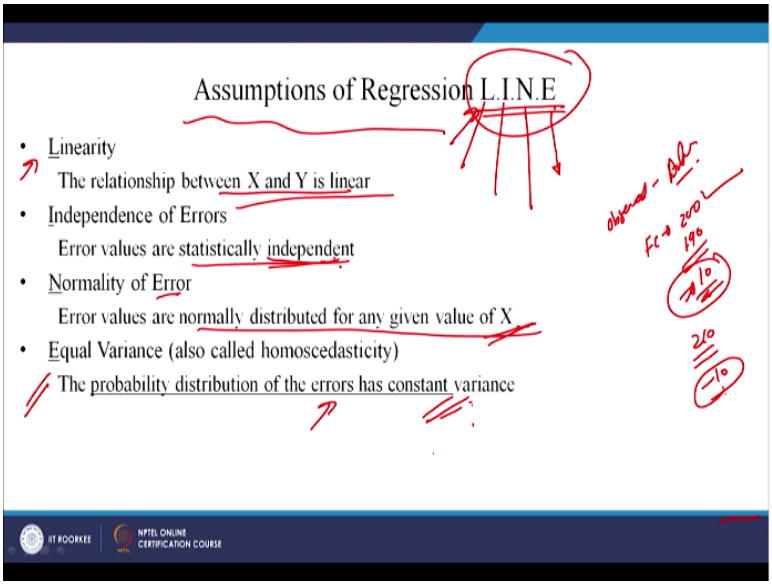
Business Statistics
Prof. M. K. Barua
Department of Management Studies
Indian Institute of Technology – Roorkee

Lecture – 58
Assumption of Regression

Hello friends. I welcome you all in this session. As you are aware, in previous session we were discussing about regression and correlation. We have seen that regression is basically association between two variables and correlation is nothing but it is a measure of strength of relationship between two variables. We did work out an example as well. In today's class, we are going to talk about assumptions which you should fulfill before applying regression to any data set.

(Refer Slide Time: 01:03)



So there are 4 basic assumptions, unless until you fulfill these assumptions you should not apply regression. Otherwise, you would be getting wrong results. In fact, you would get results without in fact looking for these assumptions but that could be wrong result right. So the first is linearity. So we are calling it as LINE, linearity independence of error term normality and equal variance right.

So first is there should be linear relationship between X and Y right. This is first and foremost assumption. If you do not have linearity of X and Y, if it is non-linear then you should convert it into linear function first and then apply regression. Independence of error terms so error values are statistically, they should be statistically independent. Error means it is

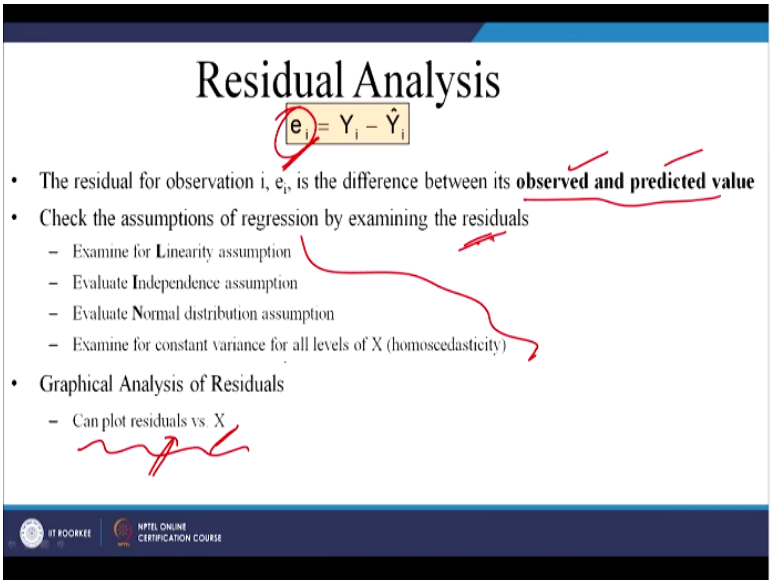
whatever is basically the difference between the observed value and actual value right, observed and actual right.

In fact, let me give an example of error term. Let us say if you forecast something, let us say the forecast of a particular product in let us say next month would be this much right and the actual demand turns out to be 90 right. So will say there is an error of 10 units right, is not it? So error 10. Now let us look at so this positive error let us say if your forecast is 100 units but actual sales comes out to be 110.

So the error would be -10 so error values are statistically independent right. So the error means what error it is a difference between actual and what you have predicted right. Normality of error right, so error values should normally be distributed for any given value of X, so we will see. In fact, we have seen several methods of checking normality of data; let us say normal probability plot, beat and normal distribution curve and so on right.

The fourth one is equal variance right, so the probability distribution of the error has constant variance. So these are 4 basic assumptions which you must fulfill whenever you solve any question using regression.

(Refer Slide Time: 03:56)

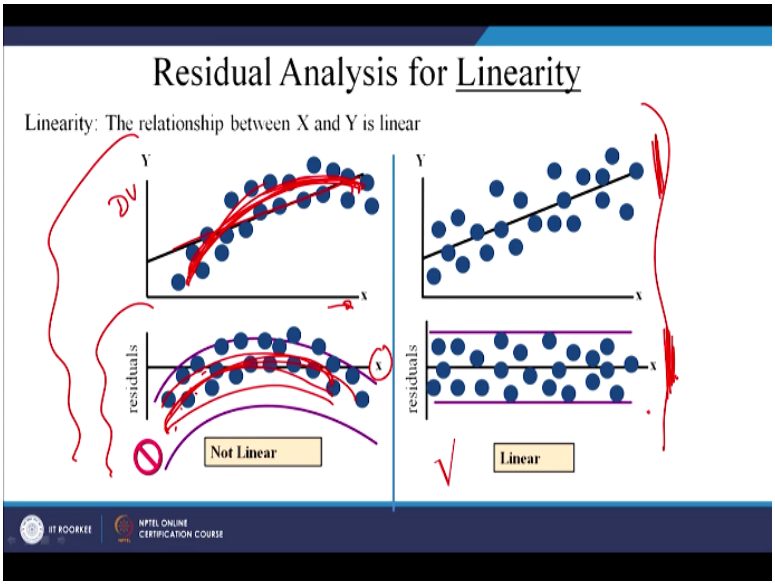


Let us look at residual analysis in fact residual analysis are this error terms will help you in knowing whether you are fulfilling those assumptions or not. So observed and predicted value. So the differences between them, it can be positive or it can be negative. So check the

assumption of regression by examining the residuals or errors. So these are 4 assumptions which we have talked about right.

