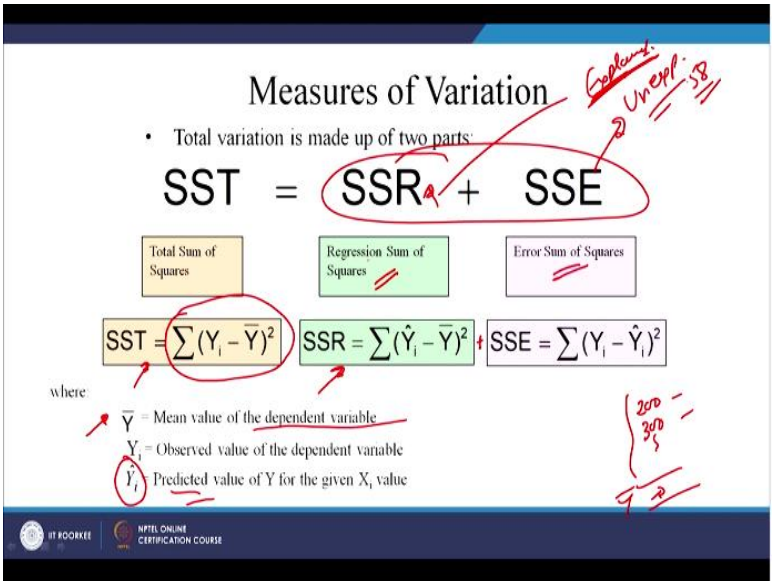
Business Statistics
Prof. M. K. Barua
Department of Management Studies
Indian Institute of Technology – Roorkee

Lecture – 57
Simple Linear Regression

Hello friends. I welcome you all in this session. As you are aware, in previous session we were discussing about basics of regression. We have seen what is regression, we have seen what is dependent variable, what is independent variable, what is strong correlation, weak correlation. We have worked out an example as well on regression wherein the question was like this. There was house price which was dependent variable and the square feet of the plot was independent variable.

So we did find correlation between house price and area of the plot which were measured in terms of square feet right and we did see there was some correlation between those two variables. Now let us look at some more points related to regression.

(Refer Slide Time: 01:38)

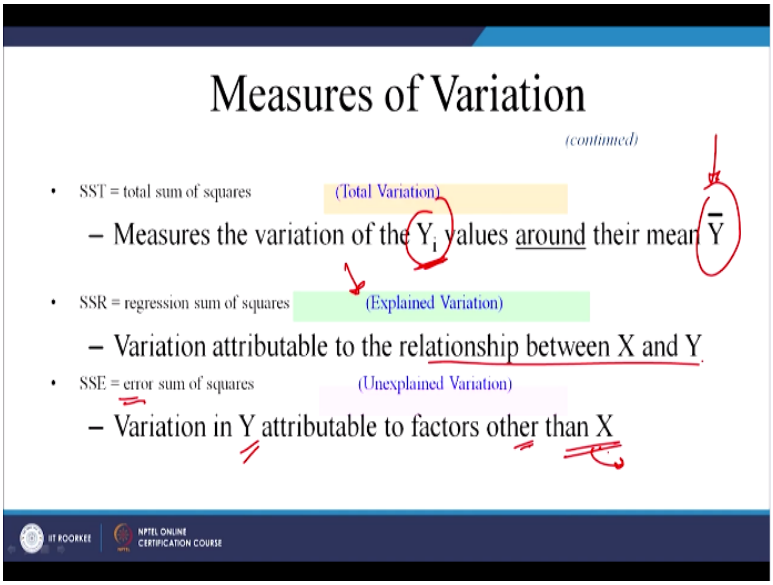


In fact, in previous session we said that correlation was let say it was 0.58 or 58%. So we said that the 58% variation in dependent variable was explained by independent variable. So let us talk about variation today. So there is something called total variation which is nothing but regression sum of square and error sum of square. So just add these two, regression sum of square is nothing but the explained variance.

You can write it like explained variance and this is nothing but unexplained variance by the model right or when we say explained variance it means this much of variance is explained by independent variable. So this is total sum of square which is this summation of $(y_i - \bar{y})^2$. So Y bar is nothing but the mean of all your independent variable data right. So let us say in previous question, we had let us say the price was let us say 200.

The size was something, then price was 300, the size was something. So you just take the mean of all these values so this is Y bar mean of the dependent variable. Y is nothing but observed value of the dependent variable right. Yi hat is nothing but the predicted value of Y for given value of X right. So this is how you can calculate total sum of square if you have got regression sum of square and you just add error sum of square to regression sum of square right. You will get total sum of squares right.

