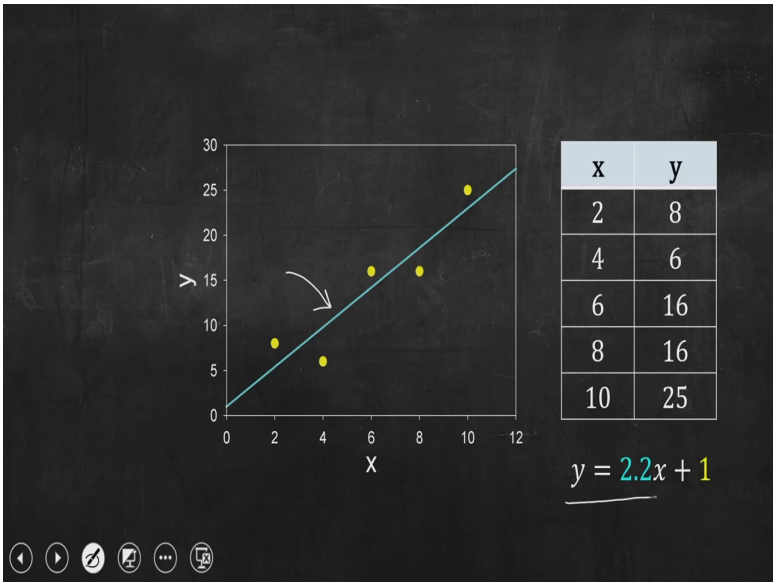
Data Analysis for Biologists
Professor Biplab Bose
Department of Biosciences & Bioengineering
Mehta Family School of Data Science & Artificial Intelligence
Indian Institute of Technology, Guwahati
Lecture 28
Linear Regression - II

(Refer Slide Time: 00:51)

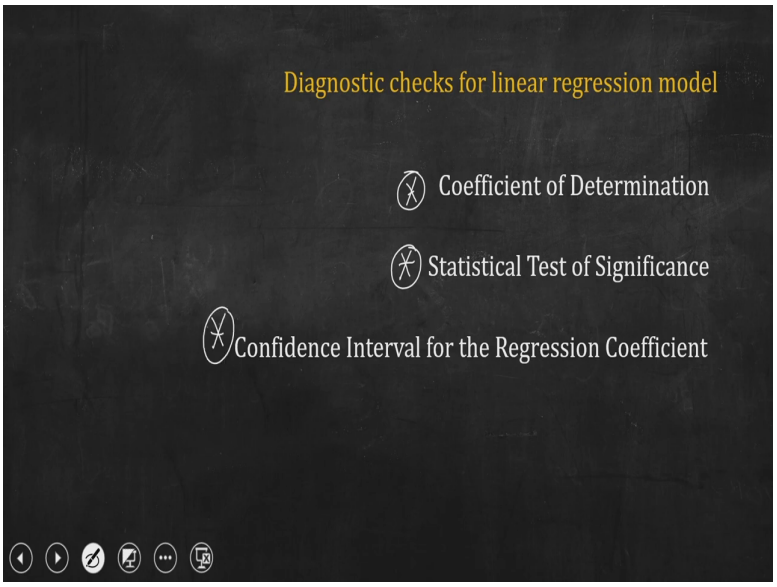


x	y
2	8
4	6
6	16
8	16
10	25

$y = 2.2x + 1$

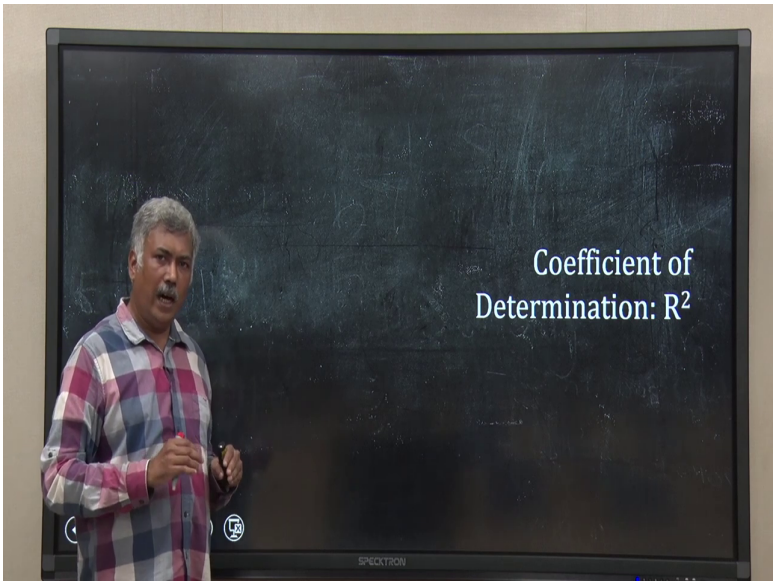
Hello everyone. Welcome to the second lecture on linear regression. In the last lecture, we have learnt how to perform linear regression and the problem statement we have discussed and I performed a linear regression on a small data set also. So, let us start with that, we have a data set and I have fitted that data by linear regression Least-square method and I got a linear relation y equal to 2.2 into x plus 1. But our business of linear regression or building a linear model out of this data does not end just with this equation. Every time you perform a regression and create a model out of regression, you have to do some diagnostic checks. There can be many types of diagnostic checks and they have a purpose, there are different purposes.

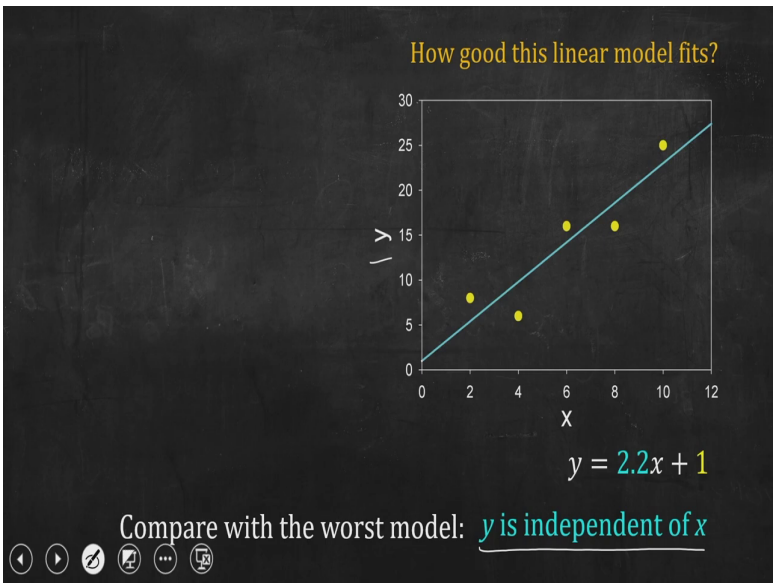
(Refer Slide Time: 1:26)



In this particular case, I will discuss three specific diagnostic tests which are must, every time you do linear regression you should check do them. The first thing is you have to calculate coefficient of determination, R squared, second thing you should always do, you should do a statistical test of significance on the regression coefficient, y equal to mx plus c equation has two regression coefficient m and c, the slope and intercept. So, you have to perform statistical test of significance on those two. And the third one you must do, you should also check the confidence interval for those regression coefficient.

