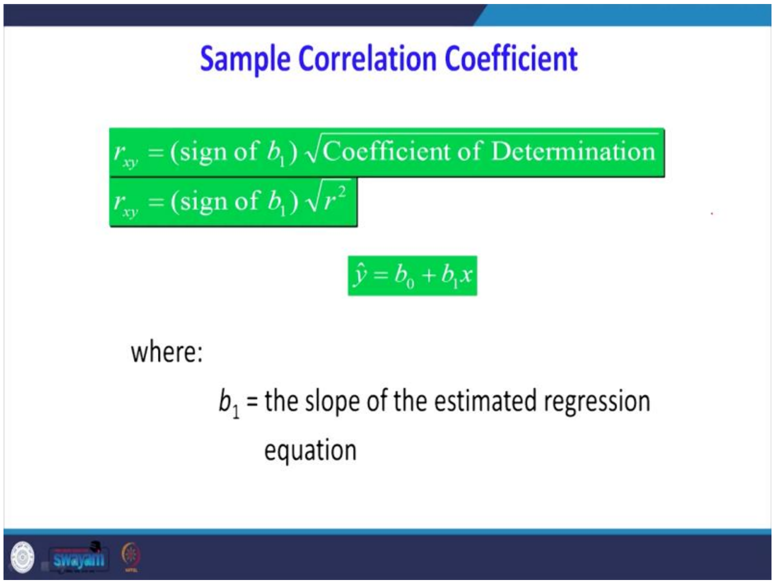
So how what is the formula for SSR we have seen previously how we are getting SSR you see that SSRis nothing but Sigma of y cap - y bar whole square first you have to find out the regression equation in that when you substitute first value of x you will get y1 cap y1 cap is when substitute the value of x into that then y - y bar whole square then when you substitute x equal to 2 they will get y 2 hat so then y bar whole square when you sum that one that is nothing but your SSR.

SST is y minus it is a numerator of that variance of y - y bar whole square so from that you can find out SST. So, in this our problem it is 100 order by 114 now what is the meaning of this r square so the meaning of r square is as we know that it is 0 to 1 the regression relationship is very strong what is the meaning is 88% of the variability in the number of cars sold can be explained by the linear relation between the number of TV ads and the number of cars sold.

So what is meaning that 87 it is 1.7 88% of the variability of y can be explained by the help of this independent variable there is a remaining 13% that we are not able to explain that may be due to two reasons one is we might all miss you that some other independent variable there may be some other variable that affects the car sales. Another reason is that we have fixed a linear regression but the actual data may follow non linear regressions so that is why we are not getting exactly one.

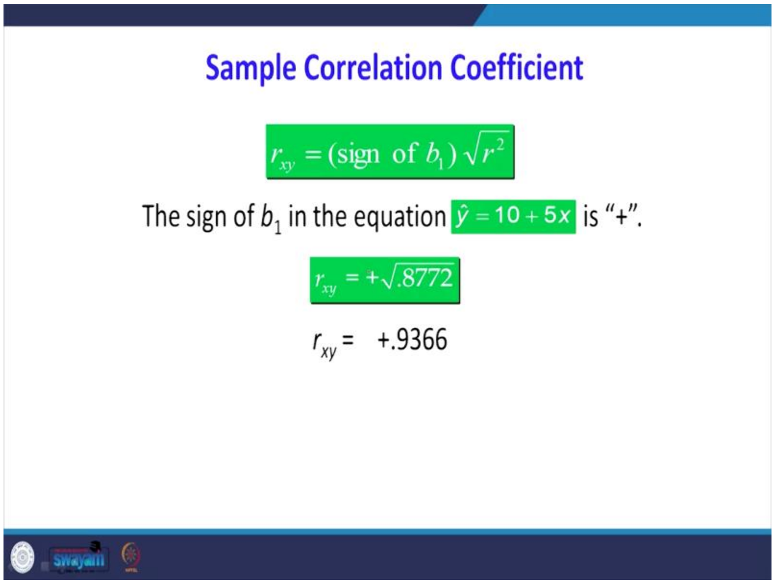
In Python output you see that when you see the r square is the 0.77 this is r square is 0.77 that is the meaning of that is 87.7% of the variability of car sold can be explained with the help of number of TV ads that is our independent variable.

(Refer Slide Time: 10:56)



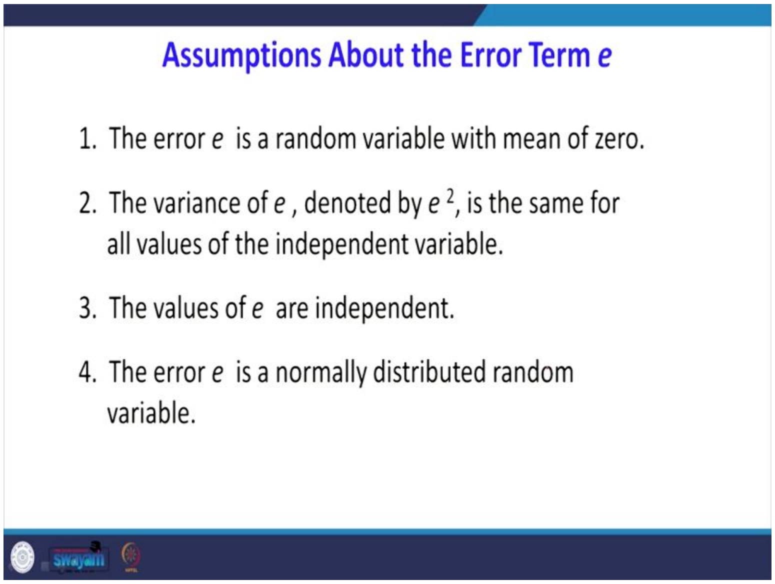
From r square we have to find out the r that is a correlation coefficient. So, the sample correlation coefficient r x y equal to sine of b 1 in our problem it is the sine of b 1 is positive root of coefficient of determination r square so sine of b 1 into r square. So, what is this b 1 is that the slope of the regression equation.

