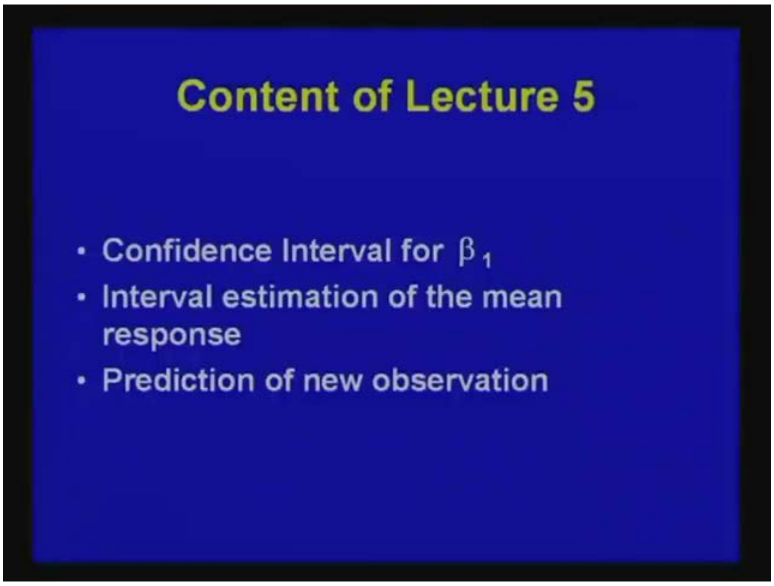
Regression Analysis
Prof. Soumen Maity
Department of Mathematics
Indian Institute of Technology, Kharagpur

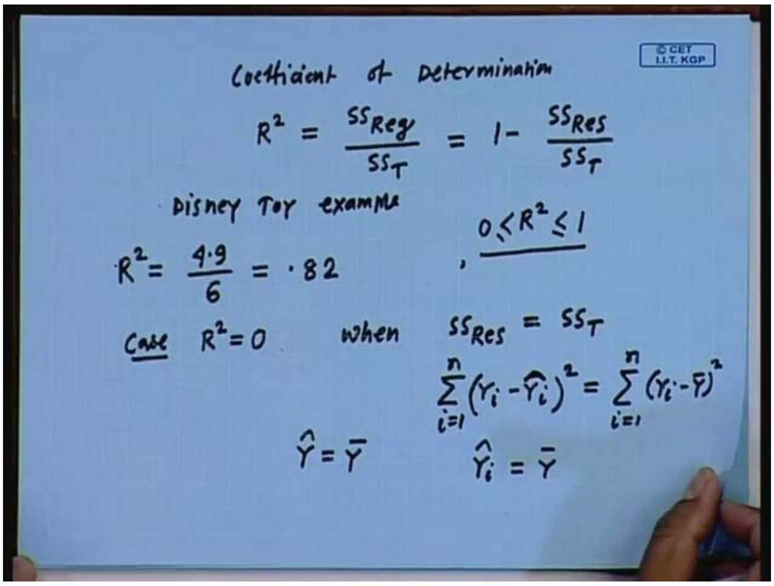
Lecture - 5
Simple Linear Regression (Contd.)

(Refer Slide Time: 00:57)



This is my 5 th lecture on Simple Linear Regression. Content of today's lecture is, now we will be talking about confidence interval for the regression coefficient beta 1. And then we will be talking about interval estimation of the mean response, and the finally will be talking about prediction of new observations for a given value of a regressor variable X equal to x naught. So, before I start talking on interval estimation for beta 1, I want to just recall the important parameter I talked about in the last class, that coefficient of determination that is R square.

(Refer Slide Time: 02:00)