(Refer Slide Time: 2:07)





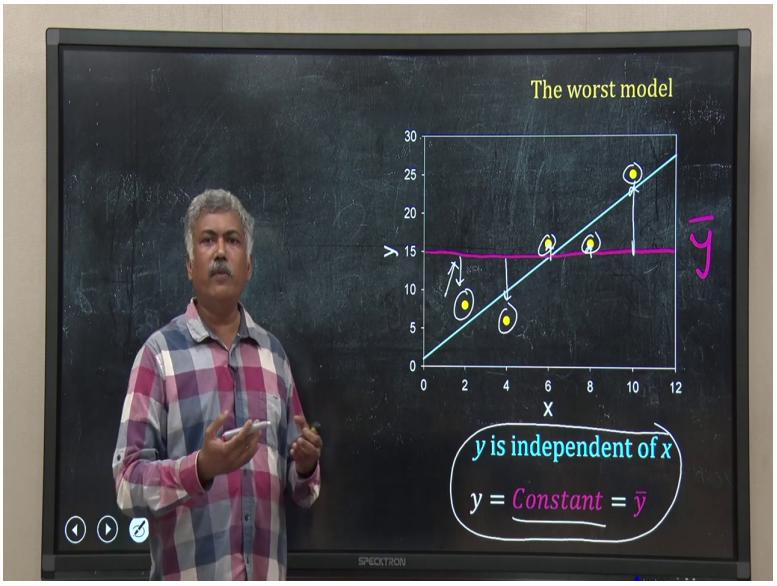
$$y = 2.2x + 1$$

I will discuss and explain each of this one by one. Let us start first with coefficient of determination, R square. So, what we have done in linear regression? We have five data point in this example and I have done used Least-squares method to find a straight line which we call best fit because it is has a minimum error.

Now, I am asking you a question here the, how good is this linear model. Now, remember when I say something is good or bad that is always relative in this case because we know we do not have a unique solution for this, we are just optimizing. So, when I say something is good, there must be something bad also and how good means how good it is with respect to that bad thing.

So, let me first imagine what could be the worst model for this data? This data 5 data point of x and y is given to you and what could be the worst model? The worst model could be that y does not depend upon x at all, that means y is a constant. It does not vary with x.

(Refer Slide Time: 3:13)



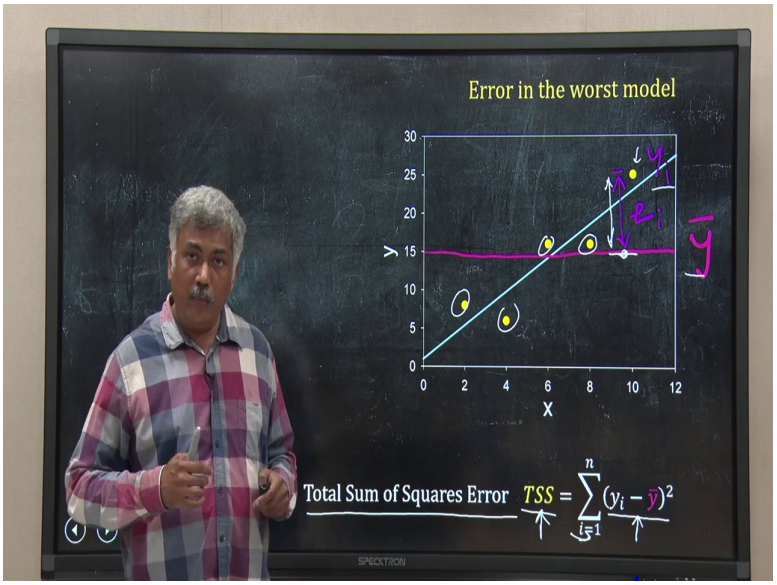
y = Constant = y^

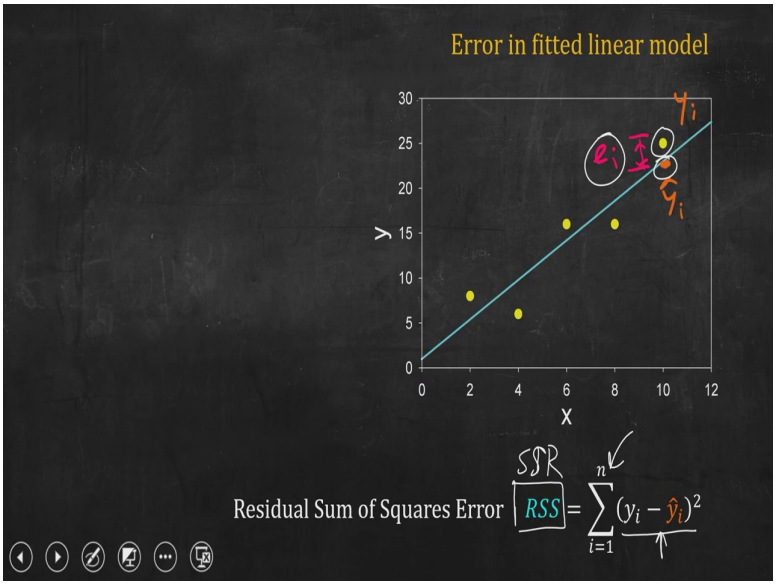
And if I visually see that, I should get a straight line, y equal to constant. Now, y equal to constant, you can imagine any value, y equal to 100, y equal to 0, y equal to minus 10 something like. Those will not be reasonable, the reasonable would be if I collect the mean of this value of y from this 5 data point which I have got from experiment and take that mean as the value of y as the constant.

Because then you can imagine that the real relationship between x and y is such that y is constant at that y bar and during experiment as my experiment has noise, observation has noise those values has got deviated slightly from that real value. This pink line. That makes this worst model reasonable, is it? So, whatever, it is the worst model because this is linearly independent, x and y is completely independent. This is called worst because my model, my belief is there is a linear relation between x and y.

So, now I want to compare these two, I want to compare this worst model, so called worst model with my good model, the y equal to mx plus c that I have fitted to my data. So, how can I do that? One-way, simple way I can think of is that, why do not I find, what is the error of this model? I have assumed this model and I have calculated y bar. So, this is my model and what is the error in this model? Means how much deviation is this model had from the original data.