(Refer Slide Time: 03:32)

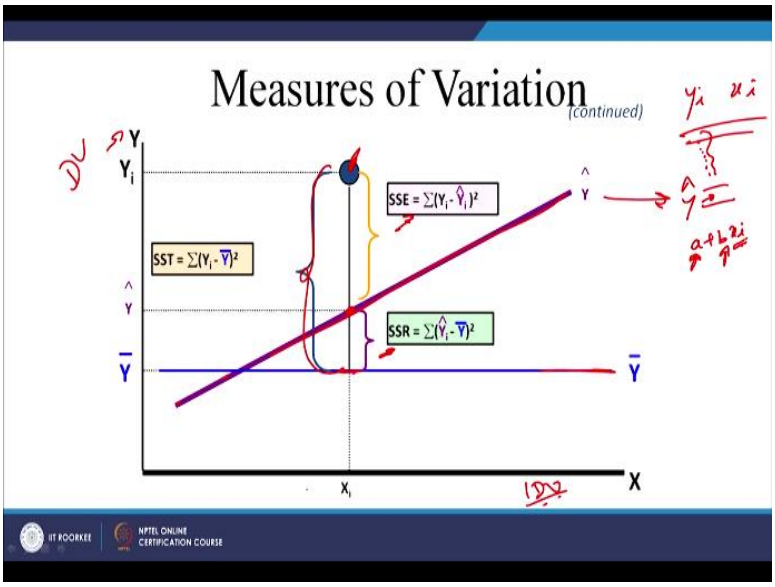


So what is total sum of square? It measures the variation of Yi values around the mean right. This is nothing but mean of dependent variable. So it measures the variation of Yi individual observations right around the mean. Regression sum of square as I said this one is nothing but explained variation. It is the relationship between X and Y. The variation attribute table to the relationship between X and Y.

And error is the value of Y which is getting affected by factors other than X is not it, other than independent variable given in the question right. So variation in Y attributable to factors other than X that is why it is called error sum of squares so what those variables are called

which are other than X. We have seen in case of ANOVA, they are called concomitant variables or confounding variables right. So let us look at one more slide on variation.

(Refer Slide Time: 04:50)

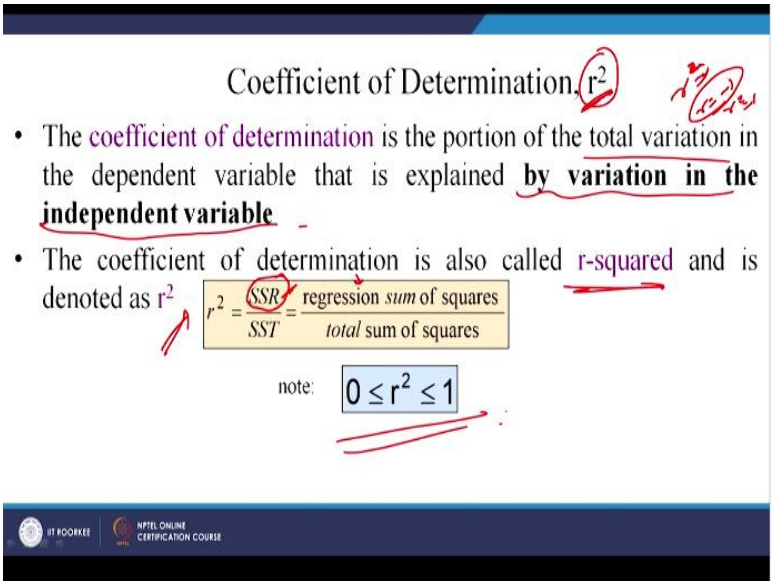


So this is your independent variable X right, Y is your dependent variable, so you have got different values of X and you will get from the question you will be given all Yi's and all Xi's right. Find out line of best fit like this, is not it, which is nothing but prediction line right y is nothing but prediction line right. Now this is your Y bar, as I said sum of all these dependent variable values right.

Just sum of all these and take the average right. So let us say this is your observation ith point and this is your point which you have estimated from this line of best fit is which you drawn over here. So regression sum of square is base, from this point point on the Y hat line or estimation line to the mean of the dependent variable right. You can see this in this this. This is Y bar which is the mean of dependent variable and Y hat.

So this difference is regression sum of square and from Y hat let us say Y hat 0.2, this point the given Xi value is nothing but error sum of square okay and if you add these two this becomes total sum of square, is not it? So the point is you need to calculate a and b or in other words the constant and the slope of independent variable. Let us call it X or you can have several values of X right.

(Refer Slide Time: 07:08)

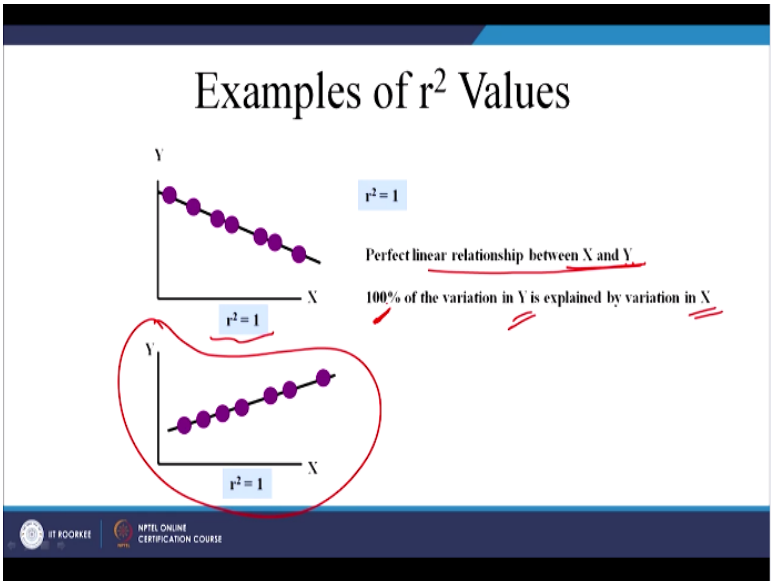


Let us look at coefficient of determination. So far we have seen regression, let us look at what is coefficient of determination or r square is the portion of the total variation in the dependent variable that is explained by variation in independent variable. So this is nothing but coefficient of determination. We always represent it by r square right and its value can never be negative because it is square term right.