(Refer Slide Time: 11:23)



In our problem it is y equal to 10 + 5 x so the sine is + the root of 0.87 7 is this is your correlation coefficient and remember that the range of correlation coefficient is - 1/2 + 1 but the range of r square is 0 to 1. Here in the correlation coefficient if it is a - 1 it is a perfectly negative correlation if it is a + 1 it is perfectly positive correlation if it is 0 there is no correlation. In the context of r square if it is 1 it is a perfect model that means all variability of y can be explained with the help of independent variable. If it is 0 there is no relation relationship between x and y.

(Refer Slide Time: 12:13)

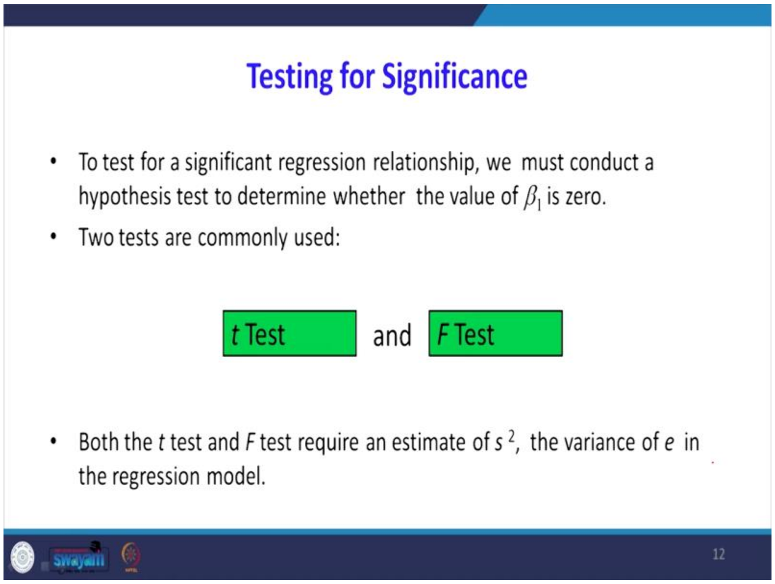


Another important point is assumptions about the error term e, here the two tests the goodness of the model not only r square is important you have to plot the error term. When you look at the error term we have to look at the behavior of that error is nothing but actual minus predicted value, so what are the assumption of the error term the error E is he a random variable with mean equal to 0, so the error has to be appear in a random manner where the sum of positive error should be equal to sum of negative errors so that sum will be 0.

The variance of e denoted by e square is the same for all values of the independent variable, so that a concept called a homoscedasticity what is the meaning of that one is if there are many x1 say x 2 x 3 independent variable these variants of x 1 variants of x 2 variants of x 3 should be same then only there is a meaningful comparison otherwise the variance of the error should be the same then only there is a meaningful comparison.

And the value of e is independent another important there should not be any pattern in the error term sometime what will happen when you plot the error term sometimes there is an increasing trend sometimes there may be in decreasing trend this kind of this kind of pattern is not allowed the error term has to be distributed randomly. And the another point is the error e is normally distributed random variable now testing for significance.

(Refer Slide Time: 14:03)

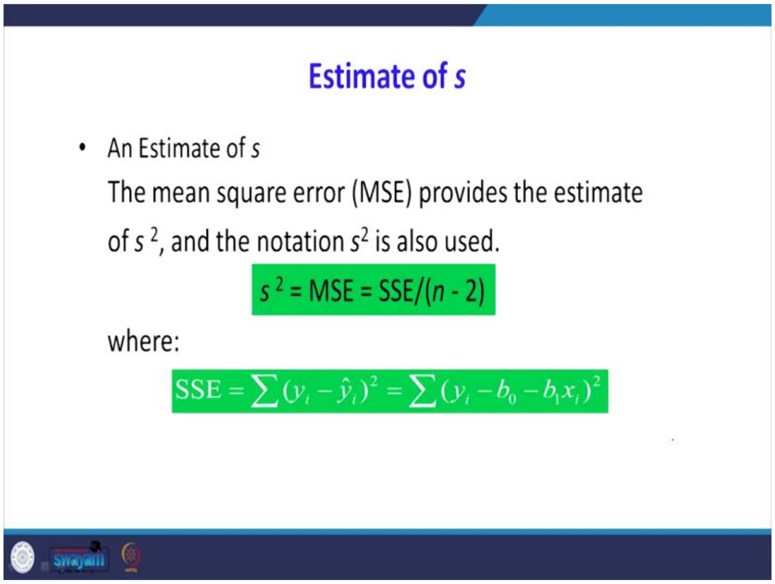


So far as I told you in the beginning of the class whatever regression equations and the goodness of model which you have tested only for the sample data what is the sample data y equal to b 0 + b 1 x so this capital y equal to beta 0 + beta 1 x whatever we have know done is only for the sample. Now we are going to see whether this model is valid even at the population level for that purpose we are going to do some assumption we will see what is that assumption that is a hypothesis?

To test for a significant regression relationship we must conduct a hypothesis test to determine whether the value of beta 1 is 0. What will happen if the beta 1 is 0 there is no relation between x and y at the population level. But there is a possibility that there may be a relation between x and y at the sample data it is not necessary that even at the population level there will be a relation between x and y. so, that testing can be done by two methods one is a t-test another one is F test.

Both t-test and F tests require an estimation of S square. S square is called the variance of the error otherwise if you say S it is the standard error the variance of e in the regression model.