(Refer Slide Time: 4:52)





$$RSS = \sum_{i=1}^{n} \left(y_i - \hat{y_i} \right)^2$$

Similar thing we have done, when you are doing linear regression. If you remember from the first lecture, so let me define that error. So, suppose this is my y1, yi point and this has this deviation from my y bar. What I think is correct? The real one. So, that means I can subtract yi from y hat or the reverse one. So, that is what I have done, yi minus y bar and again I have squared it because if you remember from the last lecture it is better to square because otherwise it can be negative and positive both and then I have done the same thing for all the n data point that is, I have 5 data point here. So, I sum them all, this summed error that is the

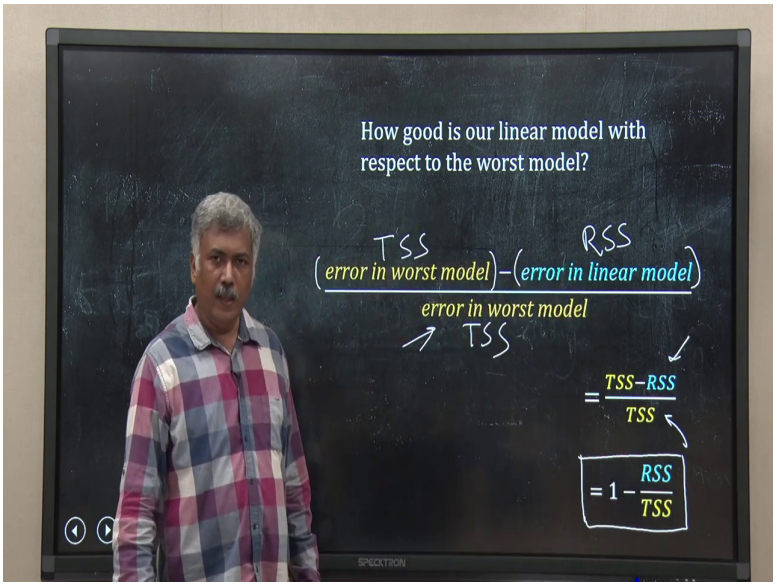
total error in model in this case is called TSS in mathematics or statistic literature that is Total Sum of Squares Error.

So, this is my error in my worst model. What is the error in my linear model that I have used for linear regression? You know that, we have done that in our last lecture on linear regression that is, this is suppose yi and this is yi hat that I get from my linear regression. So, the deviation is ei or eta i, I take the square of that and then I sum it for all the data point n, in this case n is 5.

So, I have summation of yi minus yi hat square and that is actually SSR. In the last lecture, I have said but in this lecture I am calling it RSS just to point out in different book, in different text actually they used these two terms, exchanged this term. So, this is also called RSS, SSR is also called RSS and this is called residual sum of squares error. They are the same thing, there is a name some people use SSR, some people will call RSS, do not get confused.

So, I have RSS, which is the error when I am fitting my linear model and I have TSS when I am fitting my worst model to the same data. Now, I will compare, so I will now create a term or a parameter which will tell me something about how good is my linear model with respect to this bad worst model.

(Refer Slide Time: 7:12)



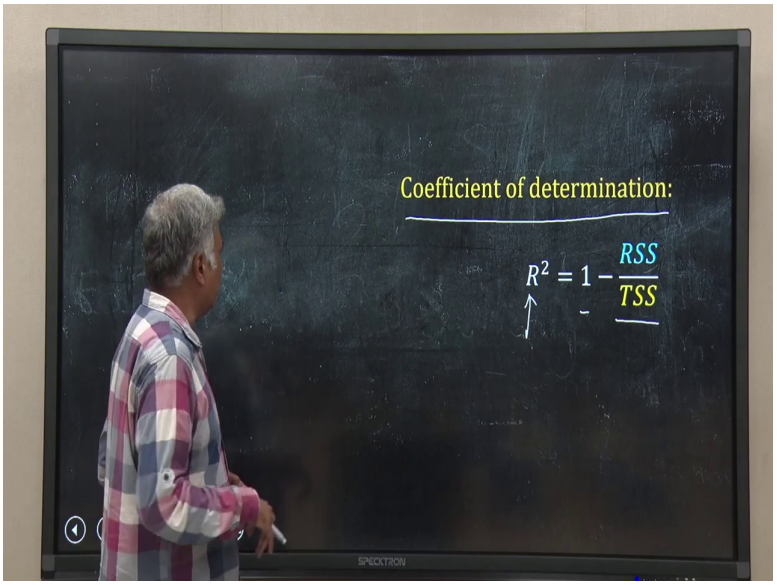
$$\frac{(error\ in\ worst\ model) - error\ in\ linear\ model}{error\ in\ worst\ model} = 1 - \frac{RSS}{TSS}$$

How do I define that parameter? Let us see. So, I define in this way. I take the error in the worst model and subtract the error in my linear model and then I divide the whole thing, the

difference by error in the worst model. So, what is the error in worst model? This is TSS, just now two slide back we discussed.

What is error in linear model? That is RSS or SSR. Somebody will say it is SSR and the denominator is TSS. So, I get TSS minus RSS divided by TSS and if you do simple arrangement what I land up is I have this parameter which I will try to use as a measure of how good my model over the bad model that is equal to 1 minus RSS by TSS.

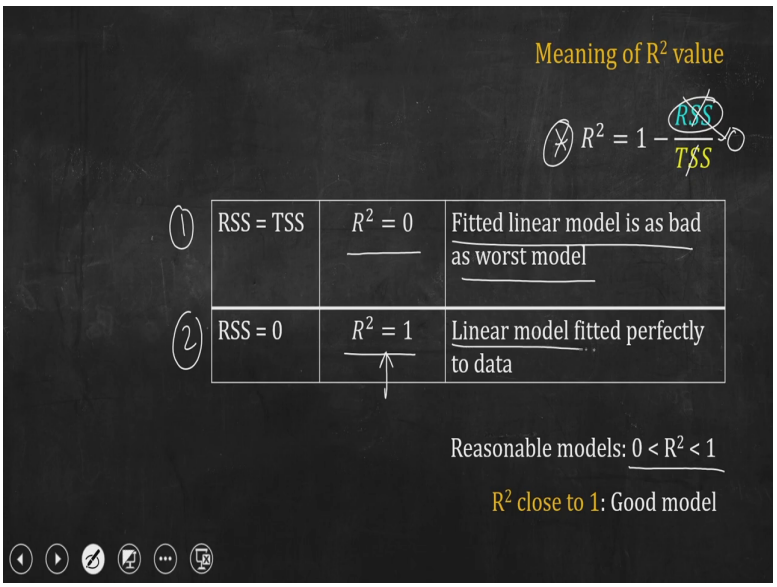
(Refer Slide Time: 8:09)



$$R^2 = 1 - \frac{RSS}{TSS}$$

This term 1 minus RSS by TSS is called R squared, you must have heard of it earlier. It is coefficient of determination. So, R squared is equal to 1 minus RSS by TSS. I have defined it. Now, let us dig into this R squared and try to check out how this value of R squared can tell me something about how good my model, linear model is over the worst model, that is the purpose.

