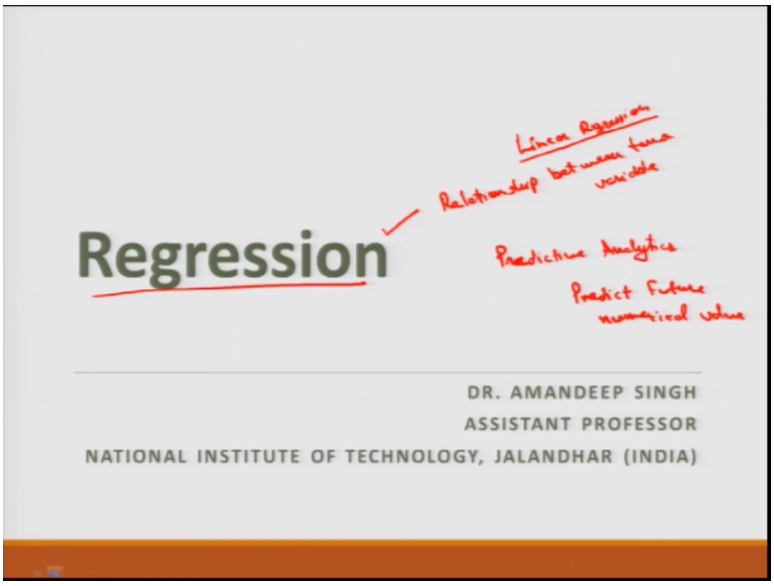
Practitioners Course in Descriptive, Predictive and Prescriptive Analytics

Prof. Deepu Philip
Dr. Amandeep Singh Oberoi
Mr. Sanjeev Newar
Department of Industrial & Management Engineering
National Institute of Technology, Jalandhar
Indian Institute of Technology, Kanpur

Lecture – 22
Regression

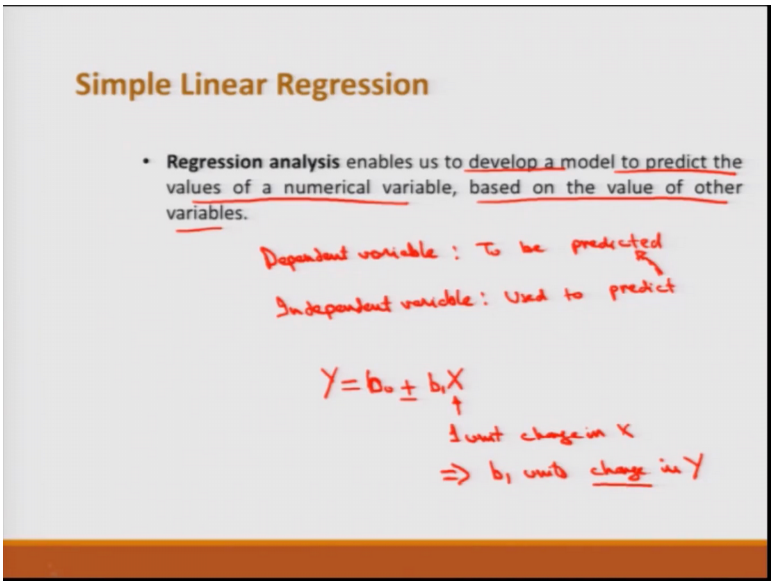
So, welcome back to the course on analytics here. So, we have tried study the descriptive, predictive and prescriptive analytics. So, I will cover regression in this session.

(Refer Slide Time: 00:25)



So, what is Regression? Regression is a statistical method that analysis and finds relationship between two variables. If I say two variables this is linear regression here also we can have multiple regression the when there is one dependent variable and more than one independent variables. So, this is used in predictive analytics to predict future numerical value of a variable, this predict future numerical value.