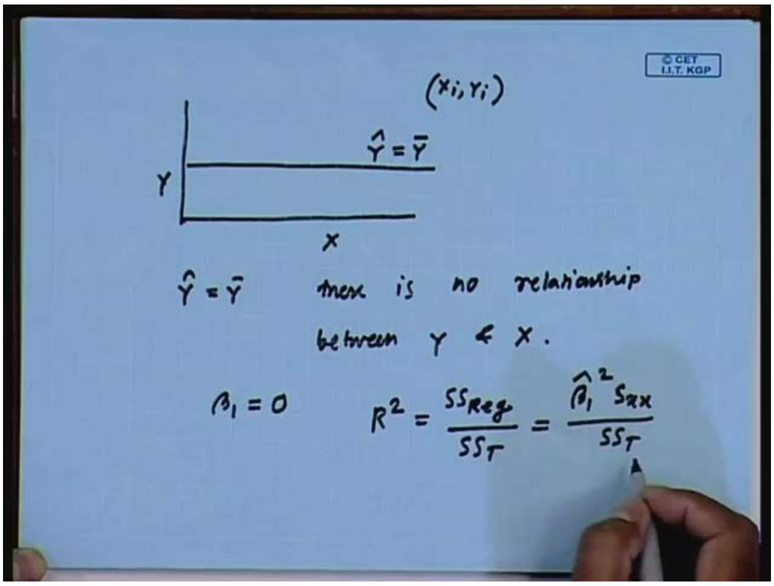


Coefficient of determination, which is a denoted by R square, R square is equal to SS regression by SS T. So, this one is basically, this R square, it measures the proportion of variability in the data or in the response variable, that is explained by the model or that is explained by the regressor variable. For example, if we consider the Disney toy example, there R square is equal to, SS regression was a 4.9 and SS T is 6, so this is equal to 0.82.

So, the meaning of this one is that, 82 percent of the total variability in the response variable or the total variability in sales amount is explained by the amount of money spent on advertisement. We know that, the range for R square is from 0 to 1, we discussed when R square is equal to 1. Let me consider the case R square is equal to 0, so R square is equal to 0, so R square can also be written as 1 minus SS residual by SS T. So, this quantity is going to be equal to 0, if SS residual by SS T, if this ratio value is equal to 1 that means, R square is equal to 0 when SS residual is equal to SS T.

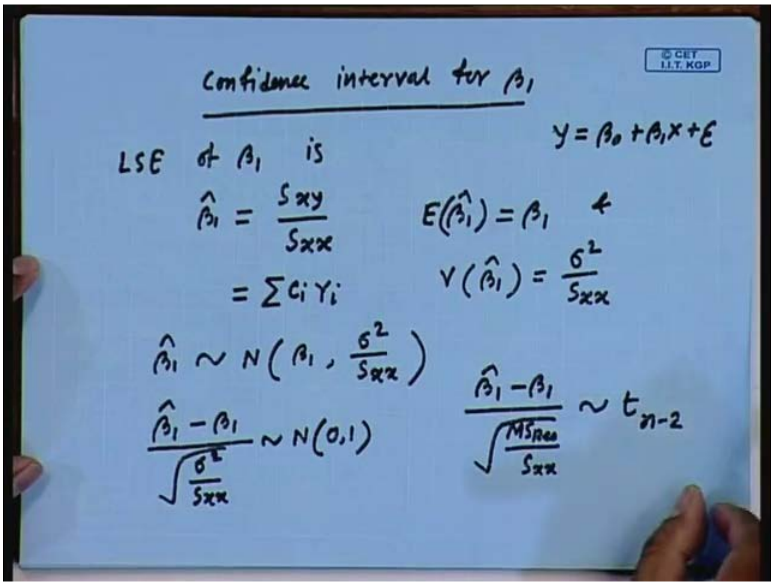
So, what is SS residual, SS residual is summation Y i minus Y i hat square, which is equal to summation Y i minus Y bar square, i is from 1 to n. So, this to quantity are equal when Y i hat is equal to Y bar, so basically if the fitted model is Y hat equal to Y bar then R square value also is equal to 0. That means, this fitted model, it does not depend on the regressor variable.

(Refer Slide Time: 07:00)



So, the situation I mean, suppose we have some data and we have a set of observations X i Y i and the fitted model is this one, which is Y hat equal to Y bar. That means, the significance of this one is that, X Y this Y is that, there is no relationship between the response variable and the regression variable. I mean, this will happen, Y hat equal to Y bar will happen when there is no relationship between Y and X. Also in other way we can think about I mean, this will happen when beta 1 is equal to 0. So, these also says I mean, R square is equal to SS regression by SS T and we know that, SS regression is beta 1hat square S xx by SS T. So, this will be equal to 0 when beta 1 is equal to 0, so next we move to further confidence interval for beta 1.