(Refer Slide Time: 8:43)



$$RSS = TSS \Rightarrow R^2 = 0$$

$$RSS = 0 \Rightarrow R^2 = 1$$

$$Reasonable\ models:\ 0 < R^2 < 1$$

$$R^2\ close\ to\ 1:\ Good\ model$$

So, let us see the extreme conditions. So, this is my definition of R square, coefficient of determination. Now, imagine the first case, this is the first case. RSS is equal to TSS, that means the error in my linear model that I got from linear regression is same as the error in the worst model that y is independent of x.

That means the fitted linear model is as bad as the worst model. So, what will happen? If RSS equal to TSS, then these two will get cancelled, I will get one minus 1 R squared will be 0. So, when my linear model is same as that y is independent, that worst model that I am talking of, then R square should be equal to 0. Good enough, I have got a handle now.

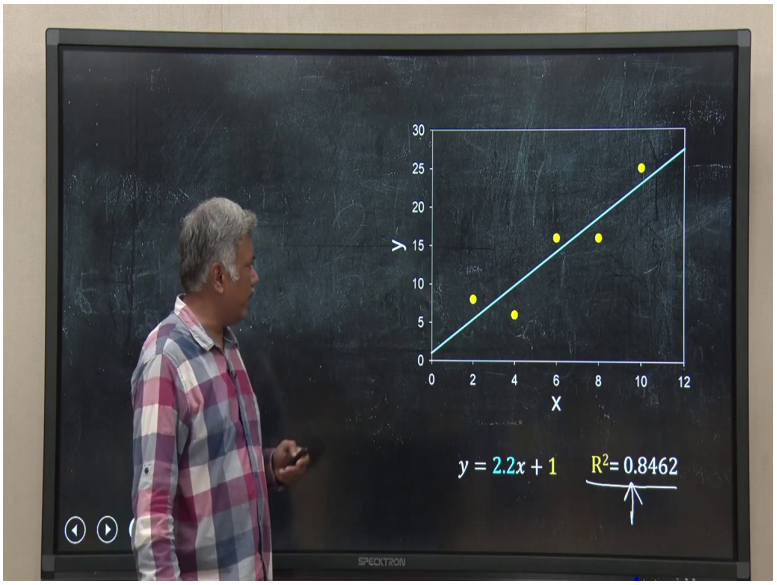
Now, say consider the second case RSS is 0, that means SSR is 0, RSS or SSR is 0, that means my model my y equal to mx plus c that equation, that line I have fitted has fitted perfectly with the data. That means all the data point, observation are on the same line. So, in that case this is 0. So, I get 0 by TSS that means 0 that means my R square is 1.

Now, that will happen rarely, if I have all the data point exactly on the straight line then you really do not need a linear regression, is it? You can simply find it very easily. But let us consider this extreme, so for all realistic problem, all realistic data what will happen, my models R square will lie between 0 and 1. If R square is close to 1 that means my linear

model has fitted good with the data, not perfectly good with the data. If R square is close to 0 that means my linear model is as bad as the worst model.

Remember what is the worst model? Worst model is y and x has no relation, y is independent of x. So, R square, if R square is very low that means my linear model does not hold, it is as bad as the independent model, worst model. So, in this way R squared or the coefficient of determination is my first diagnostic check.

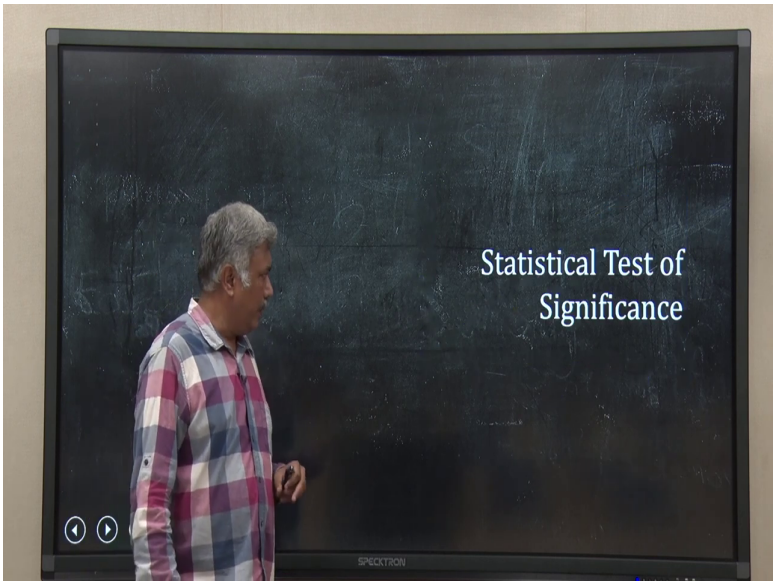
(Refer Slide Time: 11:13)

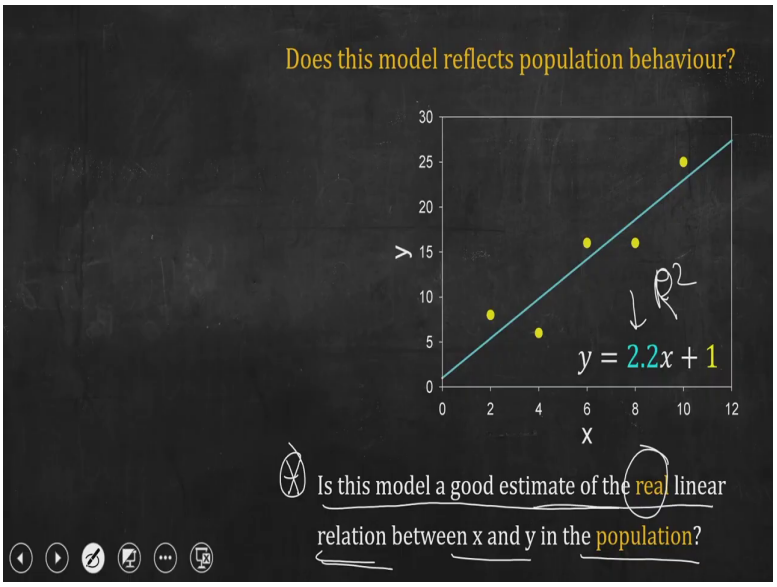


$$y = 2.2x + 1 \Rightarrow R^2 = 0.8462$$

I will like to have a R square value close to 1. And I have done that for my data and I have calculated the R square that turned out to be 0.8462, quite reasonable for the dispersion that I have in the data, that is quite a reasonable R square and I am happy with this. So, this is the first diagnostic.