So if r is let us say -1 but r square becomes 1 right so the coefficient of determination is also called r-squared and is denoted as this. So this is nothing but the regression sum of square to the total sum of square. Why we are calling it regression? Because this is nothing but the explained variation right, so what would be the unexplained variation, that would be the error right SSC, error sum of square/total would be the unexplained variation right.

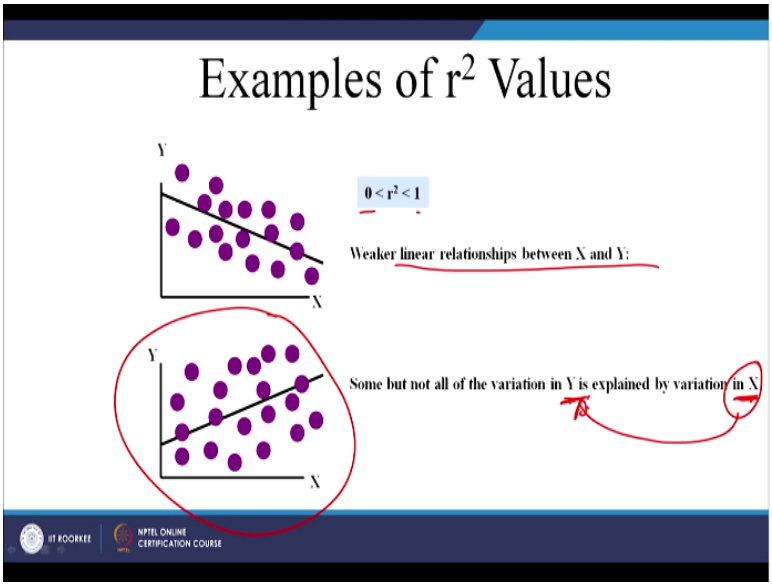
And when you say it like this SSR/SST it is explained variation right. So its value is always between 0 and 1 right.

(Refer Slide Time: 08:34)



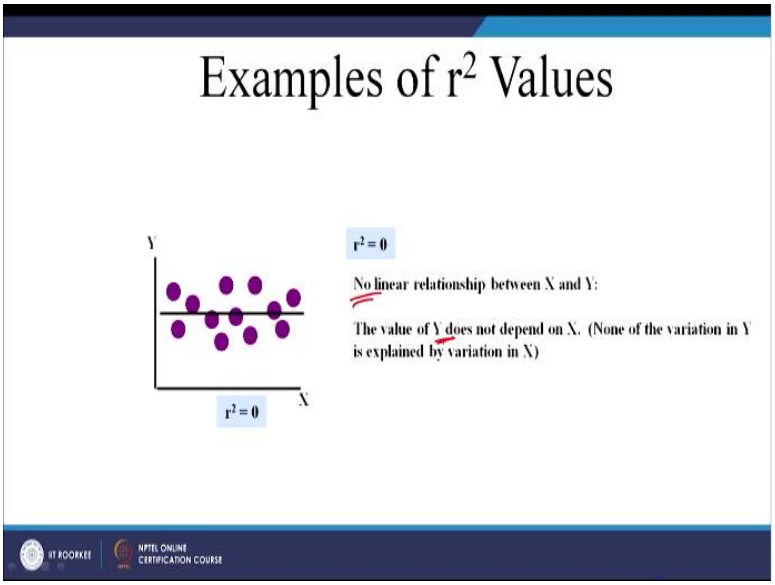
So you can have couple of examples of r square values. So r square is 1 which means there is perfect correlation between X and Y. So perfect linear relationship between X and Y when we say r square is=1. Again, one more example of r square is=1. So 100% of variation Y is explained by variation in X right. Generally, it will not be the case; you will always have this value less than 100 okay.

(Refer Slide Time: 09:17)



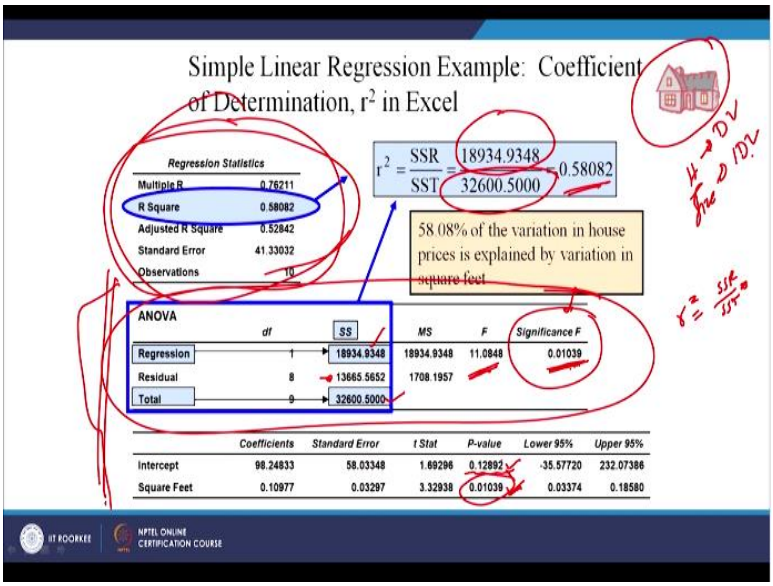
So let us look at some more examples. So r square is between 0 and 1 so weaker linear relationships between X and Y. One more example of weaker linear relationship some but not all the variation Y is explained by variation X. So independent variable X is explaining dependent variable Y but not fully right. There are some other variables which are also explaining Y right. So these two are examples of weak relationship between X and Y.