And you can get little idea about linearity of X and Y just by having plot of X and residual values right. So we can graphically see, it is not only linearity but independence of error term normality and equal variance okay.

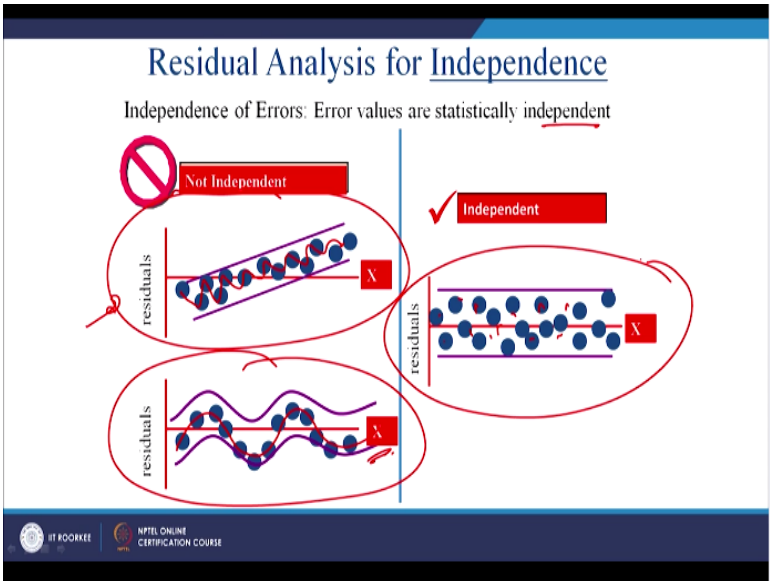
(Refer Slide Time: 04:50)



So this is how the relationship looks like, so you have got independent variable X, this is dependent variable Y, this is your if you see carefully this is not a linear relationship right, it is a curvilinear relationship and if you look at this it is relationship between X and residuals right. So you plot X on x-axis and residuals on y-axis. So if the residuals are like this, there is a pattern right or it is a nonlinear one right.

So it should not be like this, is not it? You should have linear data as well as the relationship between X and Y should be linear and the residual should be linear right with respect to X. So this is linearity assumption.

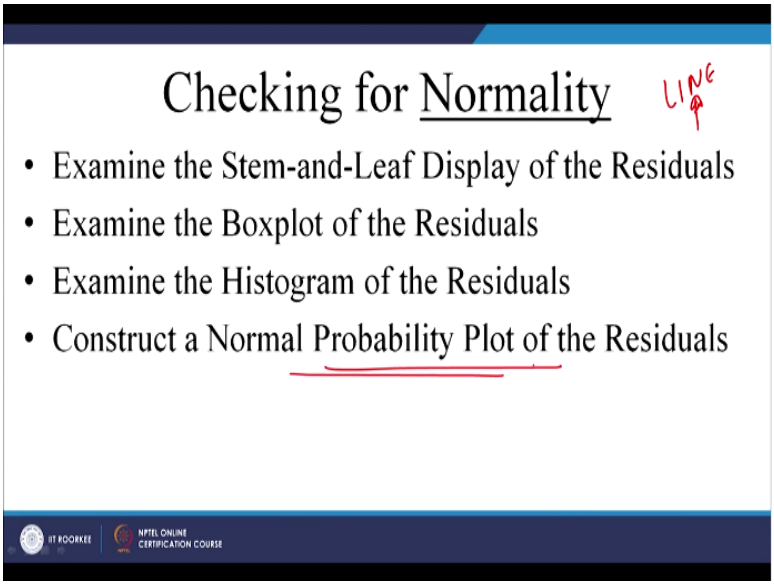
(Refer Slide Time: 05:51)



Independence of error terms so the error term should be independent of each other right, so let us look at this. So these are let us have one example, residuals versus X, again residual versus independent variable X. So if you look at carefully then there is a pattern over here, it not it. Up and down, up and down right similarly here, just see there is a pattern. So there should not be any pattern as far as residuals and X plot is concerned right.

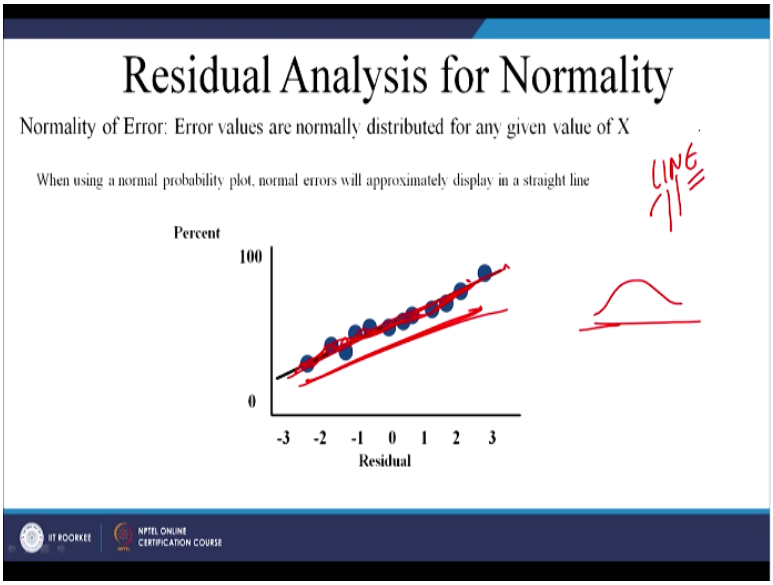
If there is a pattern, it means they are dependent on each other, so error terms should be independent of each other just like this. So here you are getting different values right. There is no pattern as such, so this is independence of error terms, second assumption.