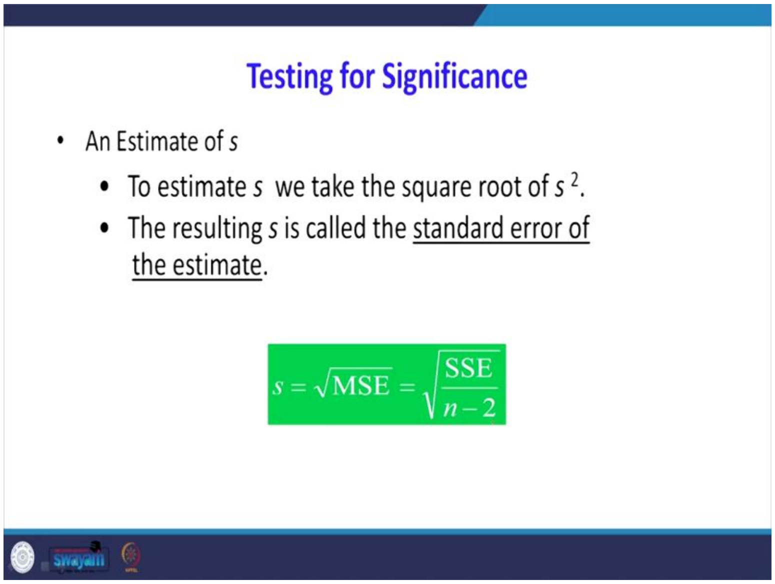
(Refer Slide Time: 15:22)



What is the estimation of the standard error suppose in years normal data set if there are two data set data set 1 and 2, 1 is having lesser variance than the other one so the first data it is more homogeneous in the same way when you do in regression model suppose there are two model is there model 1 model 2 for the model in which there is a lesser standard error that is a more suitable model. So, the mean square error provides the estimate of S Square and the notation S square is used so s square is nothing but MSE mean squared error.

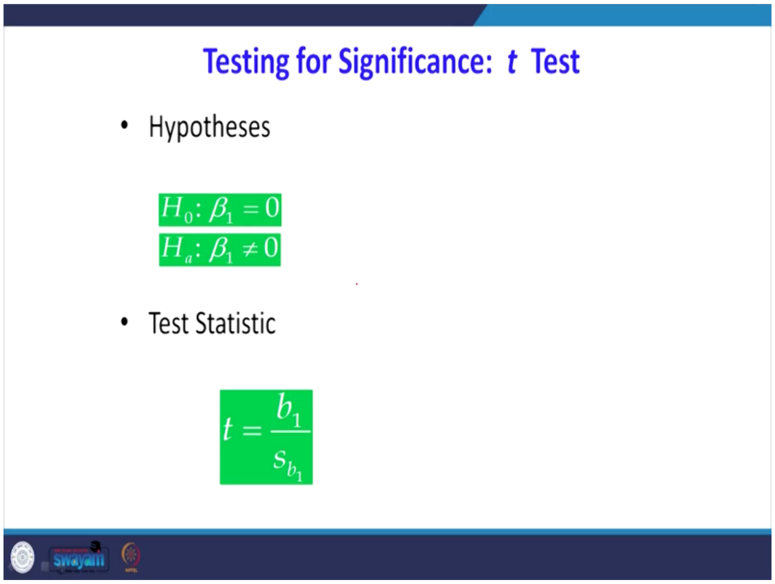
We know that how we got to MSE, MSE is SSE a divided by n - 2 here what the degrees of freedom n - 2 so the logic is n - 1 - K there is a logic of degrees of freedom. K is number of independent variable in this we are having only one independent variable we know that already the degrees of freedom is n – 1, so n - 1 - K will be n - 2 and SSE also you see that we can find out that formula which I in the beginning of the class which I am saying yi – y hat whole square. The y hat you can substitute b 0 actually it is a b 0 + b 1 x when you bring - inside b 0 - b 0 - b 1 x.

(Refer Slide Time: 16:51)



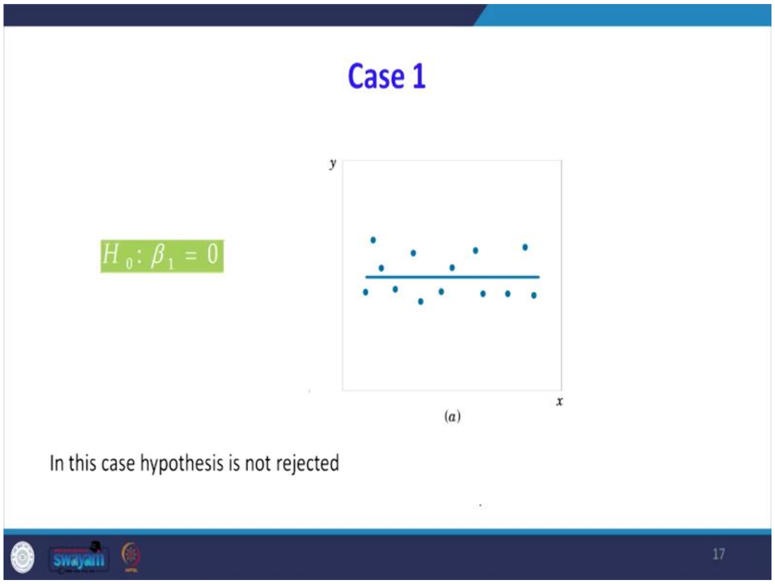
The term S we make the square root of s square the resulting is called standard at other term. So, when you take the square root of this mean squared error so you will get the standard error see in our problem I will go back I will see what is the standard error here where it is standard error as I told you by using shortcut method you can use SSE divided by n - 2 so that is MSE so S xy – S xy S xy whole square by S xx divided by n - 2 that is the standard error of the estimate. So you have to take the square root of that then you look at the standard error.