(Refer Slide Time: 09:14)

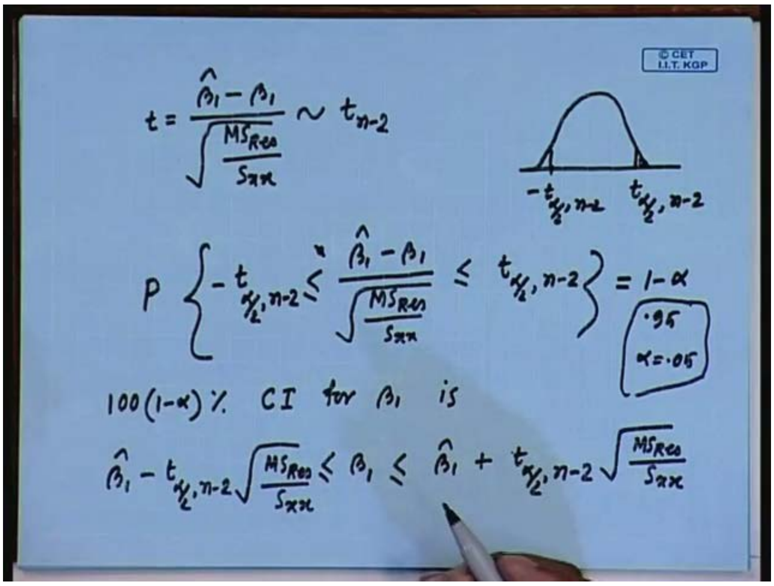


So, our model is y equal to beta naught plus beta 1 X plus epsilon, so this is called the intercept and this is beta 1 is basically the slope. And we know that, the least square estimator of beta 1 is beta 1 hat, which is equal to S xy by S xx. So, this estimator is the result of least square estimate and this is in fact, more precise, this is called the point estimation of beta 1. So, the concept of ah interval estimation is that, instead of giving a point estimate of some population parameter, the interval estimation gives an interval such that, the probability that the population parameter will lie in that interval with high probability that means, may be with the probability 0.95 or 0.99. So, the technique to get the interval estimation is that, first you find the point estimate or you find the point estimator of the population parameter and then you find the sampling distribution of the point estimator.

So, here the beta 1 is equal to S xy by S xx, we need to find the sampling distribution of beta 1 hat. We know that, beta 1 hat is an unbiased estimator of beta 1, so this is that means, beta 1 hat is equal to beta 1 and also we know that, the variance of beta 1 hat is equal to sigma square by S xx. And also we have proved in previous lecture that, this beta 1 hat, this one is basically it is a linear combination of random variables Y i.

And we have assumed that, Y i’s are normally distributed, so any linear combination of normal variable also follows normal distribution. So, beta 1 hat we know, beta 1 hat follows normal distribution with mean beta 1 and variance sigma square by S xx. From here, we can say that, beta 1 hat minus beta 1 by sigma square S xx root of this, this follows normal (0,1), but the situation is that, most of the cases, sigma square is not known then we replace sigma square by it is unbiased estimator that is, MS residual. So, if we replace MS residual, if we replace sigma square by MS residual then this random variable beta 1 hat minus beta 1 by MS residual by S xx, this follows t distribution with degree of freedom n minus 2.

(Refer Slide Time: 14:38)

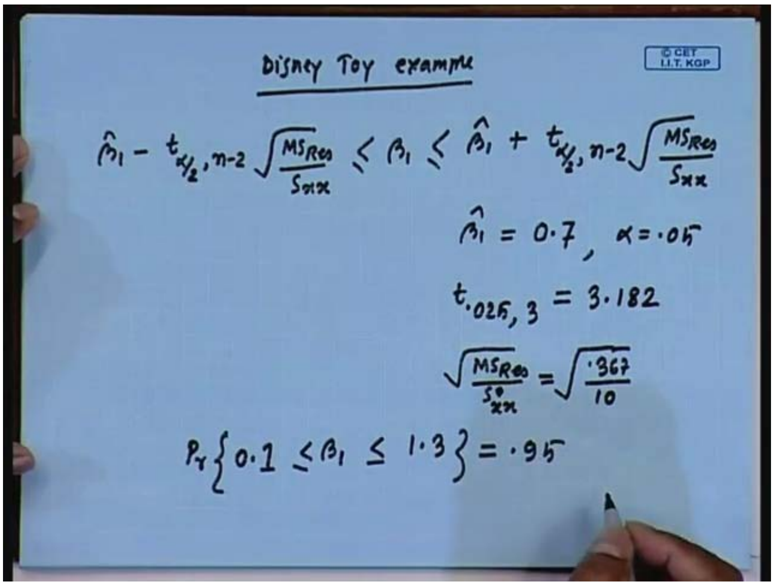


So, what we got is that, we got that beta 1 hat minus beta 1 by MS residual by S xx, this follows t distribution degree of freedom n minus 2. And let me call this equal to t, now we need to have a confidence interval for beta 1, suppose this is t distribution, now we will take two points, this point is t alpha by 2, N minus 2. The meaning of this one is that, t greater than this one and the probability of this area I mean, this portion is alpha by 2, so from here and this point is minus t alpha by 2, n minus 2.