(Refer Slide Time: 07:02)



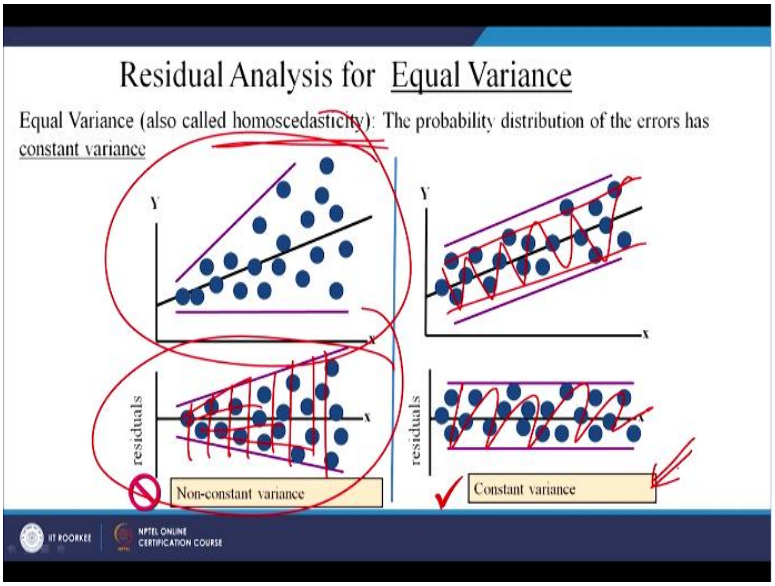
The third one is as I said LINE, so the third one is normality of data. So normality of data can be seen by multiple ways which we have seen in very few initial classes. So one of them is you know you can construct a normal probability plot.

(Refer Slide Time: 07:30)



So this is how you plot a normal probability plot. So if all these points are on the straight line, this is nothing but normal data right. In fact, this is nothing but your normal distribution right. So it is a bell-shaped curve but since you have taken log normal so that is why this is the shape of distribution. So we have seen 3 assumptions, LIN right linearity, independence of error terms, normality and the fourth is error term should be equally distributed right.

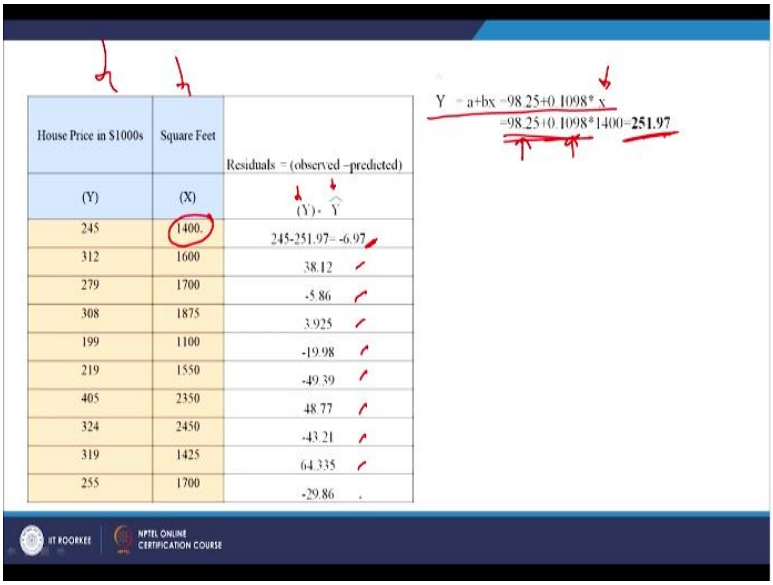
(Refer Slide Time: 08:06)



So equal variance, it is also known as homoscedasticity. So you should have homoscedasticity not the heteroscedasticity. So these are your error terms okay. So constant

variance is there and just see even in this plot as well as in this plot you have got this error terms are whereas these terms are spreading right. So this should not be there. It should be just equal, so this variance is constant here right, is not it? Similarly, here so the variance should be constant as far as error terms are concerned.

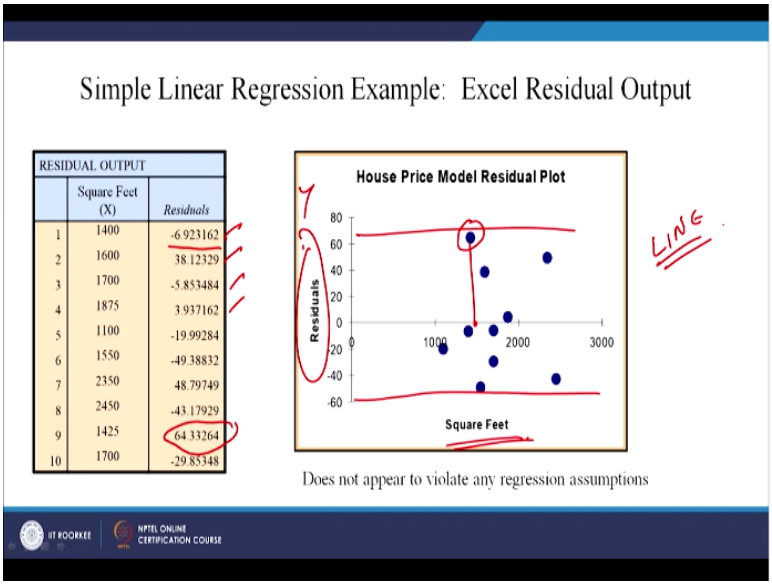
(Refer Slide Time: 09:02)



So let us find out those LINE assumptions and we will see how to get the residuals. So this is the question which we have worked out. So house price and the size of the plot and this was our line of best fit or prediction line right. So a is = 98.25 and b is 0.1098, so let us say what is the value of y or predicted value when X is =1400 right. So in this equation just put all X values. So when you put X is =100, this Y hat really becomes 259.

So y - y is residuals is not it? So residual over here is -6.97, this is plus, this is minus, this is plus, minus, minus, plus, minus, plus. So if you look at carefully there is a pattern, you have got plus, minus, plus, minus is not it? So these are residuals right and we will draw these residuals against X values okay.