(Refer Slide Time: 09:56)



So there is no relationship so X and Y are independent right. So no relationship, the value of Y does not depend on value of X right.

(Refer Slide Time: 10:09)



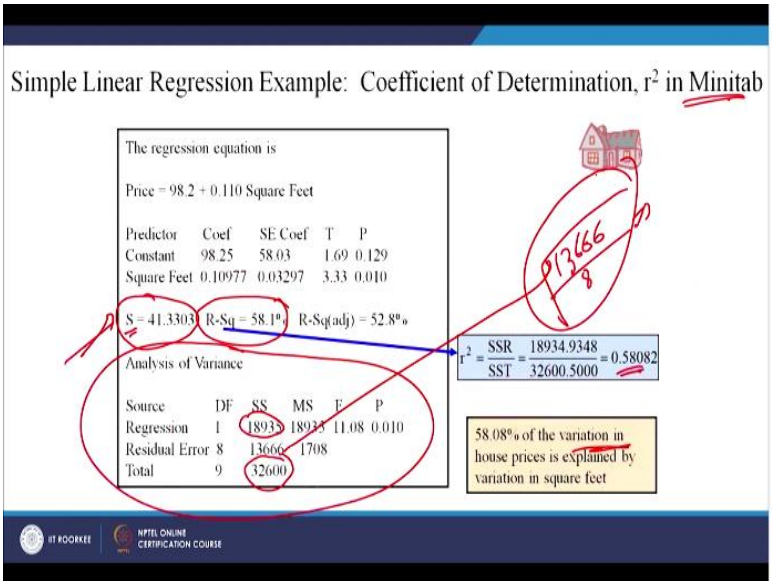
So this is the example which we worked out wherein we had house price as dependent variable and the size of the plot it was independent variable right. So we have seen in that question, our r square was 58.08%, so the same this now in fact we did see how to interpret this table, let us look at how to interpret this part of the table. Now this is ANOVA table right. In ANOVA table, you have got regression sum of square and total sum of square and this is nothing but error sum of a square or you can call it residual right.

So what we have said what is r square, explained variation right. So that is regression sum of square divided by total sum of square, is not it? So regression sum of square is 18934 which

is this, total is this and the r square value is 0.58. So if you are not given this particular part of the table, then you can calculate r square from ANOVA table as well. If you look at this, this is nothing but P value of this model right.

F value is this but this is nothing but P value of this model, so this 0.01 that means the model is significant one right which is less than 0.05. Now if you look at the P values over here for these two, for these two unknown variables that is intercept and coefficient of independent variable. For intercept P values 0.12 so will say that intercept is not significant in this regression equation but square feet which is an independent variable is significantly affecting dependent variable okay. So this is one more output from Minitab.

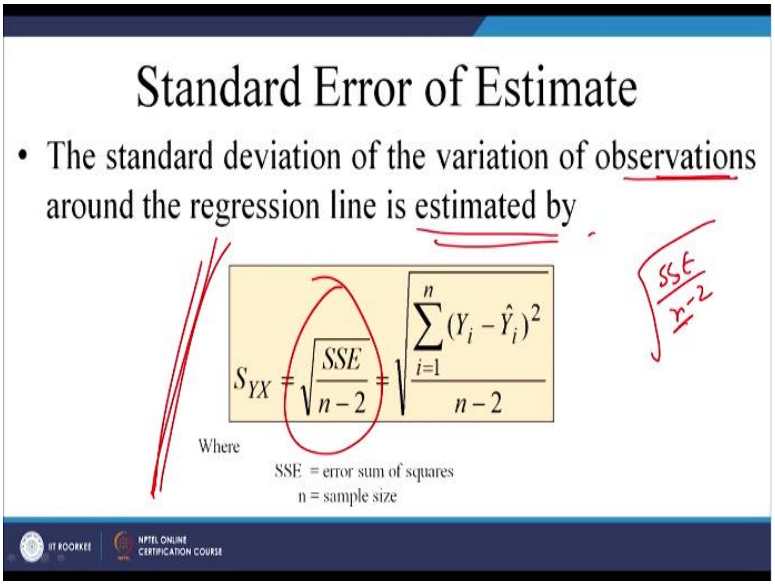
(Refer Slide Time: 12:32)



You can see r square is 58.1, this is standard error. We will talk about standard error after some time. This is your ANOVA table right, so you have got regression sum of square, you have got total sum of square the ratio of these two is r square. So will say that the 58.08% of variation in house price is explained by variation in square feet, so if the size of the plot increases the house price will also increase.