Now, beta 1 hat minus beta 1 by MS residual by S xx, we can say that, this t, this is basically the t, so t is in this interval t alpha by 2, n minus 2 greater than t alpha by 2, n minus 2 with probability 1 minus alpha. So, to make this probability high, we have to choose alpha accordingly. For example, if we want to make this probability say 0.95 then we have to choose alpha equal to 0.05. So, from here, we get the confidence interval for beta 1, so we say that, 100 into 1 minus alpha percent.

That means, if we choose alpha equal to 0.05, this quantity is going to be 95 percent confidence interval for beta 1 is obtained from here, just simple algebra, you write beta. So, the range of beta 1 is beta 1 hat plus t alpha by 2, n minus 2 root to over MS residual by S xx and the lower bound is beta 1 hat minus t alpha by 2, n minus 2 MS residual by S xx, so this one is basically 95 percent confident interval for beta 1. In other word, we can say that, the population parameter beta 1, which is basically the slope for the simple linear regression model. This will lie in this interval with probability 0.95, let me explain this one in the toy example.

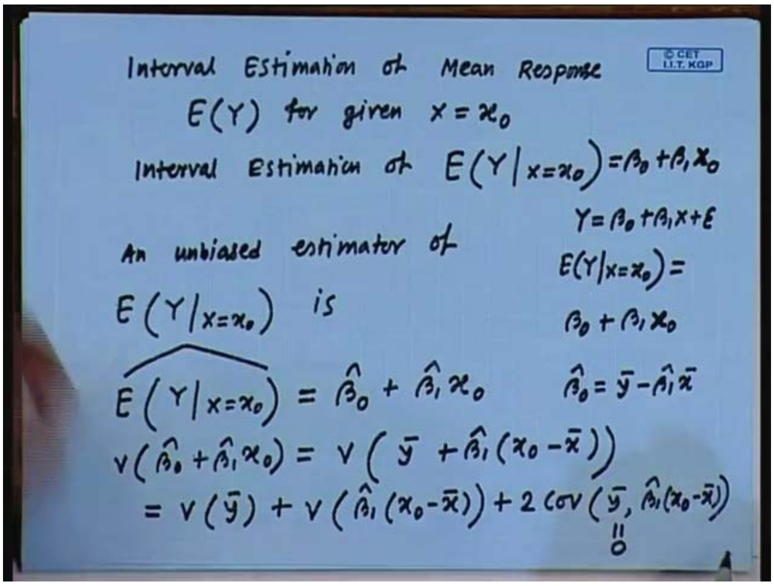
(Refer Slide Time: 18:46)



In the Disney toy example, so what we got is that, the upper bound was beta 1 hat plus t alpha by 2, n minus 2 root over MS residual by S xx, this is the upper bound for beta 1 and the lower bound is beta 1 hat minus t alpha by 2, n minus 2 MS residual by S xx. So, for the Disney toy example, beta 1 hat is equal to 0.7 and there we have 5 data points, so will choose alpha equal to 0.05 to make probability 0.95. So, t 0.025, 3 is, you see the value of this one from the statistical table, this one is equal to 3.182.

So, we only need to compute this quantity root over of MS residual by S xx, which is equal to 0.367, S xx is 10. And it is not difficult to now check that, beta 1 will lie in the interval 1.3 to 0.1 and beta 1 will lie in the interval 0.1 to 1.3 with probability, this probability is equal to 0.95. So, this is what the interval estimation is, instead of giving one estimate of a population parameter, here we give an interval and the use of this interval is that, we can say that the population parameter will lie in this interval with high probability that is, 0.95.