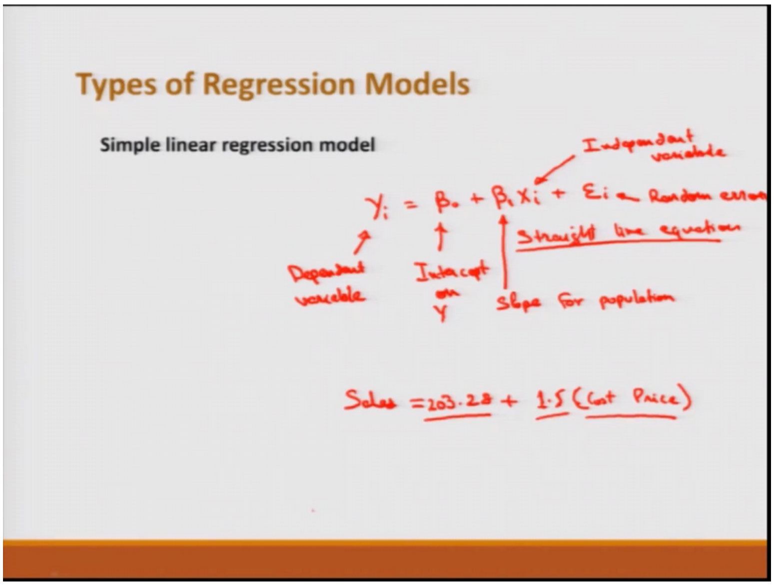
(Refer Slide Time: 01:31)



So, let us the see how does regression work, simple linear regression. A simple regression enables us to develop a model to predict the values of a numerical variable based on the value of other variables. So, in regression analysis the variables we wish to predict is called dependent variable and the variable used to make the prediction is call independent variable to be predicted used to predict this variable.

So, regression analysis allows us to identify the type of mathematical relationship that exists between a dependent variable let me say dependent variable is Y and an independent variable X and this quantify the effect that changes in the independent variable that it has on the dependent variable. It says that Y is b naught plus b 1 into X that is with 1 unit change in X 1 unit change in X implies b 1 units change in Y, this sign here might be plus or minus I have put the word change not increase or decrease the change can be increase or decrease. So, this is a simple linear regression model.

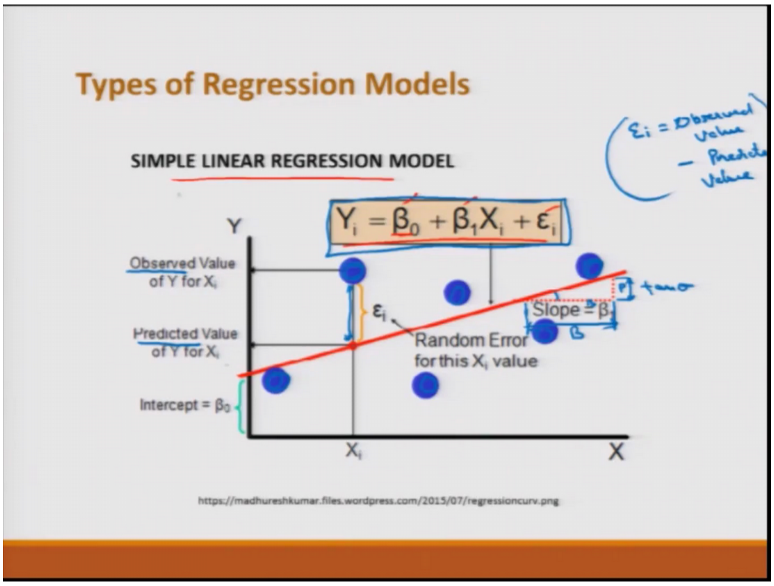
(Refer Slide Time: 03:40)



As I just put it Y i is equal to beta naught plus beta 1, X i plus the error term. The simplest relationship consist of a straight line of linear relationship according to this equation of line. This is a straight line equation with error term where this beta naught is the intercept that is intercept on Y this beta one is the slope of line slope for population Y i is my dependent variable X i is the independent variable and epsilon i is my random error. So, let us see a few details of regression model linear regression model here.

So, an example here can be the sales if my Y i sales is equal to maybe I put some number here 3.28 plus or let me put a bigger number 203.28 plus 1.5 times 1.5 times the cost price or this can be maybe twice. So, here the cost price is the independent variable we are trying to predict the sales. So, this is the intercept this minimum this much of sale would happen, but with one unit change in cost price 1.5 times the sales is increasing. So, this is a regression model.