(Refer Slide Time: 11:33)





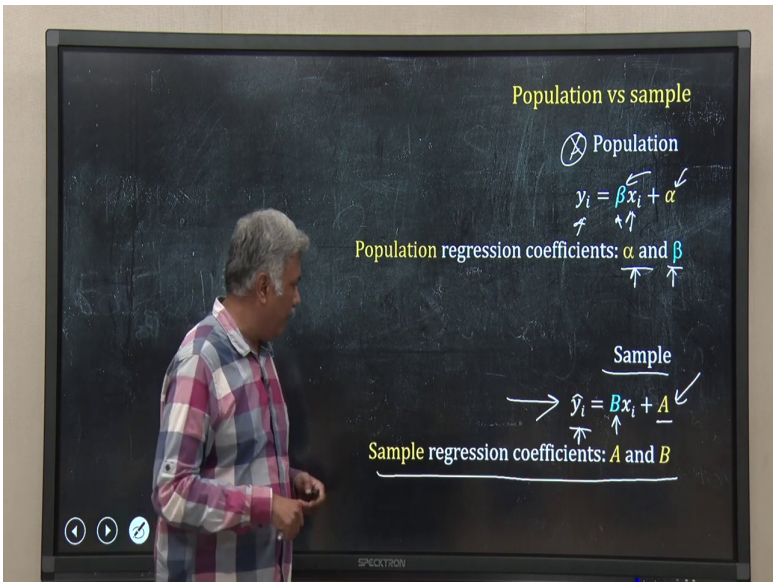
Now, the second diagnostic check. What is that? Statistical test of significance. Now, in this case I will remind you that we have a discussion in the first week where we have discussed about test of significance, test of hypothesis and t test. If you have forgotten that part, please go back and revise that one because those ideas will be used here. I will not go in detail of those again because we already have a detailed lecture on t test.

So, what is the problem here? What is the…how do I formulate the question here? I have done linear regression, I have got the equation slope m and c and I have also got the R square and I am happy with it. But remember here, all this thing that I have done, is based on the sample data. This 5 data point that I have got is nothing but a sample.

So, the question is I have derived a model, linear model. I have fitted that to the sample data, does that model hold for the whole population? So, let me word correctly, so the question is

this model a good estimate of the real linear relation between x and y in the population? So, just to remind you in that, in the test of significance lecture, we have discussed about this issue of population and sample. Please recheck that once again. So, the question is, is this model a good estimate of the real linear relationship between x and y in the population. So, I have to jump from sample model to population model and that is why I have to use test of significant or test of hypothesis.

(Refer Slide Time: 13:30)



$$Population: y_i = \beta x_i + \alpha$$

$$Sample: \hat{y_i} = Bx_i + A$$

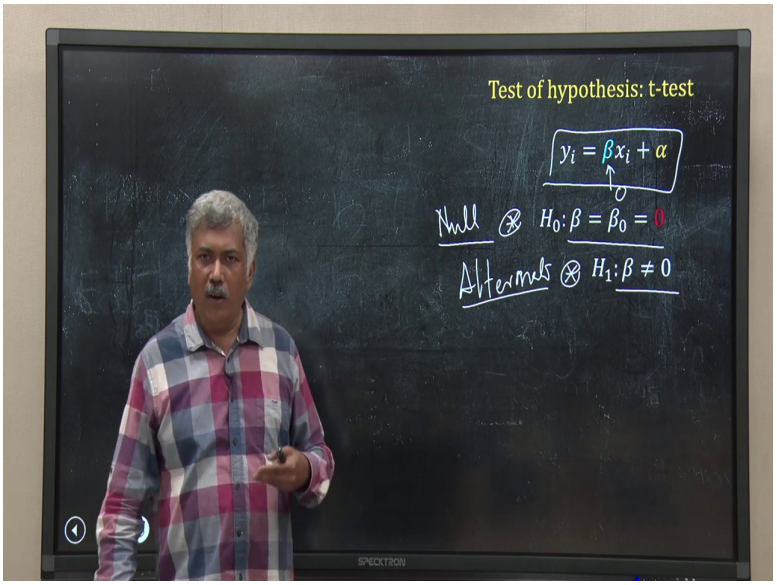
So, how will approach this? Let me write down a few terms so that we can move forward. So, I have sample and I have done a linear regression to fit equation which has b, which is the slope. Earlier I was using m, I have used it b here because I am using beta in the other case. So, the same thing, b into xi plus a is equal to yi hat. Why I am using yi hat? Because that is not the observed data, that is coming from the regression equation.

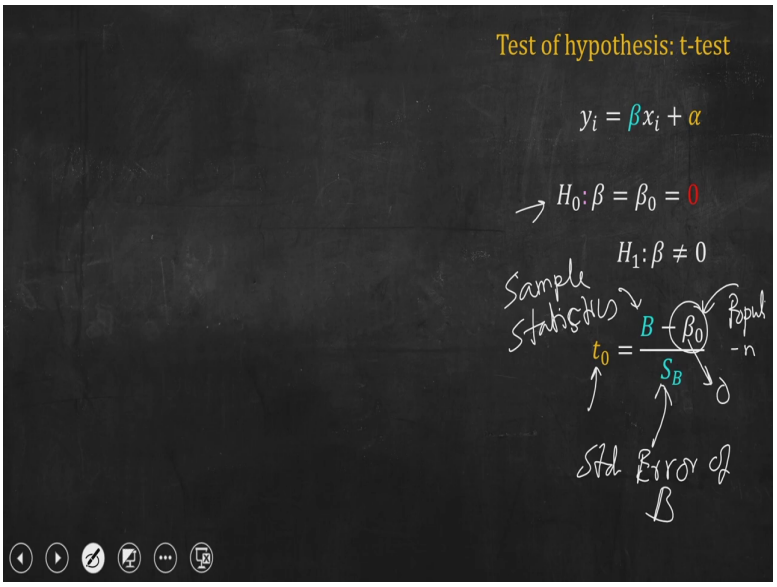
So, this is my model from the sample. I have five data point that is a sample, I have fitted this equation and I have got b and a. So, this b and a are sample regression coefficient. They are not the real population level regression coefficient because I do not know the population. I have not seen the population and I do not know whether there is such values for a and b or not at all.

What will be the population level? In the population level imagine we have a linear relationship between x and y and in that case the relation is y equal to beta x plus alpha. So,

in this case alpha and beta are the real population level regression coefficient. So, if I have the population level data somehow and then if I do the regression, I will get alpha and beta although I will never know that because I never will have a population level data.