(Refer Slide Time: 22:19)



So, next we move to interval estimation of mean response that is, E Y mean response or expected response for given X equal to x naught. Once you have the fitted model, one important application of a regression model is to estimate the expected response for the given value of the regressor variable. And also the another important problem for the regression model. the another important application of the regression model is that, prediction of new observation corresponds to a given value of response variable, given value of a regression variable X.

So, first we will talk about the estimation of, in fact the interval estimation of the expected response or mean response at sum for a given value of the regressor variable X. So, here, we want to find interval estimation of mean response, that is or expected response for X equal to x naught. So, this looks like conditional expectation, but basically what I want to mean by this notation is that, I want estimate the expected response value for given X equal to x naught.

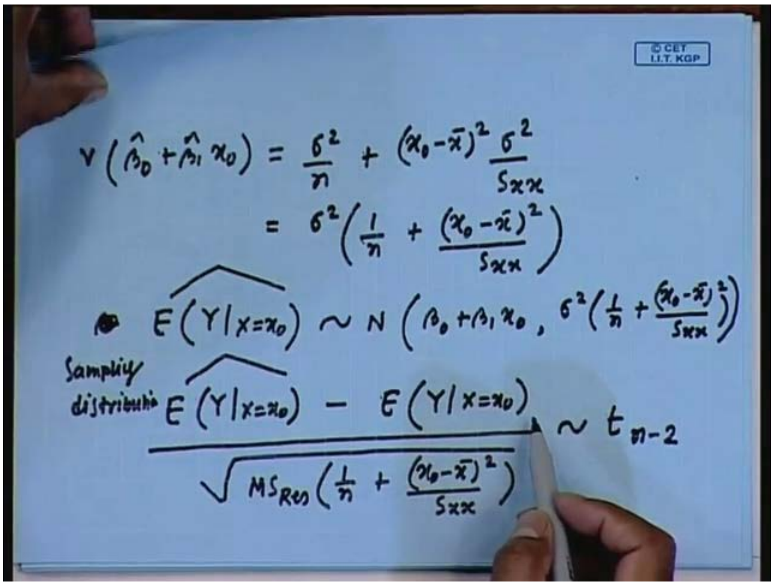
That means, at x naught point of the regressor variable, I want to find the expected response. If you recall the model, simple linear regression model, Y equal to beta naught plus beta 1 X plus epsilon, so the expected response Y at the point X equal to x naught, this quantity is equal to beta naught plus beta 1 x naught. So, we want to find an estimator of this quantity beta naught plus beta 1 X, not only I mean, we are not looking for the point estimation of this expected response, we are looking for an interval estimation of this expected response at the point X equal to x naught.

So, again we have to start from the point estimation of I mean, the estimator of this expected response. We know that, an unbiased estimator of this expected response Y given X equal to x naught is, let me denote this estimator by this expected response hat equal to beta naught hat plus beta 1 hat x naught. So, I should put x naught here, you want to find interval estimation of this expected response at the point X equal to x naught.

So, this is an unbiased estimator of the expected response, it is very easy to prove that, this an unbiased estimator. Because, both beta naught and beta 1, they are unbiased estimator of beta naught and the beta 1 respectively. Now, we need to find the sampling distribution of this quantity or this random variable I should say, to get that, I need to find the variance of this estimator. Variance of beta naught hat plus beta 1 hat x naught is equal to is equal to variance of y bar plus beta 1 hat x naught minus x bar.

What I did here is that, I just have replaced, we know that beta 1 hat is equal to y bar minus beta 1 hat x bar, so I have replaced beta naught hat by this quantity. Now, variance of this one is equal to variance of y bar plus variance of beta 1 hat x naught minus x bar plus twice covariance of y bar beta 1 hat x naught minus x bar. And it is not difficult to prove that, this quantity, this covariance term is equal to 0.

(Refer Slide Time: 30:27)