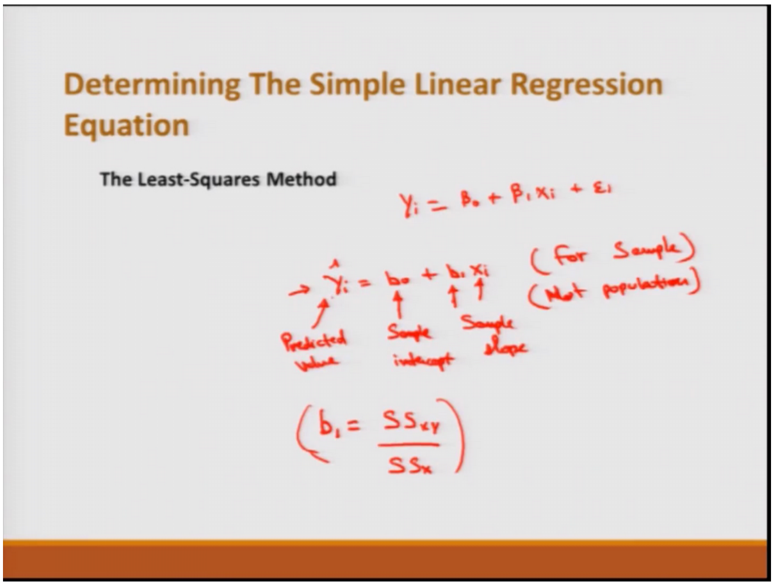
(Refer Slide Time: 06:41)



So, let us see how is this seen in a graph. So, this is a graph of simple linear regression we have this relation dependent variable, independent variable, error term, slope and intercept. So, this is intercept beta naught, this is intercept here and this is slope beta 1, slope is actually tan theta that is perpendicular upon base, this length by this length, this is slope.

So, these are my observed values these are my observed values. So, this difference from the predicted model, this is the model regression model this difference from the predicted model that is epsilon i is known as random error. So, this is the observed value, this is the predicted value. Observed value minus predicted value is my error. So, epsilon i is equal to observed value minus predicted value.

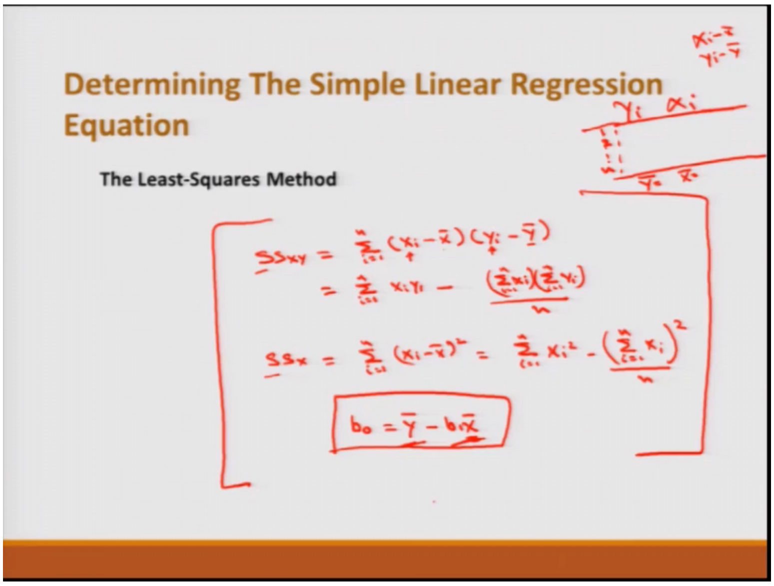
(Refer Slide Time: 08:35)



So, next is how do we determine the simple linear regression equation. The least-squares method is one. We have this equation Y i is equal to beta naught plus beta 1 into X i plus error. The least square method says that the predicted value that is Y i cap is equal to b naught plus b 1 X i please note I have put b here not beta because this is for sample not population.

So, here Y i cap is the predicted value and this is X i is the value of observation b i is the sample intercept and b 1 is the sample slope. Now, what happens when we use least care method this b 1 and b 2 calculated as b 1 is equal to SS xy by SS x. What is SS? SS is sum of squares. So, what are these let us see.