(Refer Slide Time: 15:09)





$$y_i = \beta x_i + \alpha$$

$$H_o: \beta = \beta_o = 0$$

$$H_1: \beta \neq 0$$

$$t_\circ = \frac{B - \beta_\circ}{S_B}$$

$$p = P\left(t \geq t_o\right) \text{ as per suitable } t - \text{distribution}$$

$$If\ p < \alpha \Longrightarrow Reject\ H_{\circ}$$

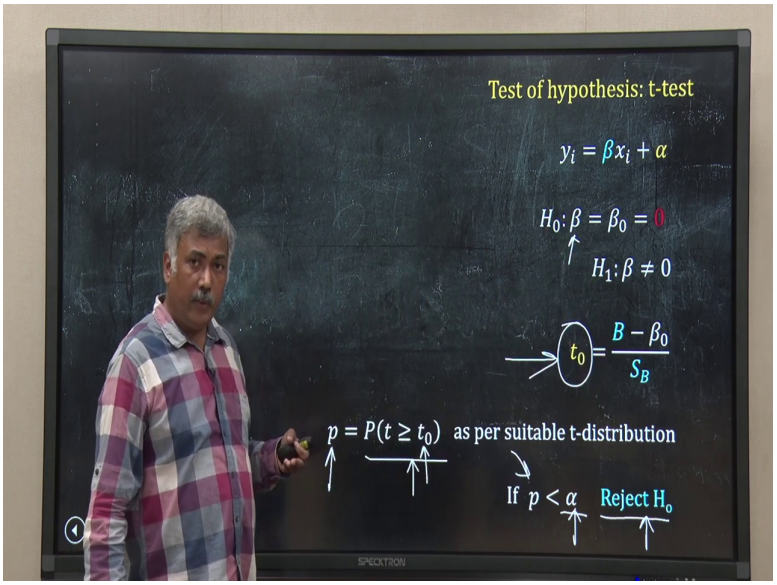
So, now based on these two definition, I will now perform test of significance or test of hypothesis. So, what I will do? I will use t test, so this is my population level model, I make two hypotheses. The first one is null hypothesis, h naught, h naught says that beta is 0 that means this 0. That means there is no relation between y and x, just like the idea that we use for worst model to get R square.

The alternate hypothesis is the idea which we have used for linear regression. We believe that beta is not equal to 0 and that is why we have fitted y equal to mx plus c or y equal to bx plus a. This type of equation. So, the alternate hypothesis is beta is not equal to 0. It is a t-test, so I have to formulate the t-statistics, t-statistics if you remember from that video, otherwise you can go and check once again, t-statistics is equal to the difference between sample statistics minus population parameter divided by the standard error of the sample statistics.

So, in this particular case what will happen? My t0 that is for the t-statistics for null hypothesis will be equal to B, B is what? B is sample parameter or sample statistics, not population statistics. This is a coefficient for the sample whereas beta naught is for the population, this is a population parameter.

And we have in the denominator, we have the standard error of B. There is a way to calculate the standard error of B, I have not discussed that but in R, we can very easily do that. So, I will skip that mathematical part. So, this is my t statistics. Now, let us look into this by null hypothesis, this one is equal to 0 that means I am landing up, we are reaching the t equal to b by standard error of b. That is as simple as that.

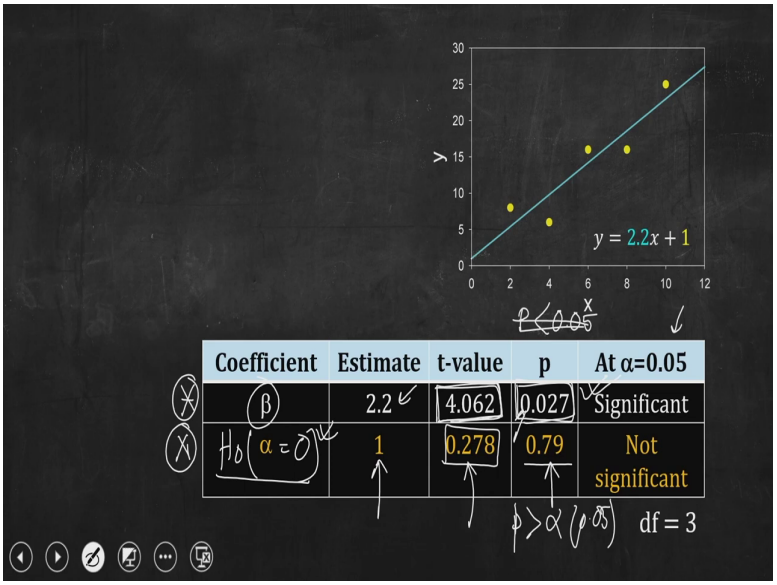
(Refer Slide Time: 17:41)



So, now once I have got this value, then I calculate the value of t naught and then I go back to t distribution and I try to calculate the probability p such that that t is, the t will be bigger equal to the calculated t naught. So, I want to calculate the probability that the value of t will be bigger equal to the t naught that I have calculated. And I use that using a suitable t distribution with a particular degree of freedom.

And if that probability, if that probability is less than a cut off which we call, level of significance, like if it is less than a cut off then what we do? We reject the null hypothesis. So, this is the basis of t test and we are applying that for our regression model, particularly in this case for beta, to test whether null hypothesis is beta equal to 0. Or the alternative hypothesis beta is not equal to zero, which one is correct.

(Refer Slide Time: 18:56)



Coefficient	Estimate	t-value	p	At α =0.05
β	2.2	4.062	0.027	Significant
α	1	0.278	0.79	Not significant

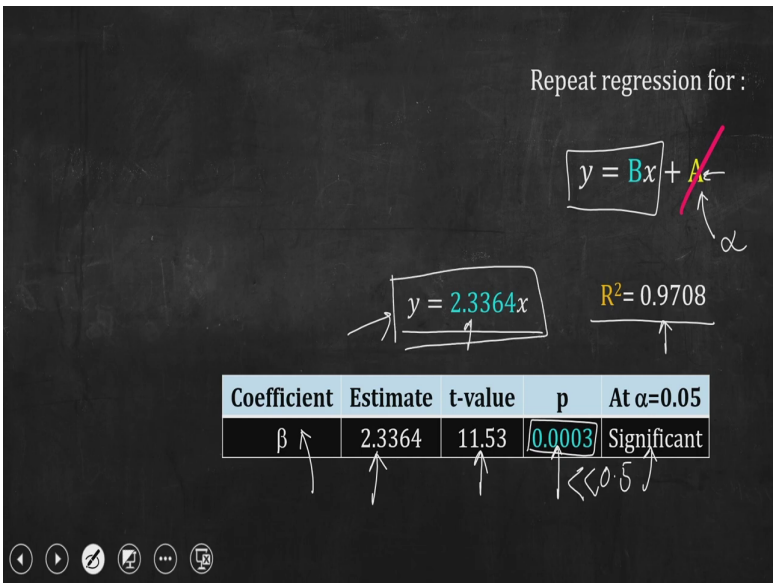
So, I have done that this calculation for both, in fact, I will demonstrate that in a separate lecture that r by default when you are doing linear regression using R by default will perform this. So, I have collected the data. So, let us see the data. So, this is my beta, estimated value you already know 2.2 and its t value for the null hypothesis came out to be 4.062, using a suitable t distribution, we get the probability that the t value will be bigger equal to this 4.062 is 0.027.