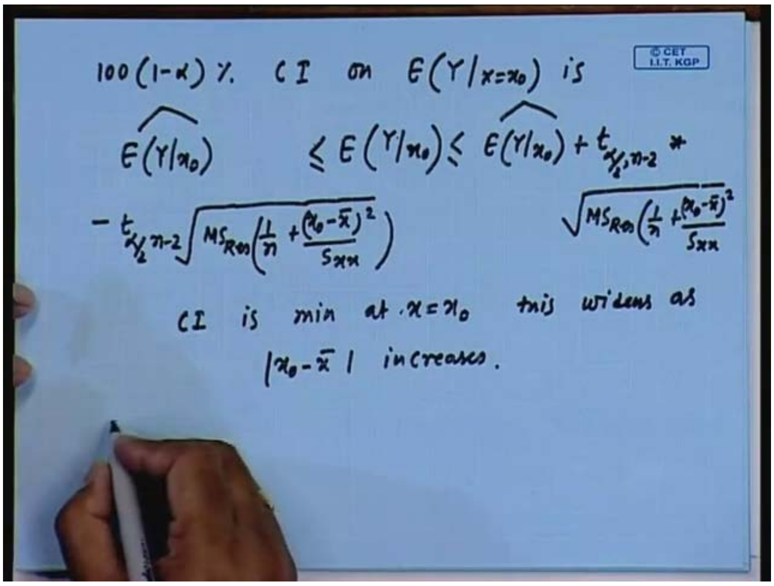


Now, what I want to do is that, I write down the variance of beta naught hat plus beta 1 hat x naught is equal to, basically variance of y bar, which is sigma square by n plus variance of this quantity, which is x naught minus x bar whole square into sigma square by S xx. So finally, the variance of this quantity is equal to sigma square into 1 by n plus x naught minus x bar whole square by S xx. So, again the same argument, beta naught hat is a linear combination of Y i's, beta 1 hat is also a linear combination of Y i's.

So, since the Y i's follows normal distribution, we can say that, beta naught hat plus beta 1 hat x naught, which is linear combination of random variable, that also follows normal distribution. So, beta naught hat or we can say that, the estimator of Y given X equal to x naught, this estimator follows normal distribution with a mean beta naught plus beta 1 x naught hence, variance sigma square by 1 by n plus x naught minus x bar whole square by S xx.

And from here, now sampling, see sigma square is not known, so we replace sigma square by MS residual. So, what we got finally is that, this estimator Y given X equal to x naught minus, this one is basically expected response at the point X equal to x naught by root of MS residual, I am just replacing sigma square by MS residual, into 1 by n plus x naught minus x bar by S xx. This follows t distribution with degree of freedom n minus 2. So, we want to find the confidence interval for this expected response at the point X equal to x naught. Now, we have an estimator for this one and we have the sampling distribution, this is called the sampling distribution of this estimator. And from here, we get the 95 percent confidence interval for the expected response.

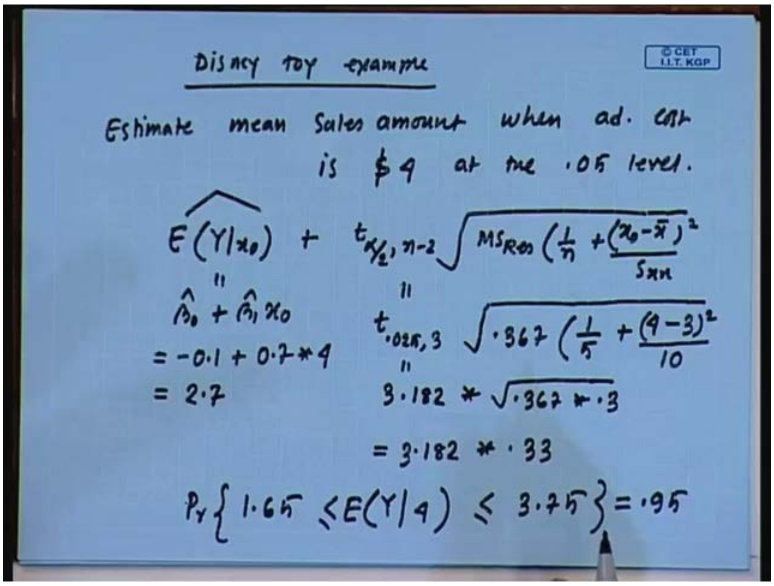
(Refer Slide Time: 35:18)



And that is given by, let it me write 100 into 1 minus alpha percent, confidence interval on expected response at the point X equal to x naught is E of Y given, I am writing just x naught. This is inbetween E of Y given x naught estimator plus t alpha by 2, n minus 2 and then that here you have root of this quantity, MS residual into 1 by n plus x naught minus x bar whole square by S xx. And similarly, the lower bound is E Y given x naught, this quantity minus t alpha by 2, n minus 2 MS residual 1 by n plus x naught minus x bar S xx.