(Refer Slide Time: 10:24)

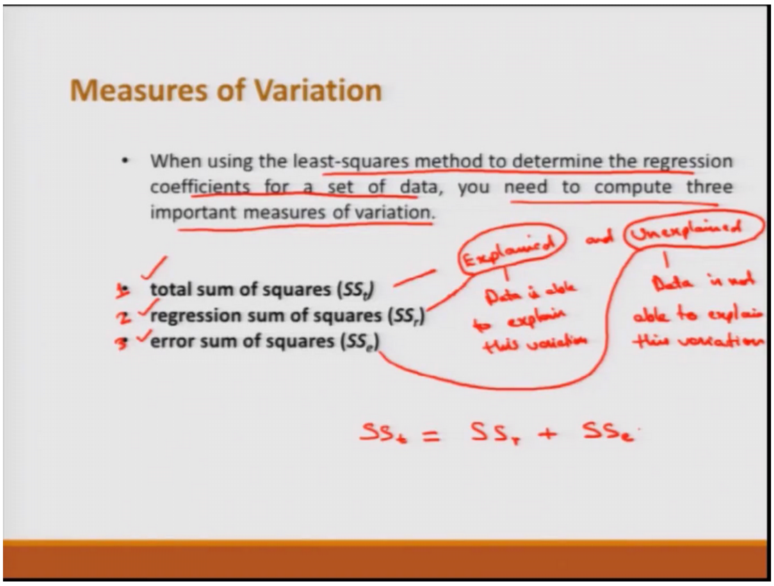


SS XY is summation i is equal to 1 to n X i minus X bar and Y i minus Y bar. So, this is actually the observed value minus mean and the observed value minus mean for X and Y two variables here. So, this comes down to summation i is equal to 1 to n X i, Y i minus summation X i summation Y i in both cases i is equal to 1 to n, i is equal to 1 to n by n and SS x is the sum of squares for variable X only that is X i minus X bar is square i is equal to 1 to n which is equal to i is equal to 1 to n, X i square minus summation i is equal to 1 to n X i whole square by n.

So, we will see how do we look this in on the graph then beta naught can be calculated as Y bar minus b 1 X bar, this is actually for the sample we have these values we have the values of X i we have the values of Y i we know what is n 1 to n values are there 1, 2 and so on up to n this tables here. So, we can obtain Y bar here X bar here that is the average value or mean, then we can get these relations X i minus X bar all in the table X i minus X bar Y i minus Y bar and we can calculate SS XY and SS X sum of squares for XY sum of squares for X.

So, what are the this is the total sum of squares and the explained sum of squares. So, we can find the value of b 1 from here that is the slope and how do we calculate b naught b naught is we know the value of b one we know the value of X bar and Y bar we will find the value of b naught. This is how we draw the line for sample then we get this equation for the sample.

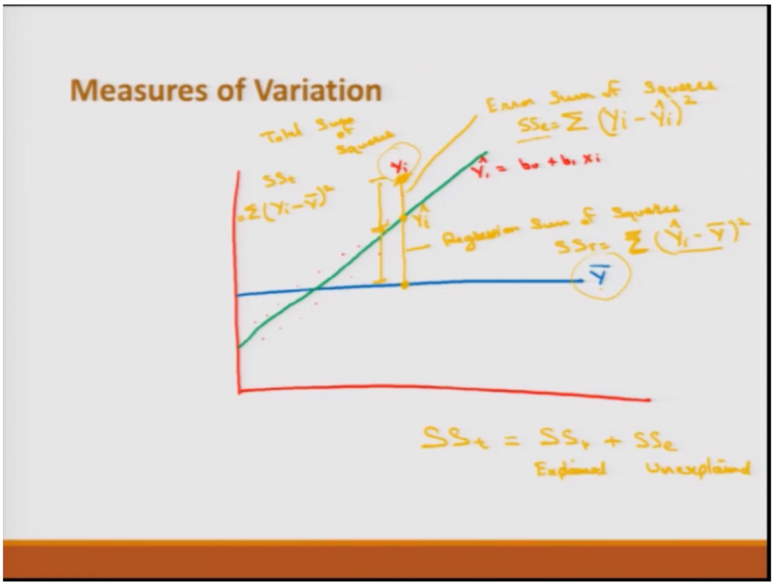
(Refer Slide Time: 10:54)