(Refer Slide Time: 17:37)



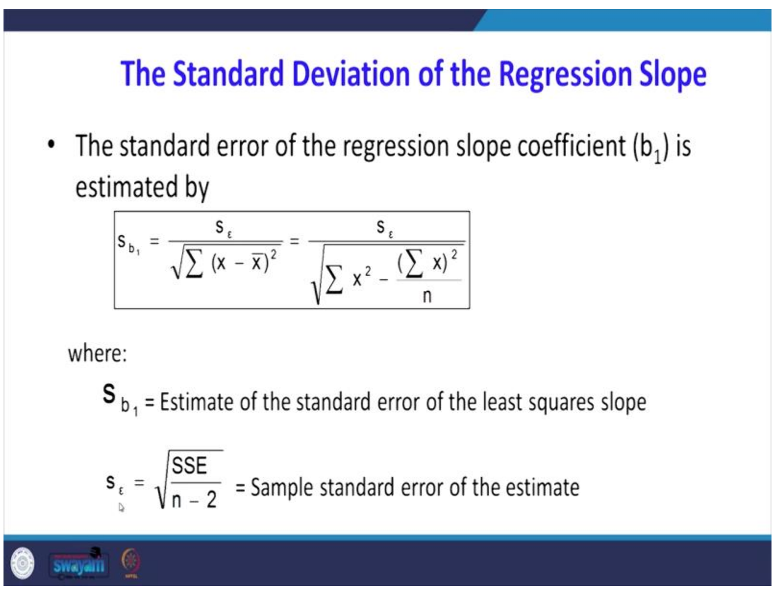
Now you go for hypothesis testing so what is the hypothesis testing beta 1 equal to 0 that means that there is no relation between x and y. Alternate hypothesis is beta 0 equal to 0 the test status is b 1 - beta 1 divided by Sb 1 Sb 1 is the standard error for the coefficient of b. So, since beta 1 we are assuming 0 it is simply b 1 divided by Sb1.

(Refer Slide Time: 18:06)



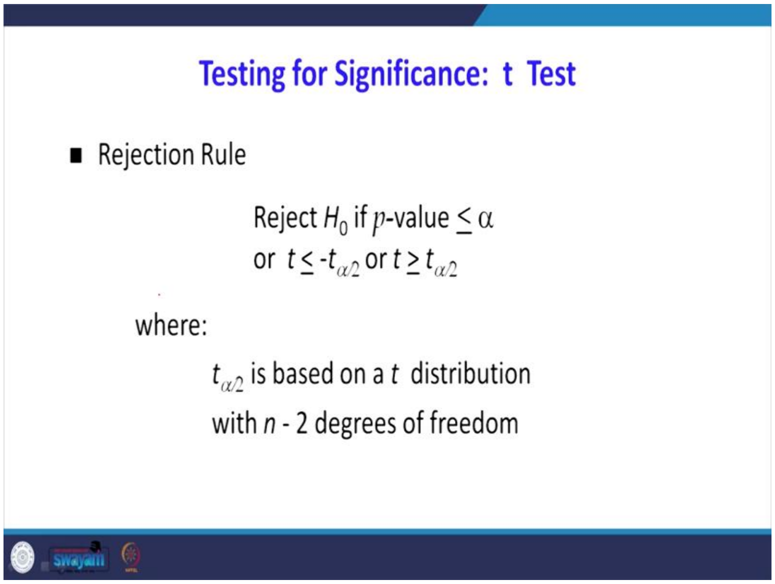
What is the meaning of b 1 beta 1 equal to 0 that means there is no relation between x and y. If it is when you plot the data in this case hypothesis is accepted because beta 1 equal to 0 now what is happening the beta 1 is not equal to 0 you see for this kind of data set so there is some relation between x and y see in this case the hypothesis is rejected so we are saying beta 1 not equal to 0.

(Refer Slide Time: 18:30)



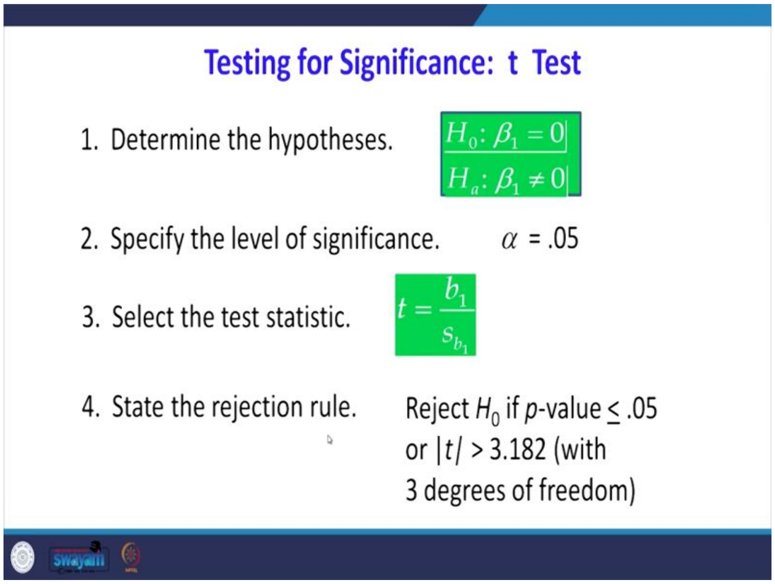
So, this is very important how to find out the standard error of the coefficient of x that is the b1, so Sb1 is y Se that is a standard error will divided by root of x - x bar whole square, you see intuitively the total error that is Se then we are dividing how much error is due to this independent variable, so total error did away portions of error from independent variable x. So, that will give you Sb1.