So, this is the confidence interval for this one and this confidence interval is minimum at X equal to x naught. And this widens as x naught minus x bar, the absolute value of this one increases I mean, this looks bit abstract. Let me give one example for this one, again you consider the toy example, Disney toy example.

(Refer Slide Time: 38:30)



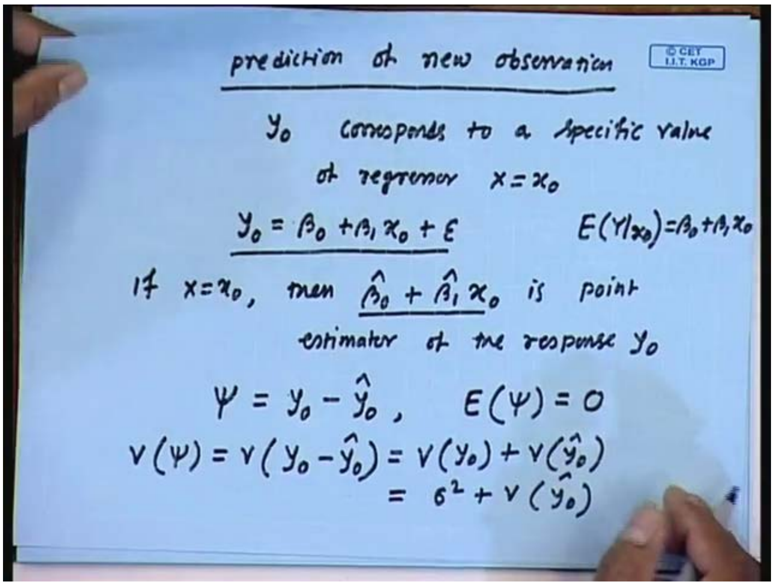
And here, what you do is that, estimate the mean sales amount when advertisement cost is say, 4 dollar at the 0.05 level. So, you find out all these things, let me just compute the upper bound for this one, the upper bound is E Y given x naught, estimator of this one plus t alpha by 2, n minus 2 into MS residual 1 by n plus x naught minus x bar whole square by S xx. You know this quantities is nothing but beta naught hat plus beta 1 hat x naught, so we know the value of this one, this one is basically minus 0.1 plus beta 1 hat is 0.7 and x naught is 4, so this one is equal to 2.7.

Now we know, that this quantity is, since in the Disney toy example n equal to 5, so this one is basically t 0.025, 3 which is equal to 3.182. Now, we need to compute this term, here what we had is that, MS residual is 0.367, n is 5 plus x naught is given 4 and you can check that, you go see the Disney toy data, you can check that, x bar is equal to 3. So, this is 4 minus 3 square by 10 and this will come out to be 0.367 into 0.3, which is equal to 0.3182 into 0.33.

So, the upper bound is going to be, for this quantity expected response at the 0.4 is going to be 3.75 and lower bound is obtained by just replacing this plus sign by minus, this will give you 2.7 minus this quantity, 3.182 into 0.33. So, this will be 1.65 and the probability that the expected response, when the cost on advertisement is equal to 4, the expected response will lie in this interval with probability 0.95. And you can go back to the, you can see the original data, there you will see that the X, the actual response value is equal to 2, corresponds to X equal to 4.

So, this is how we give confidence interval for some population parameter and here the population parameter is beta naught plus beta 1 x naught. And we have given 95 percent confidence interval for the population parameter, which is, here it is basically the expected response at some value of, for a given value X equal to x naught. So, another important application of this regression model is to predict the new observation. This one is bit little difficult, there is a slide difference between the expected response and what I am going to do now, this says that, let me the explain the thing.

(Refer Slide Time: 44:17)



We are going to predict new observation, prediction of new observation, what I want to do is that, I want to predict new observations say, y naught corresponds to a specific value of regressor X equal to x naught. The difference between the previous one and this one is that, in the previous problem, we expected response and here, we want to predict the observation at the point X equal to x naught.