So, measures of variation here are total sum of squares, regression sum of squares and error sum of squares. So, the first measure is total sum of squares when using least square methods to determine the regression coefficients for a set of data one need to compute these important measures of variation; number 1 is total sum of squares, number 2 regression sum of squares, number 3 is error sum of squares the total sum of squares is the total that is the sum of explained and unexplained variations. So, this total can be divided into two parts explained and unexplained where explained implies data is able to explain this, but something that is not unexplained that is due to error that is unexplained data is not able to explain this variation.

So, this explained variation is called the regression sum of squares, this is regression sum of squares here and the unexplained variation is called error sum of squares. So, we can say that SS total is equal to sum of the regression and error sum of squares.

(Refer Slide Time: 16:21)



So, if I try to plot this on a graph we can see that we have the average value Y bar here and our model is like this. This is the predicted model that is Y i cap this is the model Y i cap is equal to b naught plus b 1 into X i.

So, let us see how is this represented on this graph this relation. Now, let me put one value Y i here one observed Y i. This Y i is distant from the model ideally this all these points the observed points should be very much close to the model. The closer are the points in the model the lesser are the error value that is the model is trying to explain very well the data if this values are large; that means the error is large. So, this actually the error so, this Y i distance from this model this line is error this is error sum of squares.

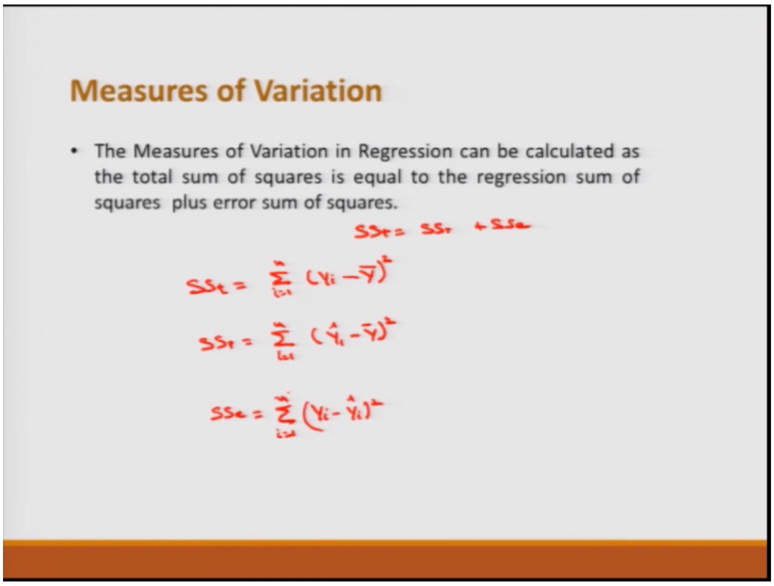
Next, we have the distance between Y bar and the model, Y bar and the model here is the regression sum of squares that is the average value of the data is Y bar my regression model is giving the predicted value Y i cap, this is the value Y i cap here predicted value. The difference between Y i cap and Y i bar it is Y i cap minus Y bar square sum of sum of the square this difference Y i minus Y bar difference square sum of summation of this squares that is my sum of squares due to regression.

Now, my predicted value is Y i cap and my observed value is Y i this comes down to the error sum of squares which is Y i minus Y i cap then sum of squares this is sum of squares due to error. So, if you see the regression sum of squares and error sum of squares makes the total sum of squares here. So, this Y i cap this length plus this length makes the total sum of squares that is Y i this Y i minus Y bar this is Y i minus Y bar sum of squares this is my SS total this is equal to or this is known as total sum of squares.

The whole idea in regression is to predict the value to predict the value of the dependent variable to find the model that is closest to the available data. So, the thing here is that the error should be minimum the lesser the error is the closer the model is to the original data. So, this value should be minimum SS error for the model to fit good SS error should be this.