(Refer Slide Time: 19:04)



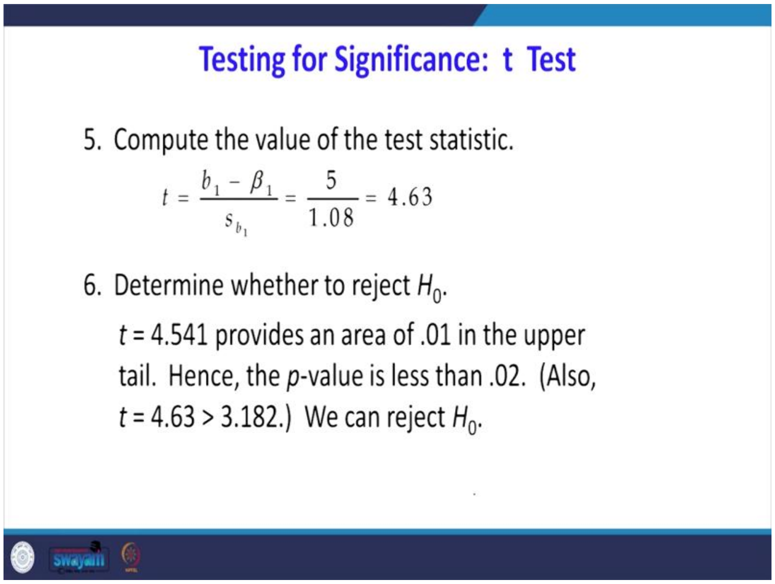
What is the rejection rule reject H 0 if the p-value is the less than or equal to alpha we have seen many times this one so what will happen this one if the p value, the p value is the see alpha the p-value is less than that you are to reject it otherwise accept it, where t alpha by 2 is the two-tailed test because we are writing beta equal to 0 Beta Beta 1 equal to 0 beta 0 equal to 0 and when you look at the t table you were to see n - 2 degrees of freedom.

(Refer Slide Time: 19:42)



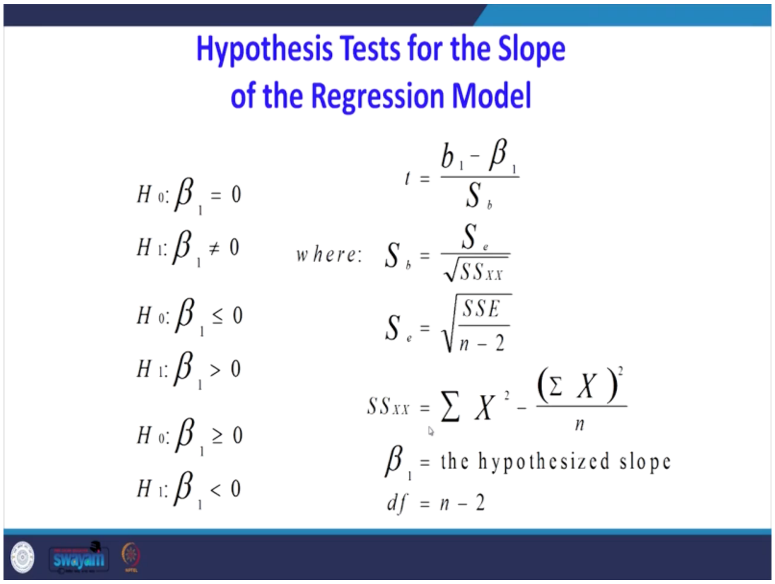
So, first determine the hypothesis beta 1 equal to 0 beta 1 not equal to 0 specify the significant level alpha equal to 5% select the test statistics b 1 by Sb1 state the rejection role reject H 0 if the p-value is less than or equal to 0.05 otherwise the t is greater than 3.182 when n - 2 degrees of freedom there are 5 data so 5 - 2 is the 3 degrees of freedom.

(Refer Slide Time: 20:16)



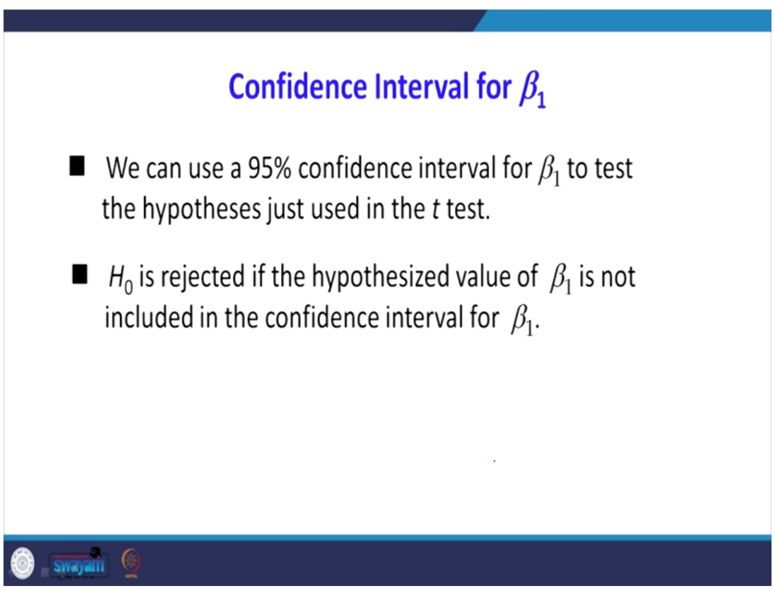
So, this was the compute the value of the test statistics so 5 data by the Sb 1 is 1.08 we get 4.63 determine whether to reject H 0 because t equal to 4.5 form provides in the area of 0.01 in the upper tail here is the p-value is less then we will see how the got the p-value less than 0.02 so P is greater than 4.63 we can reject null hypothesis and we reject null hypothesis what we are conclude there is a relation between x and y.