And if I use a level of significance of 0.05 then this p is less than 0.05. So, that means I reject the null hypothesis. What is the null hypothesis? That beta is 0, if I reject it that means I have to take the alternate hypothesis that beta is not equal to 0. That means my beta coefficient is significant. So, I should keep it in my model.

Let us look into for alpha, the intercept. If you remember this alpha is 1, the value is 1. The t value turn out to be 0.278. Now, using the probability distribution for t distribution with degree of freedom 3 in this case. What we find that the probability of having 2.78 or bigger value is 0.79. So, this is the p. Now, in this case, in this case p is bigger than alpha, alpha is 0.05.

So, as p, the probability of getting that t or bigger t is bigger than that cut off. That means, I cannot reject the null hypothesis. I have to accept the null hypothesis. What is the null hypothesis? Null hypothesis in this case is alpha equal to 0. This is my null hypothesis. So, I have to accept this one, that means in other word people will say alpha is non-significant. Alpha is non-significant, beta is significant but alpha is non-significant.

(Refer Slide Time: 21:36)



$$y = Bx + A$$

$$y = 2.3364x \quad R^2 = 0.9708$$

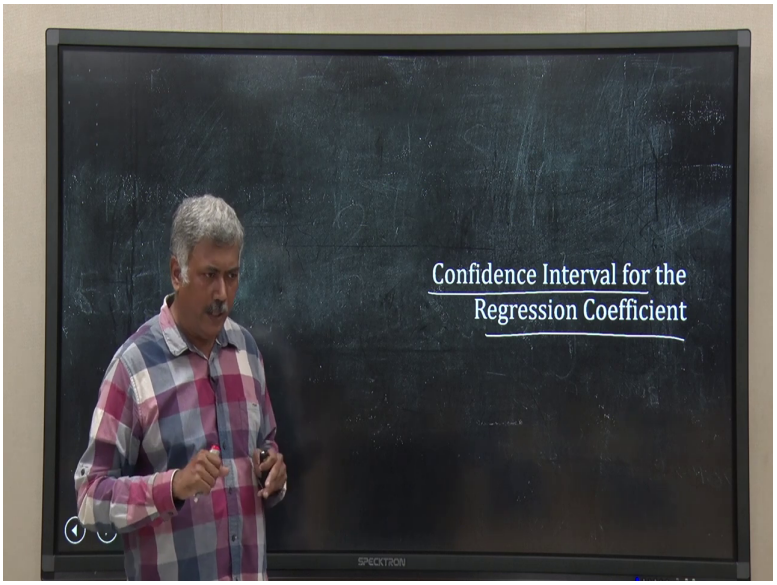
Coefficient	Estimate	t-value	p	At =0.05
β	2.3364	11.53	0.003	Significant

If something is non-significant, why should I keep that in my model? So, I will remove this from my model that is my next step. So, I have shown that alpha, this is actually equivalent to alpha. Alpha is for the population level, a is for the sample. Alpha is not important, not significant. So, that means I will remove alpha, that means I will remove a from my equation for fitting the data.

So, I will have only y equal to bx or y equal to mx. And I have done the regression again, again I have used R and I have got a new regression equation, y equal to 2.3364x and if you look R square has improved now. It has 0.9708 and again R has given me the statistical test, result also. Now, beta is 2.3364 that is what written here, its t value, it has calculated and the probability of getting that t value or bigger is 0.0003 and obviously this is much less than my level of significance 0.5, so beta is significant.

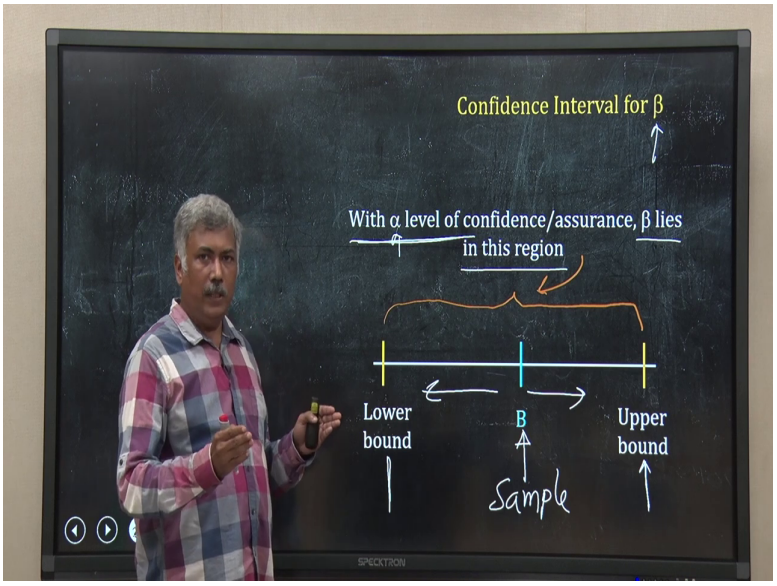
So, that means, this model is the better model than the initial model that we were handling till now starting with the first lecture on regression. So, this is I will say is the correct model. So, what I have done, till now we have discussed about two diagnostic checks R square and t test or test of significance.

(Refer Slide Time: 23:13)



Now, the last one, last diagnostic test that you should always do, when you are doing linear regression. You should calculate the confidence interval for all the regression coefficient that m and c or a and b. That you have calculated by linear regression, you should calculate the confidence interval of that. What do I mean by confidence interval? Let me explain that, I will not go detail in how to calculate that because ultimately you will use the computer to do that but we should understand the meaning of it.

(Refer Slide Time: 23:35)

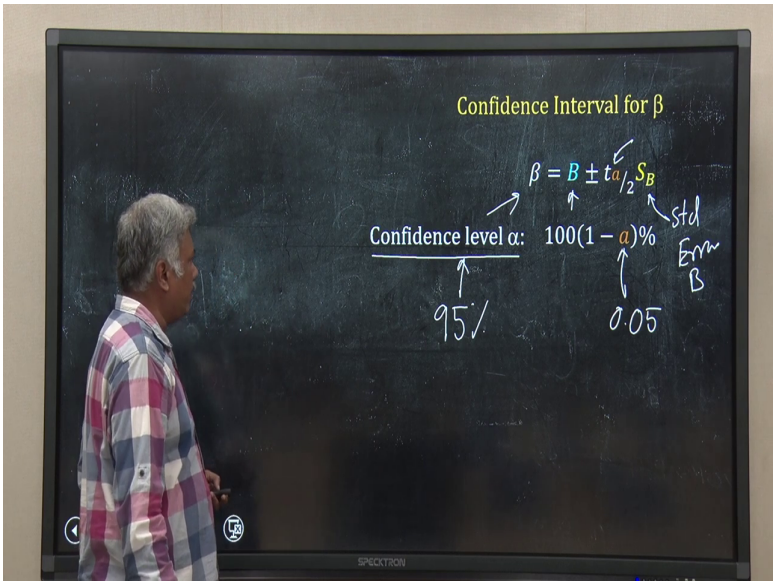


So, my regression method has given me a value of b. Remember this is sample thing, I am calculated from the sample. The population level thing is beta, equivalent thing is beta. I do not know beta but I assume that the beta must be somewhere around this b and there must be a lower bound for that and there must be upper bound for that. So, I want to know this bound,

what is the lower bound and what is the upper bound around this estimated b that beta should lie.