So, again I will put SS total is equal to SS explained that SS regression plus SS error when I say SS surface is sum of squares when I say SS r it is sum of squares for regression that is explained sum of squares, one is a SS e it is sum of squares for error that is the unexplained sum of squares. So, this is explained by the model, this is unexplained by the model.

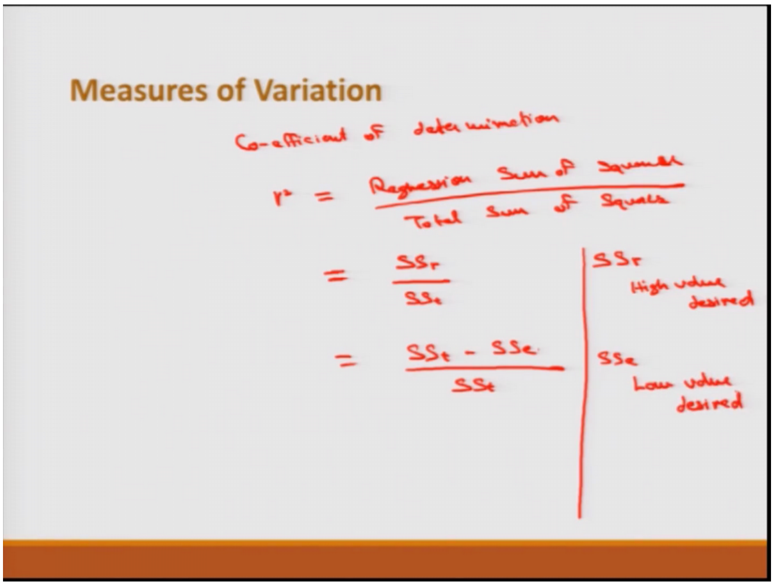
(Refer Slide Time: 22:14)



So, the measures of variation in regression can be calculated as total sum of squares is equal to the regression sum of squares plus error sum of squares putting the same thing in a statement here. So, the total sum of squares is mentioned in the graph here that is Y i minus Y bar square; that means, this SS total I am putting that again here SS total that is sum of the square of the observed value minus mean square i is equal to 1 to n and sum of squares regression is the predicted value that is Y i cap minus Y bar i is equal to 1 to n and sum of squares for error is the unexplained variation that is a Y i minus Y i cap sum of squares.

So, how does this help us? We need to find the coefficient of determination coefficient of determination.

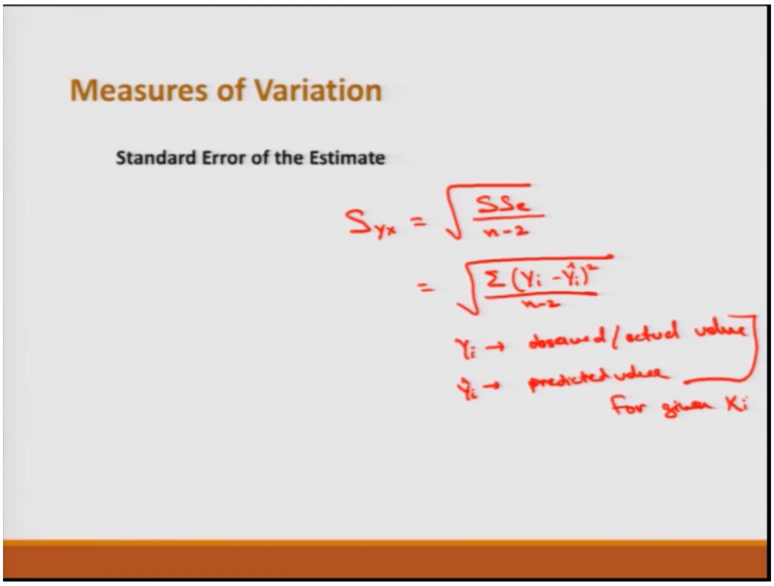
(Refer Slide Time: 23:46)



It is also known as the regression coefficient, it is r square.

Now, r square is the coefficient of determination which is equal to the regression sum of squares divided by total sum of squares that is regression sum of squares by total sum of squares that is equal to SS due to regression over SS total, which implies the higher value of regression sum of squares would lead to higher value of r square that would give the r coefficient and obviously, this regression sum of square is equal to total sum of square minus error sum of squares by SS total which implies the lower value of error sum of squares would lead to higher value of r squares so; that means, SS r higher value desired and SS e low value desired.