(Refer Slide Time: 20:51)



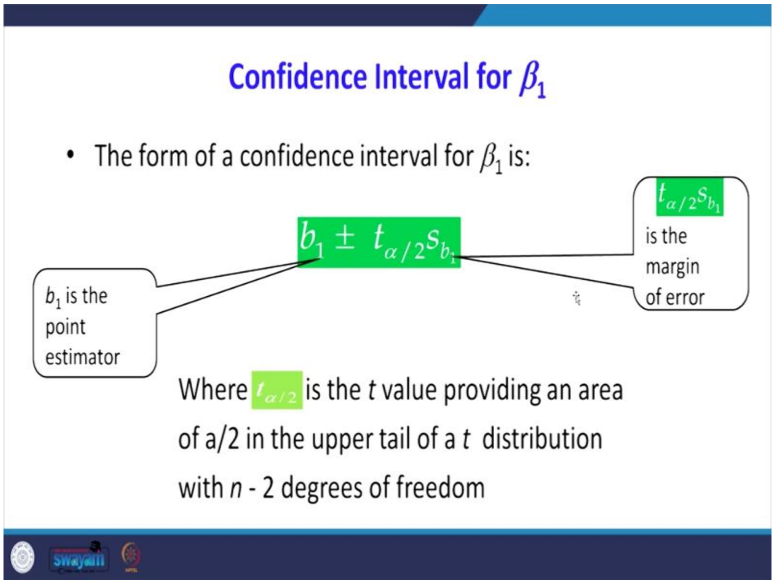
This was the here also beta 1 equal to 0 not equal to 0 2 tile test this is the right tailed test this is a left tailed test this was the formula for finding tb 1 - beta 100 is b to find Sb root S divided by Se is stranded at a root of SSX that is this is a form of SSE divided by n - 2 SSx is Sigma of x - x bar whole square remember it is n - 2 degrees of freedom.

(Refer Slide Time: 21:19)



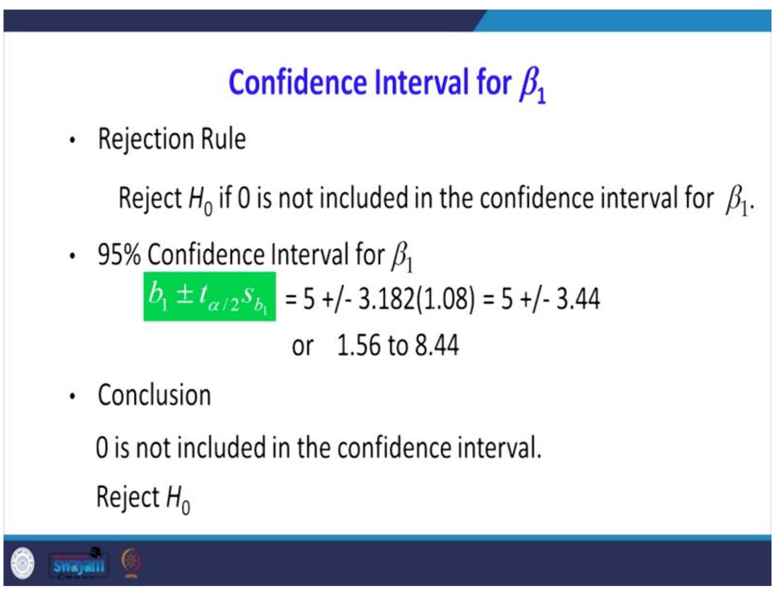
Next we can use the 95% confidence interval for beta 1 to test the hypothesis just used in the test. So, now with the help of conference interval also we can decide whether null hypothesis should be accepted or rejected it is not as rejected if the hypothesis value of b 1 is not included in the conference interval. our b 1 value what we are assuming to 0 so in that confidence interval if the 0 is appearing we have to accept the null hypothesis otherwise we have to reject null hypothesis.

(Refer Slide Time: 21:55)



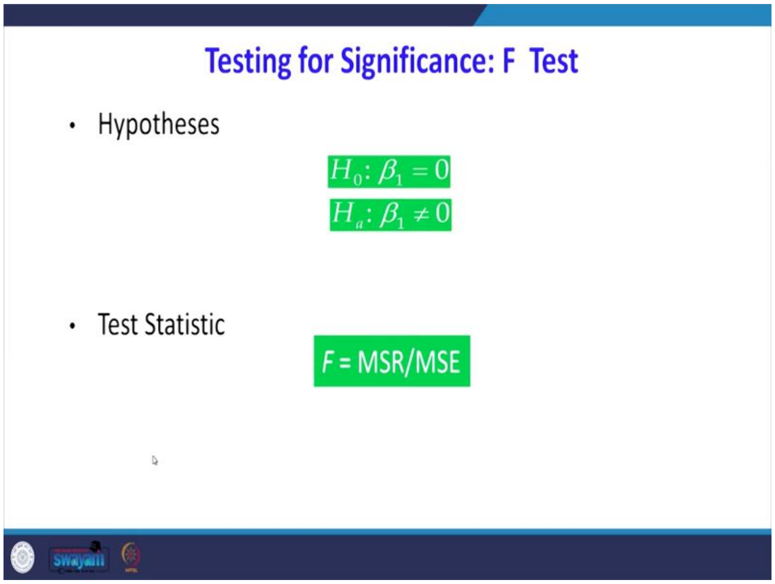
So, confidence interval for beta 1 is the form of controlled b 1 + or - t alpha by 2 is b 1 b 1 which you got from our regression equation that is a coefficient of x 1 Sb 1 previously we are getting out I told you what is the formula for getting is b 1, so when you substitute it here.

(Refer Slide Time: 22:22)



So, what is getting b 1 is 5 + or - t alpha by 2 is 3 point 1 8 - yes b 1 is 1 point 0 8 so 5 + or – 3.44 so lower limit is 1.56 upper limit is 8.44 you see in that there is no 0’s there so we have to reject our null hypothesis, conclusion 0 is not included in the conference interval so we are rejecting null hypothesis.