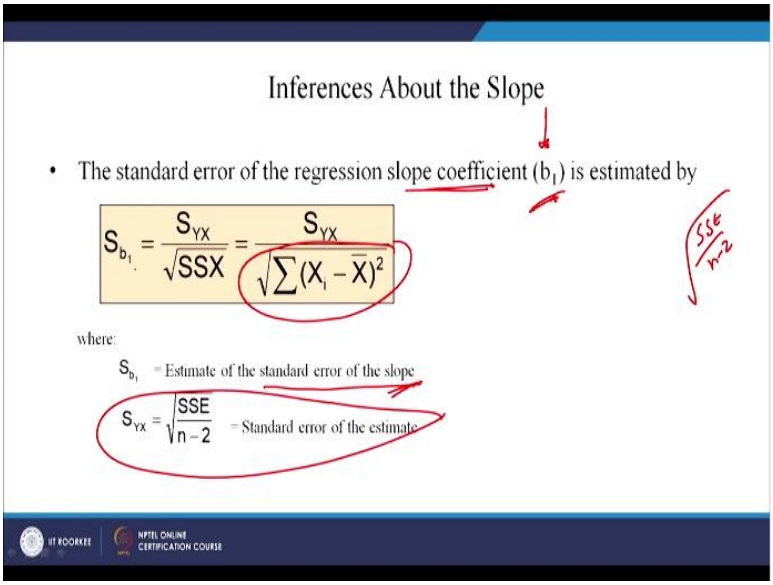
(Refer Slide Time: 10:25)



So this is how we have drawn these residuals which are here on y-axis and on x-axis you have got square feet right. So for X is =1400 you have got -6.92. So in fact you can see over here let us say the highest one is 64 right, 64.33 this is the one, 64.33 and for a given value of X is =1425 right, so this is 1425. So you can draw all these 10 points. Now if you look at all these 10 points, they are quite randomly distributed on plot.

We are not saying that even you can say once again here, so minus, plus, minus, plus. There is no such pattern right. If there is one negative value, the other is positive then negative, positive, negative to negative, positive, negative, positive and so on. So there is no such pattern over here. So will say that in fact you can see that there is equal variance so linearity is there, independence of error term is there, normality is there and error terms are having equal variance right. So we fulfill all these 4 assumptions right.

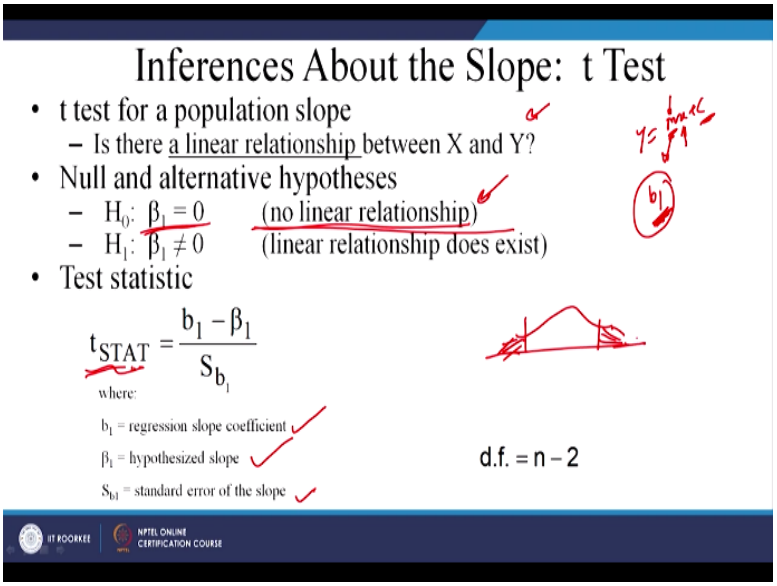
(Refer Slide Time: 11:57)



Now let us look at inference about slope. So we have seen standard error, standard error is nothing but it was sum of square error divided by n-2 under root right. So standard error, so will have the standard error of regress slope coefficient. So we have got this slope of independent variable b1, so it will have some standard error as well right. So standard error can be calculated like this, so this is estimate of standard error of the slope.

So you know the standard error very well, you just divide it by this value, you will get standard error of slope right.

(Refer Slide Time: 12:54)



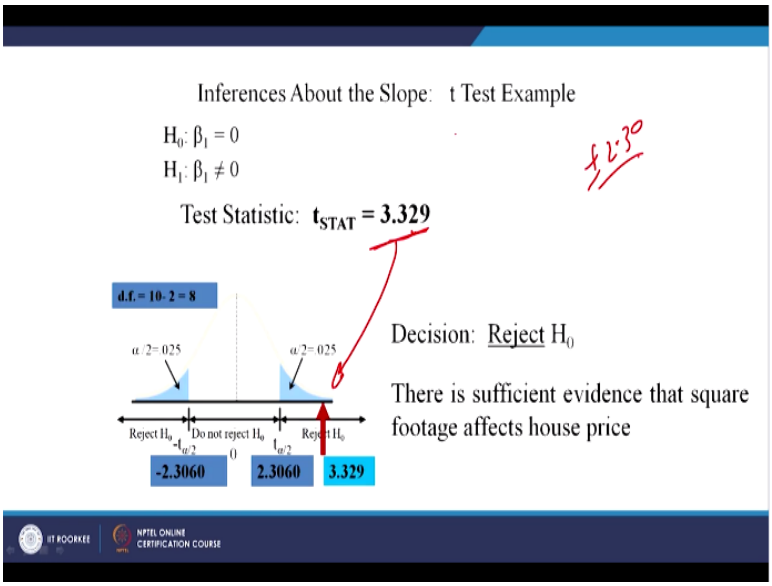
So inference about slope, now we want to know, you know we have seen y is=mx+c, is not it? This is constant, this is slope m is slope and this is your independent variable right. Now in previous example, we have seen that the square feet, the size of the plot has effect on price

of the house, so we want to see whether this slope this m value or let us say b1 value or in case of multiple regression you can have b1, b2, b3 and so on.

So all these values have got significant effect on y or not, so is this effect is statistically significant or it is just by chance, so we have to check this relationship. So we say that the null hypothesis is or in other words we want to know is there any relationship between linear relationship between X and Y. So beta is 0, will say that there is no relationship. So if you reject null hypothesis then will say there is relationship right.

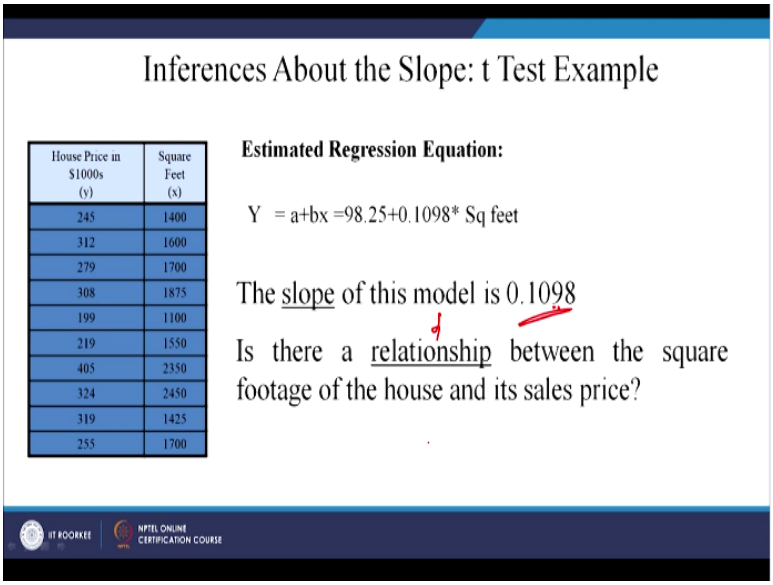
So initially we are saying there is no relationship right, so you can have just test statistics, so t statistics, you have got regression coefficient, hypothesized slope and standard error of slope at appropriate degrees of freedom. So calculate t-statistics value and since this is a case of two tailed test you will have two rejection regions, is not it? This is rejection region and this is second rejection region right. So this is first rejection region.