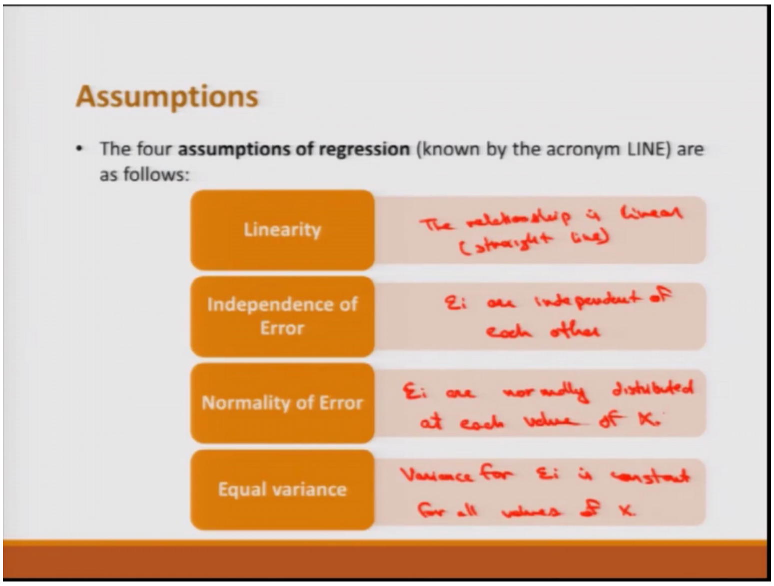
(Refer Slide Time: 25:33)



So, next is standard error of the estimate the standard error s for the estimate of Y due to X is equal to square root of these sum of squares for error that is unexplained variation divided by degrees of freedom, degrees of freedom here is n minus 2 because two parameters are known here. So, that is subtracted from n. So, degrees of freedom comes down to n minus 2. So, this is the square root is taken because this was sum of squares and we just need to find the error. So, this if I put the relations here of sum of squares that is equal to sum of squares of Y i minus Y cap Y cap squares by n minus 2.

So, this Y i is actual value or observed value again this is observed or the actual value Y i cap is predicted value the i is this is for given X i for given value of the independent variable.

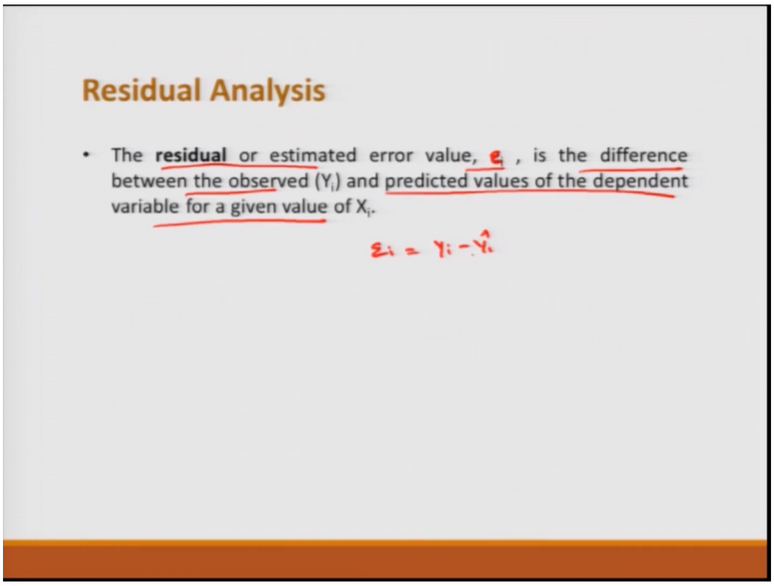
(Refer Slide Time: 27:22)



So, there are certain assumptions in regression number one is linearity. This states that the relationship between the variables is linear the relationship is linear, that is straight line. However, in realistic situation these relationships are not linear and these are also sometimes not additive in nature like in the example; we were saying sales were dependent upon the cost sales second variable here can be in the multiple regression. The second variable can be incomes here that depend upon the income and sales are depend upon the willingness to pay these are all variables we are just adding these variables. But, in real life these are not additive, so, we are just assuming that.