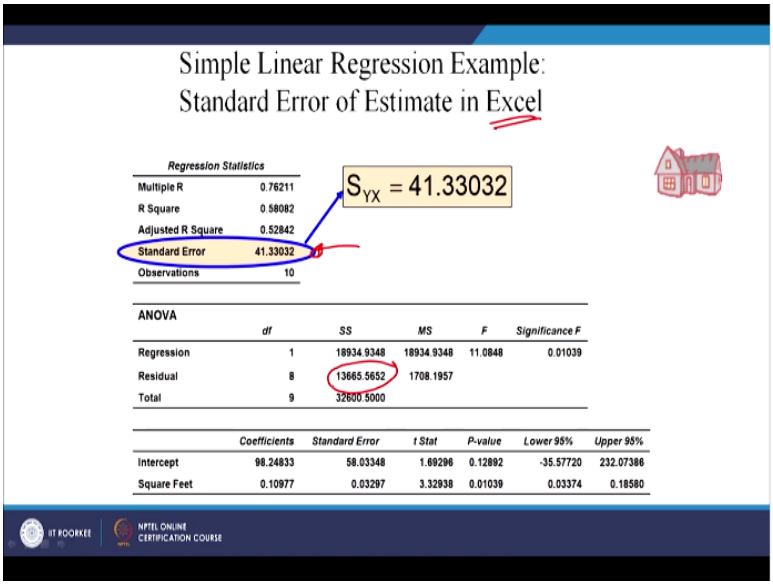
Let us look at standard error. We have calculated r square so far. We did not calculate standard error, so standard error is if we measure it using something called sigma right or standard deviation.

(Refer Slide Time: 13:30)



So standard deviation of variation of observations around the regression line is estimated by this. So this is nothing but error sum of square right divided by n-2, n is number of sample size. So if you have got let us say SSE is available right, SSE error regression is 13666 right, this value divided by n-2, n is 10 right. There were 8 data points in the question right. So this is 10-2 this is 8, so this value is equal to S. You can work out this particular value right.

(Refer Slide Time: 14:25)



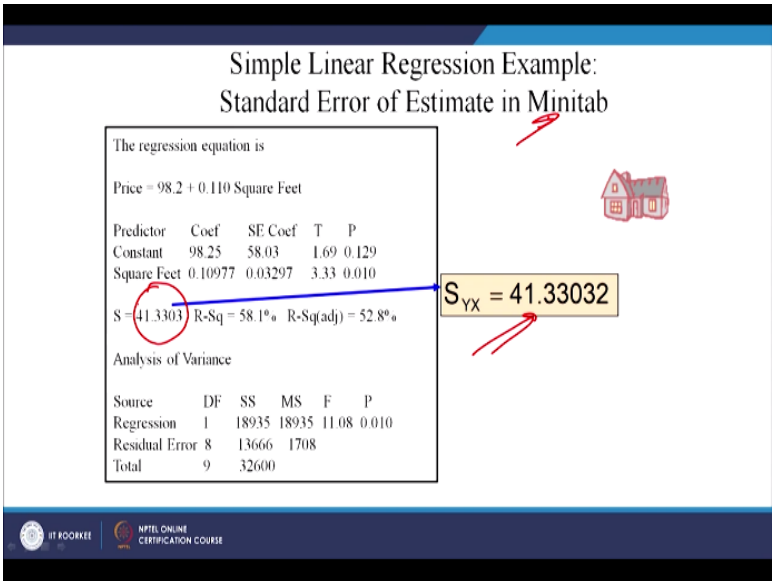
Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

ANOVA	df	SS	MS	F	Significance F
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

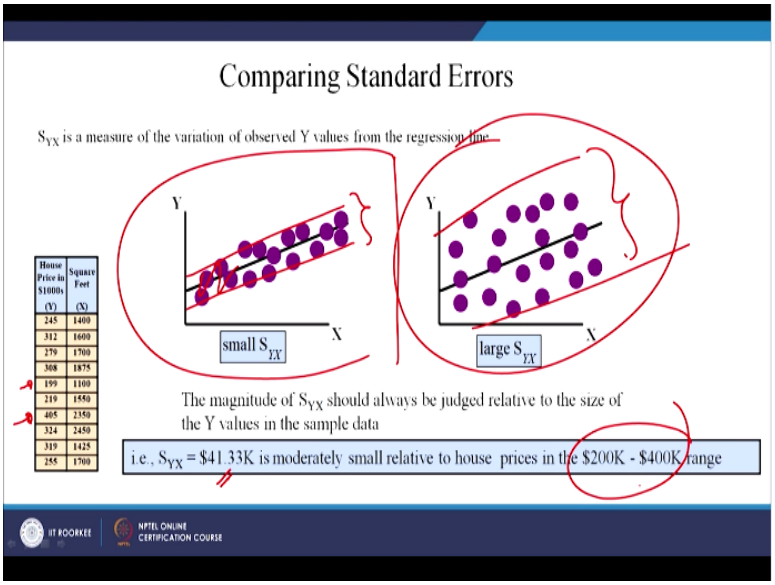
So standard error is this. Now standard error of course in excel output is available here or you can calculate using this value divided by n-2 under root is not it? So this is standard error.

(Refer Slide Time: 14:47)



This is standard error using Minitab right; the previous one was using excel right. So standard error here is 41.33 and in previous case in excel as well it was 41.33 right. So you are getting same output using excel and Minitab. Now let us compare standard errors. So if you have got let us say two models, then you should always look at standard error values for comparison. You will come to know which model is a better model.