(Refer Slide Time: 14:49)



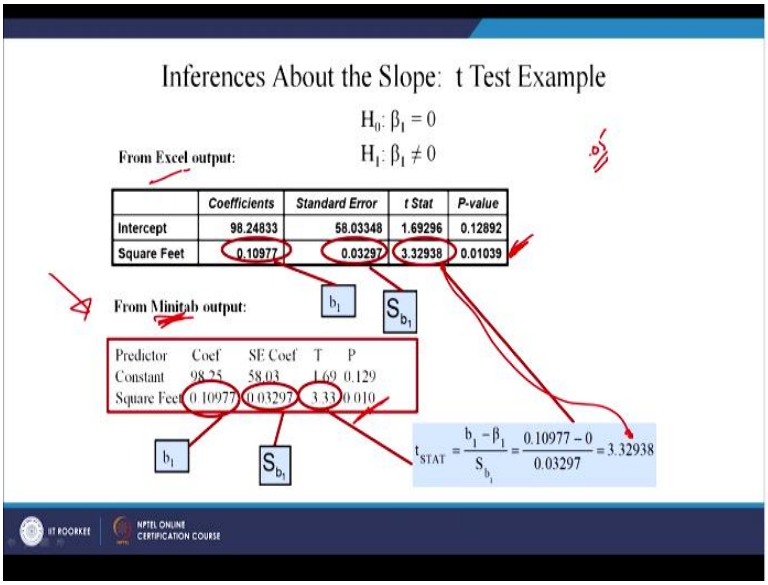
Now let us look at the output of the previous question and if you look at the t-statistics which we have calculated is 3.329 and the t-table value is ±.3 so ± 2.3 right. So calculated t-value is somewhere here right, so will reject null hypothesis. So null hypothesis is getting rejected right. So when we say null hypothesis is rejected, it means what it means there is linear relationship between X and Y right.

(Refer Slide Time: 15:29)



Now let us look at the regression line or estimated regression line for this question. So slope is this, is not it? So we have checked is there any relationship between X and Y or not. So though we have rejected the null hypothesis and we have said that there is relationship.

(Refer Slide Time: 15:51)



The same information can be seen over here in output table either you are getting output from excel or from Minitab right. So just look at the P value, so P value is <0.05 so will reject the null hypothesis right. So we will say that there is relationship. In fact, the calculated t-value is available in table as well. So this is there 3.3293. Now standard error is there, of course all these values are there in this table.

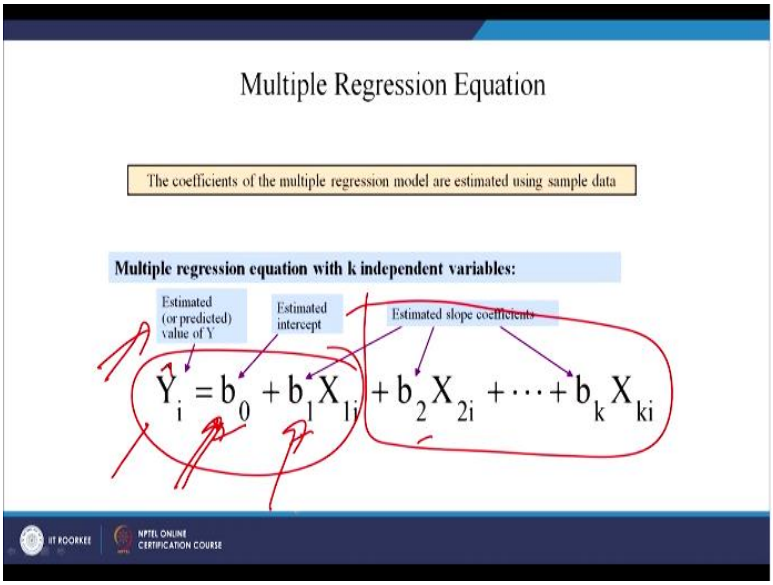
And in fact output from Minitab also can be seen. Now P value again here is <0.05. So will say that will reject null hypothesis and there is relationship between X and Y. So far we have

seen linear regression is called simple linear regression or bivariate regression because there were only two variables, one was dependent and the other one was independent variable and both of them were metric in nature. So we have seen simple linear regression right.

(Refer Slide Time: 17:10)

You may have a situation where the independent variables are more than 1, so in that case you need to apply something called multiple regression. So let us look at what is multiple regression.

(Refer Slide Time: 17:25)

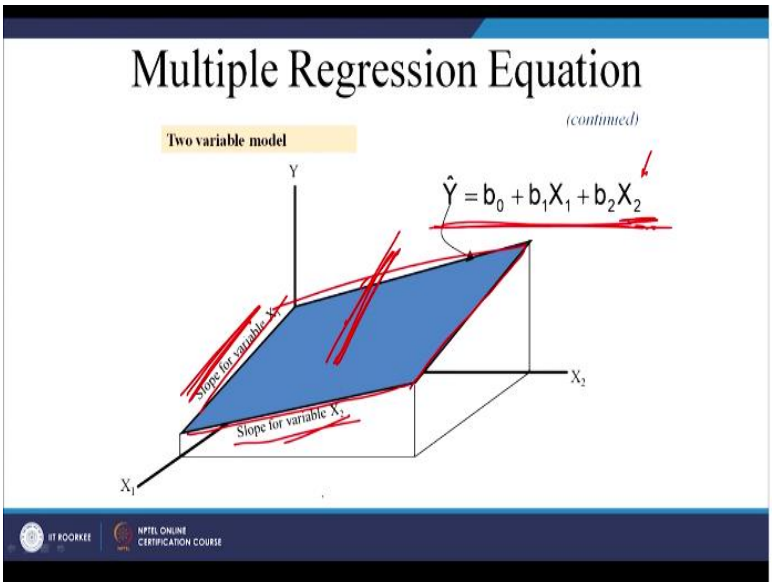


So it is very similar to simple linear regression, only the difference is you have got more than one independent variables right. So here you have got k independent variables, again you have got error term, so you have got plus and minus both in error term right. So this is so all

these are your independent variables and these are different slopes right, beta 1, beta 2, beta k and this is your intercept right.