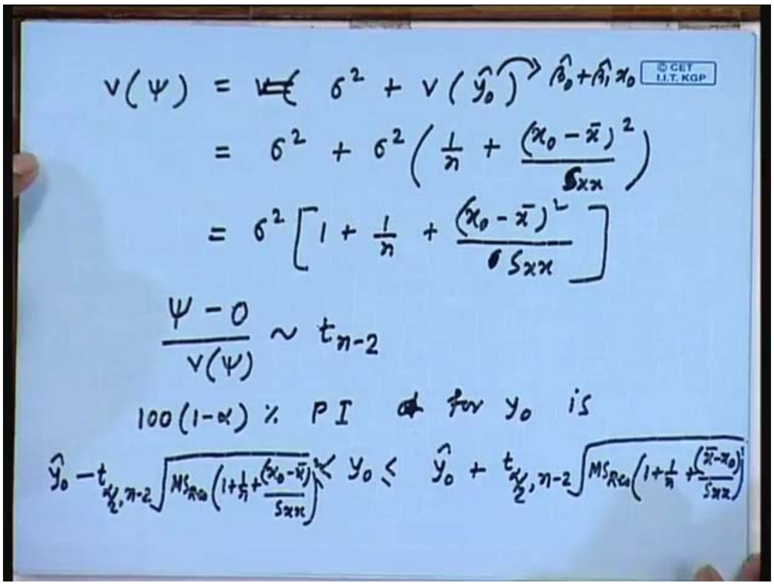
So, the difference between the previous and this one is that, y naught is nothing but it is a new observation, given the data we have fitted the model and now using that fitted model, we want to predict the response value at new point. So, you want to predict y naught, which is basically beta naught plus beta 1 x naught plus epsilon, you want to predict this one. And the previous problem was, we wanted to predict or we wanted to estimate expected response at X equal to x naught, which is equal to beta naught plus beta 1 x naught.

So, here we want to estimate y naught, which is equal to this quantity and in the previous example, we wanted to estimate expected this response, which is equal to this quantity. Now, again if X equal to x naught then beta naught hat plus beta 1 hat x naught is point estimator of the response, so we want to predict y naught. So, will start from this point estimator, now we define random variable psi, which is equal to, this is bit tricky, which is equal to y naught minus y naught hat, y naught hat is nothing but this quantity.

This is equal to, this is basically y naught hat, now you can check that, is not difficult to check that, expected value of this new random variable psi equal to 0. And the variance of this new random variable psi is equal to variance of y naught minus y naught hat, which is equal to the variance of, see y naught hat, this y naught hat is this quantity, beta naught hat plus beta 1 hat x naught. So, the whole thing is a function of y 1, y 2, y n, the given observation, but y naught is a new observation and this one is independent of y 1, y 2, y n.

So, y naught hat basically involves y 1, y 2, y n and y naught is a independent observation. So, that is why you can write, the variance of this quantities is equal to variance of y naught plus variance of y naught hat. Now, the variance of y naught hat we know, because just now you had computed the variance of this quantity, it is not difficult to check and variance of y naught is equal to sigma square, so variance of y naught hat.

(Refer Slide Time: 50:00)



So, variance of psi is going to be variance of sigma square plus variance of y naught hat, which is equal to sigma square plus sigma square 1 by n, this one we just we have proved that, this is equal to x naught minus x bar whole square by sigma xx. Because, y naught hat is nothing but this one is nothing but beta naught hat plus beta 1 hat x naught. So, this is sigma square into 1 plus 1 by n plus x naught minus x bar whole square by sigma, this is S xx, this is S xx.

And we know that, the sampling distribution of psi, minus expectation of psi is equal to 0, by variance of psi, this follows normal distribution. Now, if you replace in the variance of psi, this sigma square by MS residual then this is going to follow t distribution with degree of freedom n minus 2. So, from here, we get 100 into 1 minus alpha percent, we call it prediction interval.

Prediction interval for y naught is y naught lies between y naught hat plus t alpha by 2, n minus 2 into whole thing MS residual into 1 plus 1 by n plus x bar minus x naught whole square by S xx, so this one is the upper bound for y naught. And the lower bound is y naught hat, this plus you just replaced by minus. So, this is t alpha by 2, n minus 2 into MS residual 1 plus 1 by n and this quantity, plus x naught minus x bar whole square by S xx. I hope you can understand, basically this quantity is nothing but this one, only the sigma square has been replaced by MS residual.

So, this is the 95 percent, if you put alpha equal to 0.05 then the probability that the future observation at X equal to x naught will lie in this interval with probability 1 minus alpha that is, basically 0.95. And here, of course, this interval is minimum at the point X equal to x naught and this interval is always wider than the interval given for the expected response at the point X equal to x naught. So, this is all for today and this is perhaps my last class on simple linear regression and in the next lecture, we will be talking about multiple linear regression.

Thank you.