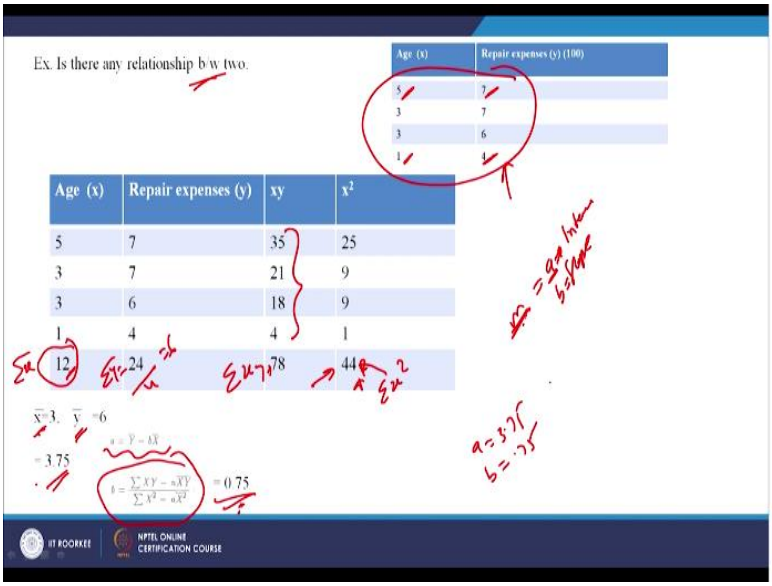
(Refer Slide Time: 15:24)



So you have got X and Y values over here, let us say this first model, one model in this, second model right. So standard error is small over here because this distance just see this value, all these points are densed right but here they are scattered, is not it? So this is nothing but standard deviation right okay. If you look at our question which we did solve in previous class so standard error is this much right as we have said 41.33 right, 41.33 is moderately small relative to house price which is in the range of 200,000 to 400,000 right, is not it?

Just see this is approximately 200,000 minimum value and this is 400,000 right, so comparative to house price, the standard error is quite small okay.

(Refer Slide Time: 16:26)



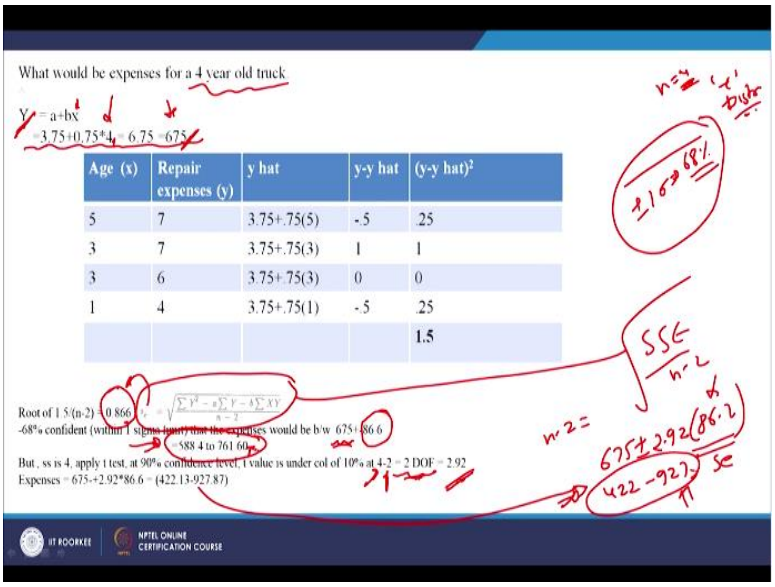
Let us look at this example. In this example what we have done or the question is like this. So there is a transportation company and the manager of the transportation company wants to know is there any relationship between age of the truck and repair expenses. Generally, we know that the moment the age of vehicle increases repair expenses also increases right. So he just wants to know the relationship.

So he collected data from his let us say register, there were several values of repair expenses having different ages of vehicle right. So he just collected 4 samples right. So when the age of the truck was 5, repair expenses was this much, when age was 1, repair expense was 4, 4 rupees or 400 rupees or whatever it is right. So is there any relationship between two right, so for this you can calculate first of all as I said it is a intercept value, b is slope right.

This is intercept, so either you call it m or a whatever it is right, is not it? If you are writing equation like this y=mx+c so this is nothing but intercept right. So you have got this is the summation of x right, this is summation of y, this is x and y, x*y, summation of xy, 78 and this is summation of x square is this right. So we know that a can be calculated like this y bar -b*x bar okay.

So y bar is nothing but you just divide it by 4 just 6 right, x bar is 12/4 is 3, so a is=Y bar-b*X bar right and b can be calculated like this, it is summation of $XY - n(\overline{X}\,\overline{Y})$ right divided by summation of X bar, summation of X square-n*X bar square. So this is how you will get b is=0.75. Once b is known a can be known right which is 3.75. So a is 3.75 and b is=0.75 right.

(Refer Slide Time: 19:27)



So this is your line of best fit right or Y hat prediction line in other words. So a+bx, now the question is what would be the expenses incurred by the manager if the truck is 4 years old. So we have got this x which is independent variable, so a+b*4 so If you want to find out x expenditure just put the value of the age of the truck over here right. So this becomes 675 right, actually all these values are in 100 right.