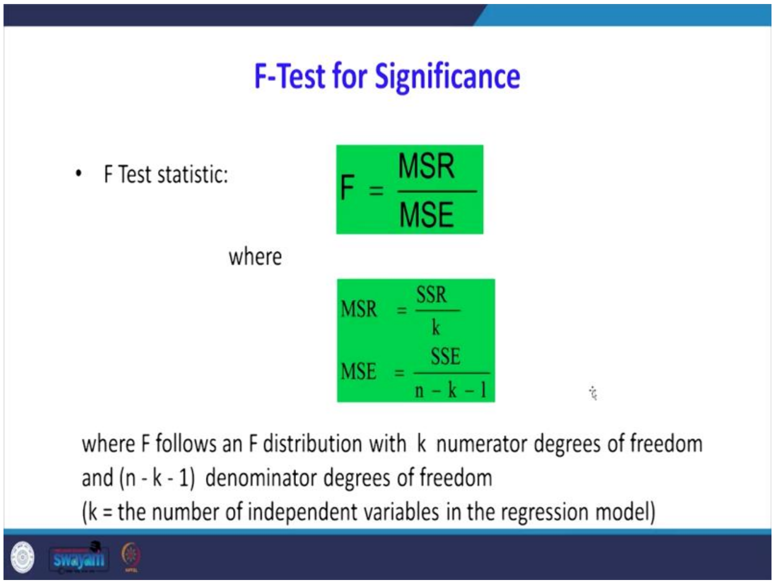
(Refer Slide Time: 22:43)



The previous way we have used the t-test some time what will happen if the number of independent variable is more than 2 we have to do the t-test to 2 times. If there are say 5 independent variable you have to do file individually as I told you whenever you are comparing more than two we should go for Anova that is the F-test so here also whenever there is a number of independent variables more otherwise a generic method for testing the beta 1 equal to 0 hypothesis is going for F test.

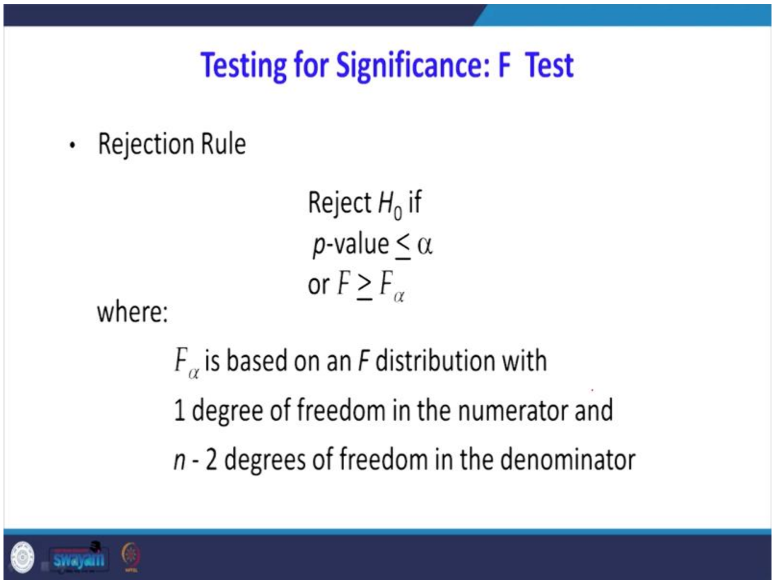
So here you have F test is a MSR divided by MSE even in Anova also you know anova what we write is we write MS treatment divided by MSE MS treatment is nothing but our regression sum of square mean regression sum of square.

(Refer Slide Time: 23:43)



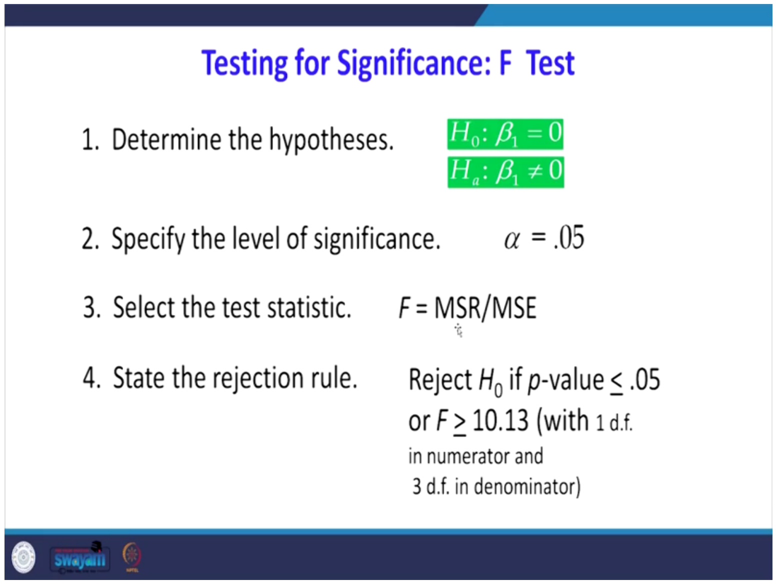
So, F equal to MSR divided by MSE you see that MSR how we are getting MSR SSR divided by K and K is number of degrees of freedom that is nothing but number of independent variable n - K - 1 is degrees of freedom for the error term.