What is intercept? Intercept is that value when all independent variables are 0 right. So this is how you can have the line of best fit or the estimation line. So *y* is=b0, b1, b2 and so on right which is similar to simple linear regression. Only the thing is you have got these extra terms over here right. Earlier we had only these three right, dependent variable, intercept and slope right.

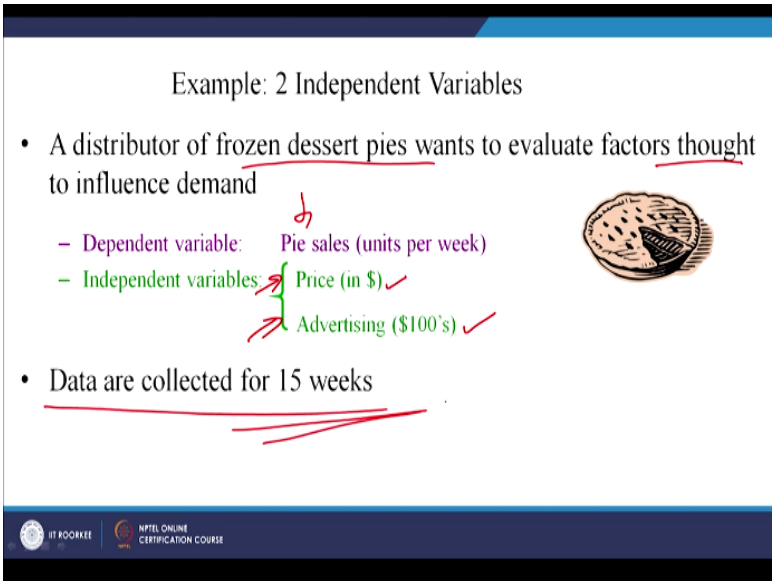
(Refer Slide Time: 18:36)



Now this is how the multiple regression or the plane of estimation, I will not call it line of estimation right. So you are getting a plane over here right, it is a two-dimensional plane right, is not it? Because there are two independent variables, so there are two slopes right. Slope for X2 and this is slope for X1. So when there was one independent variable, it was a line, two independent variables a plane.

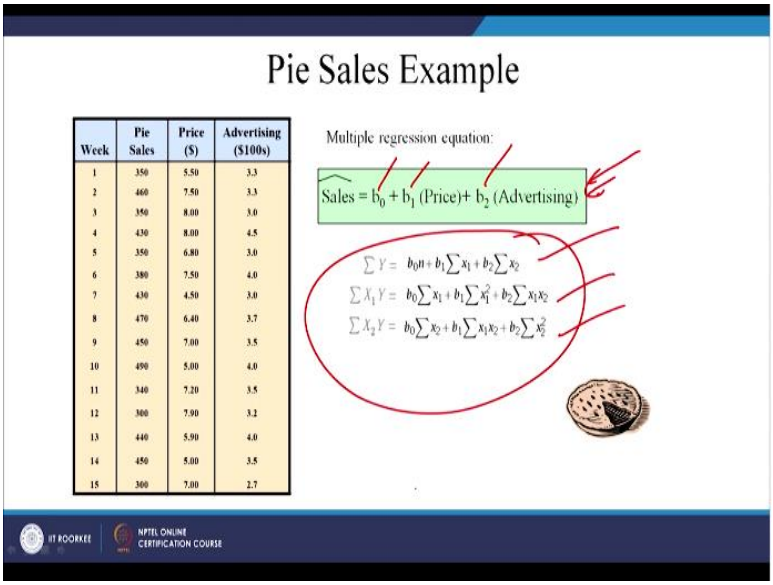
Suppose if there are 3 independent variables, what would be the shape of this graph or this plot? It will not remain a plane right, it would be a three-dimensional figure is not it? It could be cuboid rather than a plane if there are three independent variables.

(Refer Slide Time: 19:31)



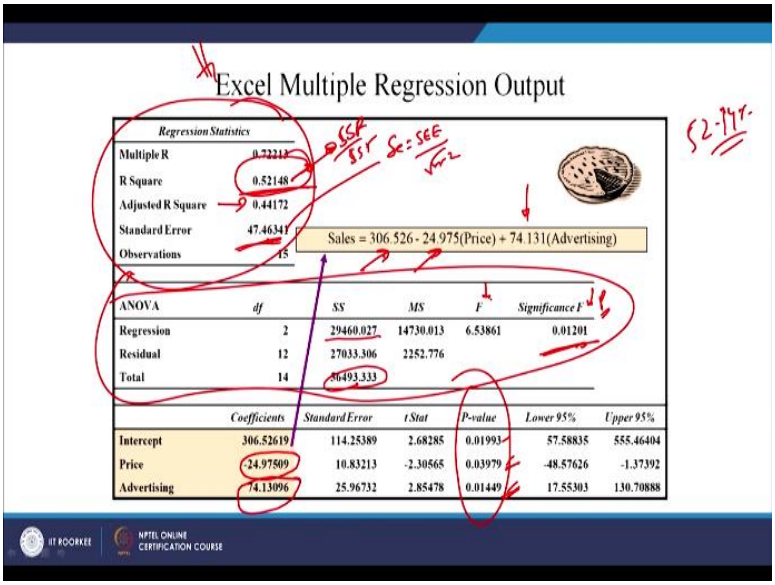
So let us look at this question wherein there are two independent variables. A distributor of frozen dessert pies wants to evaluate factors thought to influence demand. So the first factor is price, the second one is advertising. In other words, we want to know how these two independent variables namely price and advertising affecting pie sales right. So data were collected for last 15 weeks.

(Refer Slide Time: 20:13)



And regression was done so you can in fact if you look at this particular question, you can solve this question manually using these 3 equations because this is a question where in this equation there are 3 unknown variables right b0, b1 and b2. So there are 3 unknown and 3 equations. So these are nothing but simultaneous equations you can solve for b0, b1 and b2.