So that is why this expenditure is 675 rupees. So if a truck is 4 years old then it will be having expenditure of 675 rupees. Now let us find out standard error. You can calculate standard error like this or SSE/n-2 the same thing right. So when we calculate ac it becomes this, so the ac is nothing but what we have calculated here is 0.866 by this formula or if you have Minitab you just input independent variable you will get the value of standard error directly right.

So the standard error is 0.866 and we know from our knowledge of normal distribution that area within +-1 sigma limit is 68% right. So this is our expenditure right for a truck which is 4 years old right 675 ± 86.6, when you convert it into rupees this becomes 86.6 rupees. So

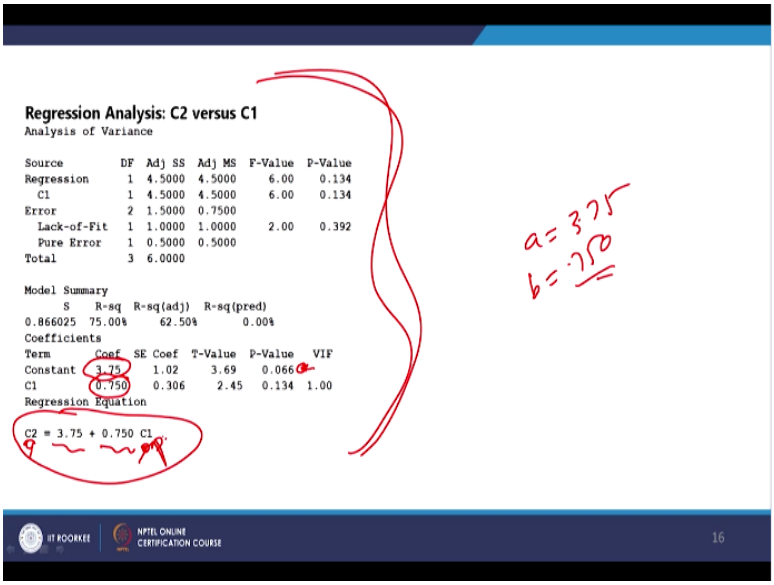
675 ± 1 sigma is this, so will say that the manager is now 68% confident that the truck for which or the age of the truck is 4 years the expenditure would be in between this range.

The minimum would be 588 and maximum would be 761.6 right but if you look at this example carefully, here we have used normal distribution right but here sample size is just 4 right. So we cannot use Z-distribution right, we should use something and t-distribution, is not it? t-distribution, so when we have got t-distribution will use at appropriate degrees of freedom at 2 degrees of freedom right, 4 is the sample size.

So n-2 let 10-2 degrees of freedom, in t-table the value is 2.92 right. So will say that the repairing expenses would not be this much, it would be 675 ± 2.92*86.2 which is nothing but standard error right, this is standard error. So this becomes 422 to 927 right. Just see this value, is not it? So now the manager is 90% confident that the expenditure for a truck which is 4 years old would be 422 to 927 rupees.

So you are getting a better answer right. Of course, this is not an incorrect answer but of course you can say it is an incorrect because you did not take care of proper distribution because you used over here z-distribution rather than t-distribution.

(Refer Slide Time: 24:15)

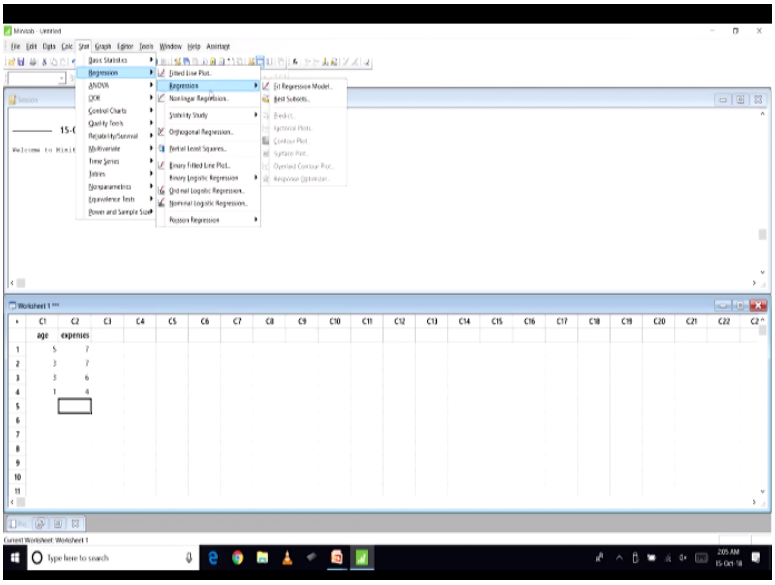


So let us and this is the output of the same question using Minitab. So here if you look at this constant values 3.75 which is nothing but a right and this is b right which is slope right 0.750 and if you look at the P value over here, then you can say that this is of course this is

insignificant because this is more than 0.05 right but at 90% confidence level this becomes significant right.