Now, you can easily understand, you cannot have a straight forward yes, no type answer for this question. There must be a probabilistic answer. So, then you can reframe your question, you can say tell me what is the lower bound and upper bound with assurance of 98 percent or 95 percent. So, what we say, tell me the with a level of significance or level of confidence, alpha beta lies between a particular region, in a particular region around b, that we want to calculate and that is called the confidence interval.

(Refer Slide Time: 24:53)



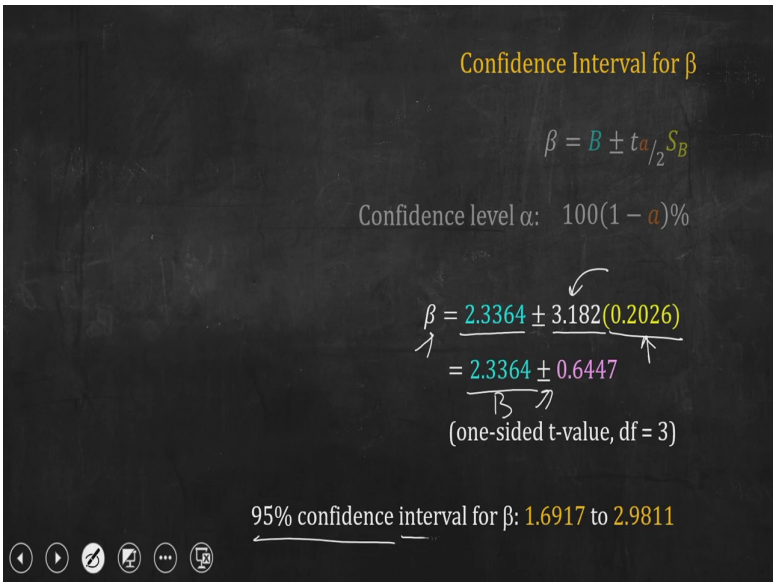
$$\beta = B \pm t_{\frac{a}{2}} S_B$$

$$Confidence\ level\ \alpha\text{: } 100(1 - a)\%$$

How to do that? Without going into the mathematical derivations, I have written down the equation here, you can use this relationship which is beta equal to b. So, beta, the value of beta will be equal to plus minus that is where saying the confidence interval, plus from b minus from b, plus minus t of a by 2. I will come what is a, into the standard error, standard error of b. What is a? Remember you have a confidence level? So, you may have said, I want 95 percent confidence level that beta will lie between this range.

So, in that case, this a will be nothing but 95 percent equal to 100 into 1 minus a, in this so, a will turn out to be 0.05. So, using the t distribution, we calculate t of a by two and then multiply that with SB, the standard error of b and then we have b term. The value of b we calculated by regression.

(Refer Slide Time: 26:01)

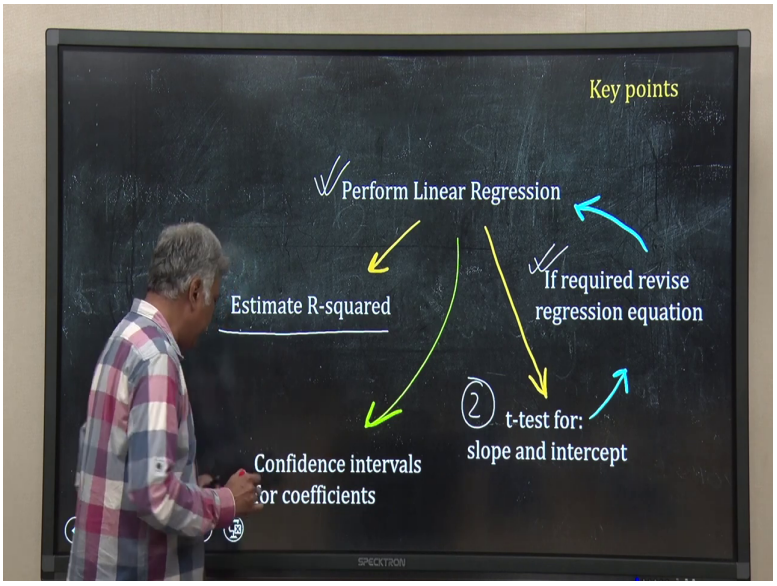


$$\beta = 2.3364 \pm 3.182(0.2026) = 2.3364 \pm 0.6447$$

For my data set that we are analysing, I have calculated that b is 2.3364, the standard error is 0.2026 and from t distribution, I have got the value of t a by 2 where a is actually 0.05 is 3.182 and the net result is, I get beta equal to 2.3364 that is b plus minus 0.6447. That means from 2.3364 plus 0.6447 and minus 0.6447 beta will lie in this range.

So, that means at 95 percent confidence interval, 95 percent confidence, the interval is of beta is 1.697 to 2.9811. So, in this way you should be able to calculate the confidence interval for alpha also. I have not done that because remember we have removed alpha or the intercept term because it turned out to be by t test, that is not significant. So, I have just performed this for beta.

(Refer Slide Time: 27:17)



So, let me jot down, what we have got till now. This particular lecture, we have discussed about three diagnostic checks that you must always do with linear regression. Performing linear regression is very easy, you can do it manually, you can use excel, you can use R or any other software for data analysis to get linear regression. But most of the time we stop there but our job does not stop there rather we are supposed to do some diagnostic check and do corrections.

So, how will start with? You should always start with the performing the linear regression. Then you check the R squared, the R square should be close to 1 and if you are happy then you go for the test of significance for the slope and intercept, that is the regression coefficient and if your t test does not pass for any of this, you remove that term and if required, you redo the regression again and that is what I have done for this particular example of 5 data point and I have removed alpha, the intercept.

Once you have done that, you are done with the regression and again you calculate the R square and you also calculate the confidence interval and that confidence interval should be should good enough, so that they're narrow and you should have a high confidence that the coefficient lies in that interval.

So, that is all for this lecture, with these two lectures on linear regression, one and two, we have learnt how to do linear regression and what are the diagnostic check you should do. In the next lecture, I will demonstrate how to use R to perform linear regression and all the

statistical test for your linear regression-based model. See you in that lecture, till then happy learning.