(Refer Slide Time: 20:48)

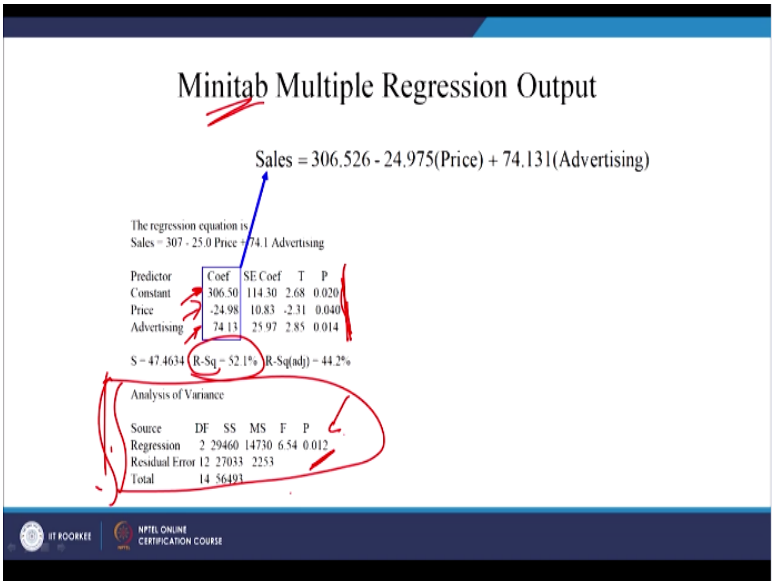


However, you can solve this question using excel or Minitab as well right. So this is the output from excel. So if you look at the first part of this table where you get the descriptive statistics, so r square is 52.14% right. We will talk about adjusted r square little later. Standard error 47.46. How did we get standard error? Standard error is Se is=sum of error term/n-2 right and how did you get this r square?

R square is the regression sum of square right divided by total sum of square right SST, so this divided by this value you will get r square. So your estimation line is this, sales is=this is your intercept -24.97, this is coefficient of first independent variable, this is coefficient of second independent variable, so 74.13 and if you look at P values then all these 3 are significant, so will say that price and advertising as well as intercept especially price and advertising are significantly affecting dependent variable.

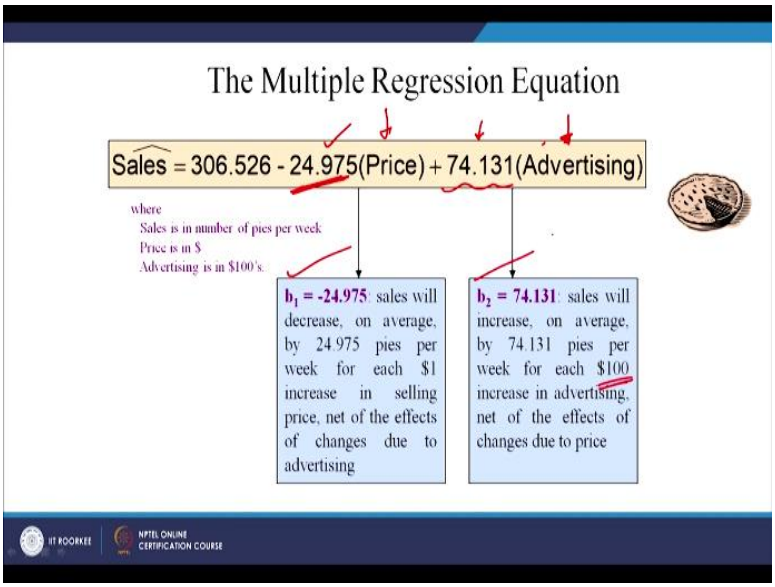
Look at this r square again, it is just 52.14% right. so will say that 52.14% variance independent variable is explained by these two independent variables. Look at ANOVA table, this part the middle part of the table ANOVA table, so this is nothing but the true value of F right so this is again <0.05 so will say that the model is a significant model right.

(Refer Slide Time: 22:58)



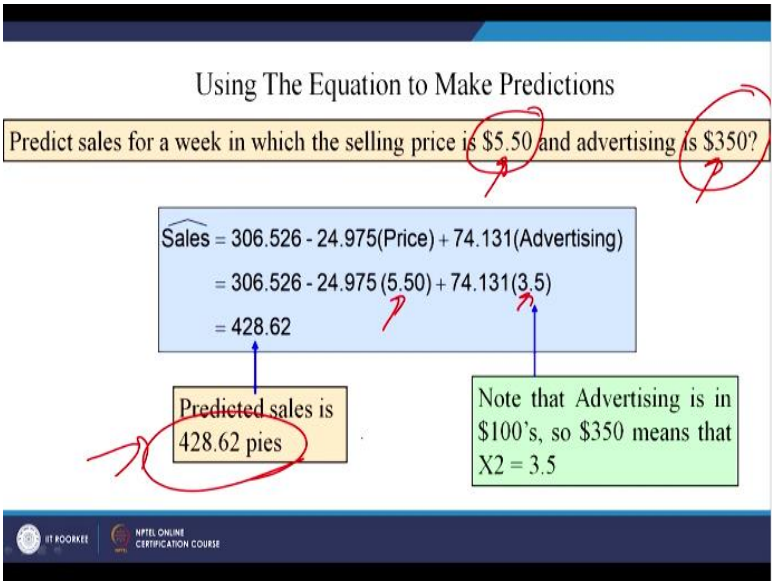
So this is output from Minitab, will get the same output right, so constant or the intercept, price and advertising coefficients right. P value is again all are <0.05, r square 52.1 same as this r square, then ANOVA table of course you can look at the how overall model is significant right. So this P is again it is <0.05 right.