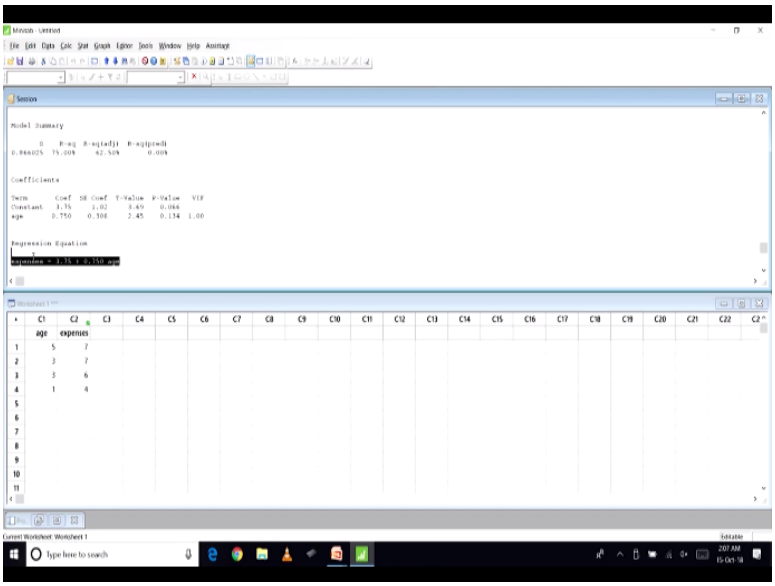
So this is your estimation line right or you can call it *y* is=0.375+0.75*C1, C1 is nothing but age of the truck right. Let us solve this question using Minitab.

(Refer Slide Time: 25:19)



So first of all age of the truck and you have got expenses, so age is 5 years, 3 years, 3, 1 so sample size is just 4 right. So for 5 years old truck, expenses 700 rupees, again 700, 600 and 400. Go to regression, go to again regression and fit regression model right, fit regression model.

(Refer Slide Time: 26:01)



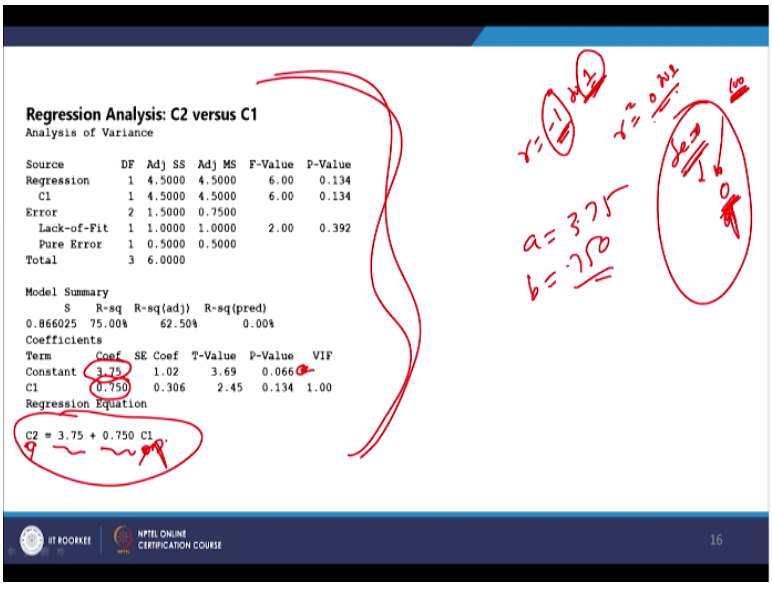
Now you have to be very careful. Now in this question you need to select dependent variable and independent variable. So we know that the repair expenses depends on age of the truck so repair expenses is dependent variable or response variable right. So expenses are nothing but responses right dependent variable, age is independent variable right. So you just click at okay and just see this answer.

You can move this side right, so let us look at these are other values, P values are given over here but let us look at this. Just look at this, standard error 0.86 right, this is what we calculated using formula right, r square 75% so what is your inference when r square is=75%, so you will say that 75% of variation in expenses is being explained by independent variable that is age of the truck right and standard error is 0.86 right.

Now let us look at P values over here, this is what we have seen earlier as well right. So these are P values and this is your regression line. So whatever is age of the truck just write over here and you will get the expenses. So this is how you can solve a question on Minitab if you are given a dependent variable and independent variable. So let me summarize what we did in today's class.

We have seen what is coefficient of determination, so coefficient of determination is nothing but the ratio of regression sum of square to total sum of square. In other words, it is explained variation and how do we measure it? It is SSE/n-2 whole under root right, so the value of r square cannot be negative, why because it is square.

(Refer Slide Time: 28:57)



So r can vary from -1 to +1 so we will say that there is perfect negative relationship between x and y so you will get r is = -1. If r is = +1, there is a positive strong relationship between x and y but r square will always be between 0 to 1 right, it can never be negative okay. Now we have seen standard error as well so standard error it measures the error term, we should have a model in such a way that the error term should be as low as possible, preferably 0.

When we say preferably 0, it means the model would explain 100% of variation independent variable and you really get such a situation, you will always have some error because you are measuring data right because of sampling right your sampling data. So because of that there will always be some error okay. Let us look at yeah we did compare standard errors so standard errors can be compared using just ac value for two different models.

So a model wherein ac is smaller than the other value would be a better model. So in next class, we will discuss some of the assumptions which you should keep in mind for solving a question on regression. For the time being, let me stop here. Thank you very much.