(Refer Slide Time: 24:02)



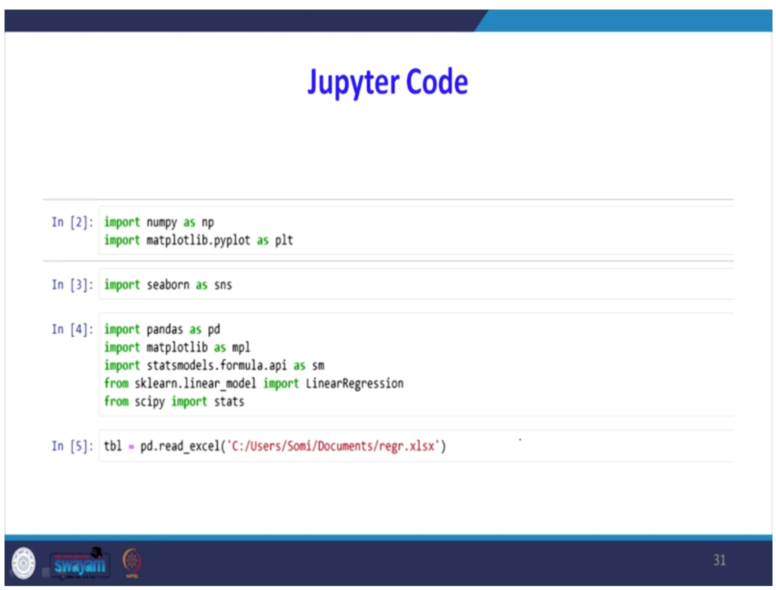
So, what is the rejection rule reject H 0 if the p-value is less than equal to alpha otherwise if the calculated F value is greater than the value each code from the table. So, if alpha is based on the F distribution we have to look at the what is the degrees of freedom as you look at you see enumerated degrees of freedom in this problem we have only one independent variables so one degrees of freedom numerator so n - 2 is 5 minutes to three degrees of freedom for denominator.

(Refer Slide Time: 24:33)



So, beta 1 equal to 0 beta 0 equal to 0 alpha equal to 0.05 F equal to MSR divided by MSE, so p value we got to find out so numerator degrees of freedom is 1 t nominated is a freedom is 3 so that p value is 10.13.

(Refer Slide Time: 24:51)

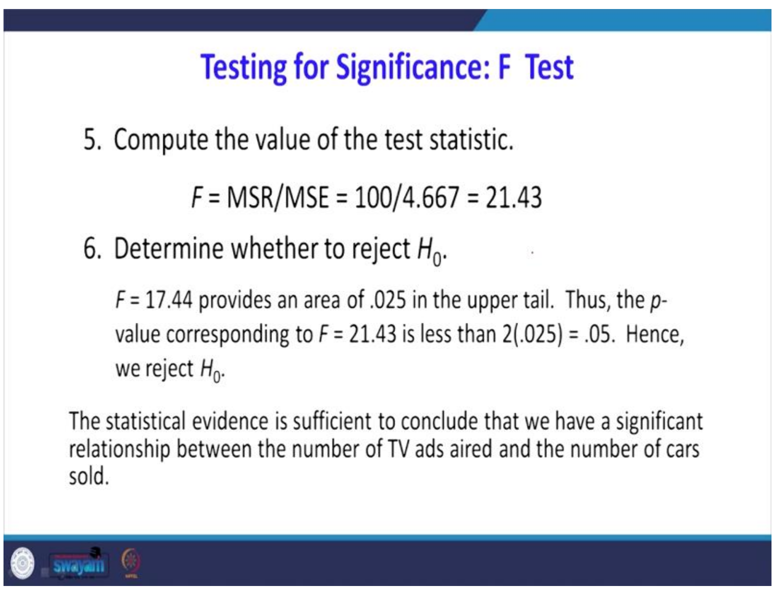


So, now we will use Python code rules that will do that import numpy as np import matplotlib ler dot pi plot recipe LT import seaborn ssn import pandas as pd input matplotlib as mpl import stats model stat formula api ssome from sklearn dot linear underscore model import linear regression from scipy import stats. So, tb1 we are going to that regression data we are going to save in the object called tb1 and we are reading that.

Now what is happening you see that the p-value for the TV ad we say alpha equal to 5% it is less than 0.01 so TV ad is insignificant variable if it is more than 5% see if it is a point 0 six the regression equation we will not include this independent variable TV ads you have statistics 21.43 I will go back will verify this answer that we can find out MSR, how we do MSR and that is what I am saying that time the first you have to find out SSR regression sum of square regression sum of square is y hat - y bar whole square Sigma divided by k, k is number of independent variable will get MSR.

So the p-value sorry the F value is 21.43 and see the probability it is less than 0.01 so that is less than 0.05 so we are saying that the model as a whole there are two things is there as a whole model the F value is less than 0.05 the model is valid and if you want to check individual independent variable also. So, see here it is less than 0.05 so this variable is significant see the lower limit upper limit there is no 0 here 1.563, 8.43. So we can we cannot accept where to reject null hypothesis.

(Refer Slide Time: 27:02)



You see MSR is 21.43 we are here 21.43 so we can verify that our Python result then what we have done it with the help of manually. Some cautions about the interpretation of significant test is so it is very important this one rejecting H 0 b1 equal to Z beta 1 equal to 0 and concluding that the relationship between x and y significant does not enable us to conclude that there is a cause and effect relationship in present between x and y. Just because of there is a correlation we cannot say there is a cause-and-effect relationship.

So, just because of we are able to reject H 0 beta 1 equal to 0 and demonstrate statistical significant does not enable us to conclude that there is a linear relation between x and y. Dear students in this class what we have seen we have taken one sample problem then we have fitted a regression equation in the regression equation we gone for hypothesis testing we have tested the significance of that independent variables.

There are two way to do the significance test one is by using team method t statistic method another one is F test method in both method we all got the same answer. Then I have explained what is the meaning of coefficient of determination that is r square from the r square I have explained how we can get the r. The next class will go for multiple regression equation where we will consider more than one independent variable and we will also ill explain some important assumptions in the regression equations. Thank you very much