(Refer Slide Time: 23:34)



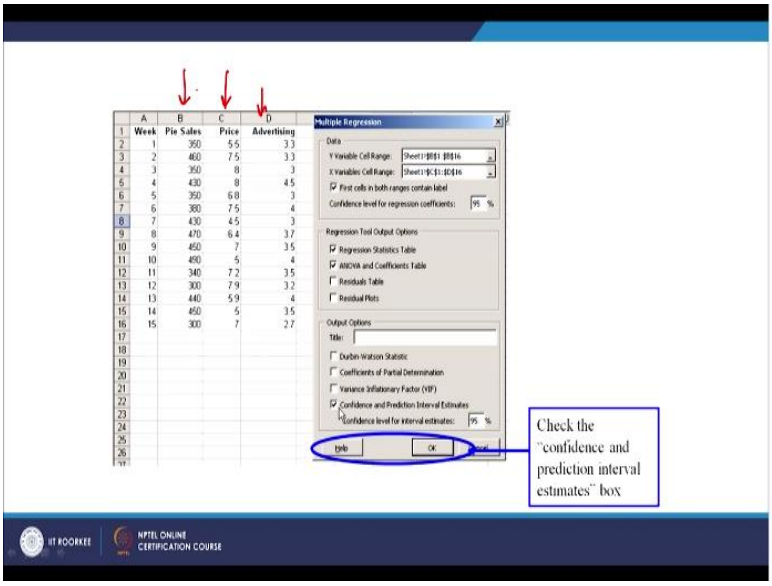
So you can solve this question using Minitab but before this let me explain the meaning of these terms right. So this is -24.97 so sales will come down by this much unit if you increase price of pie or this product by one minute. If you increase let us say advertisement, expenditure by 100 dollars then this much sales will increase, so this is the meaning of b1 and b2 right.

(Refer Slide Time: 24:12)



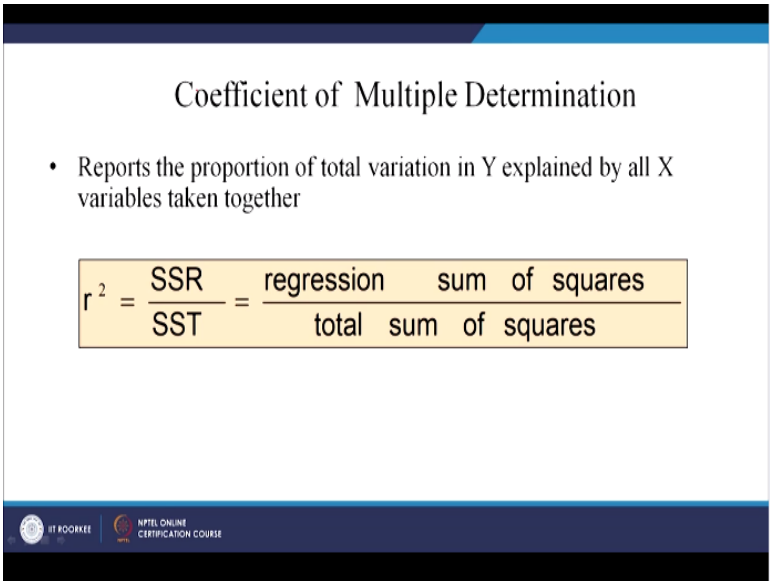
So predicted sales let us say if I ask you a question, what would be the sales when price is this much and advertising expenditure is this much, so put X1 is equal to this and X2 is equal to this, so this is X1 and this is X2 right. So will say that if the price is equal to this and advertising expenditure is equal to this, the predicted sales is 428.62 pies okay. So now we will work out this example.

(Refer Slide Time: 24:48)



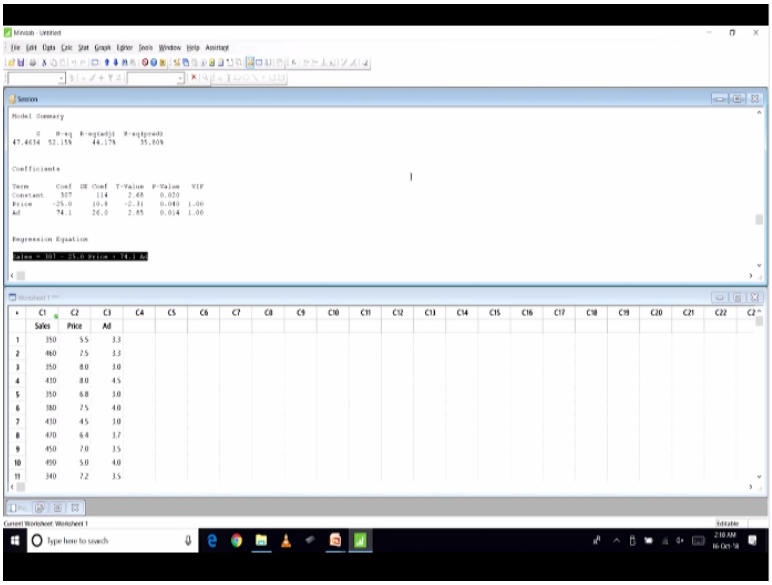
And in fact you can use excel so it is very simple, you just put input, these variables as you know predictor variables and this as dependent variables right.

(Refer Slide Time: 25:09)



So before moving onto next slide let me tell you how to solve this question using Minitab right. So this is the question we will solve using Minitab right. So pie sales is dependent variable this is X1, this is X2 right, two independent variables.

(Refer Slide Time: 25:26)



So let us look at this. So this is sales, this is price and this is advertising expenditure right, so 350, 460, 350, 430, 350, 380, 430, 470, 450 and 490. So there are some more data points 450 and 300. So there are total 15 data points. Similarly, for price now this 6 number, you have to enter it number 1, so just go up yeah so this is price 5.5, 7.5, 8, 6.8, 7.5, 4.5, 6.4, 7.0. So this is how you should enter all the prices 7.9, 5.9, 5 and 7.

Now you need to enter the advertising expenditure right and all these values are in terms of 100 right so it is 330, 100 no it is not 3.3, so the ninth one is 3.5. We are supposed to enter

data carefully otherwise there will be some difference in answer right. What I have shown you would be different and what we would be getting will be different. So we are supposed to enter data for this advertising expenditure.

Now go to stat, basic regression, yeah regression just regression, fit regression models right. So response is sales right sales then you have got continuous predictor right. So just click okay and let us look at output. So all these are P values right. So less than 0.05 will say these two are significant independent variables. Let us look at r value is 52.15%, then standard error is 47.46% and this is your line of best fit or estimation line right.

So 307-25+74.1 which is exactly the same which you are getting over here right, this is 24.975 which is as good as 25. So this is how you can solve the question on multiple regression and if you have got let us say more than two independent variables, then we will have one more independent variable noted over here right with its positive or negative sign depending upon what is that independent variable.

So with this let me stop here. In next class, we will work out couple of examples on multiple regression and will see how to include a categorical independent variable in any problem. Thank you very much.