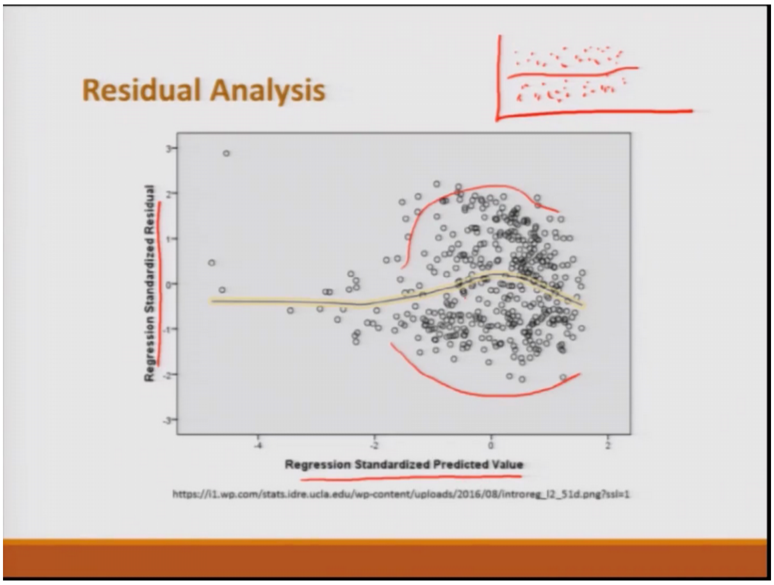
Then independence of error is the second assumption here which says that the errors epsilon i are independent of each other. So, we try to draw the plot for this for independent of error and also for the normality check this say that the normality of the error. Normality of error tells that the errors are normally distributed at each value of X. Then last is equal variance that is the variance for error is constant for all values of X, variance for E i is constant for all values of X, here at each value. So, these four assumptions if are met the model is good.

(Refer Slide Time: 29:44)



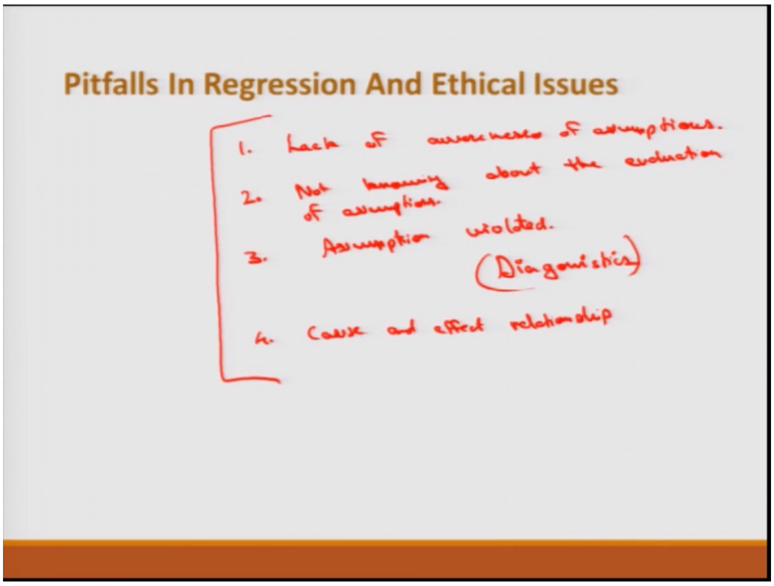
The next is residual analysis the residual or estimated error value is the difference between the observed value and the predicted value of the variable. So, the residual graphically a residual appears on a scatterplot as the vertical distance between the observed value and a prediction value. Now, we have put the notation e here it it is a epsilon i is equal to Y i minus Y i cap.

(Refer Slide Time: 30:21)



So, this residuals are to be distributed like this. So, what is happening here. So, this is telling that there all the all have equal variance. So, this is standardised regression residual versus standardized prediction value. So, this is distributed evenly along this line, distributed evenly. So, though the ideal distribution would be like this one if this is my model ideal distribution will be like this it is all constant. So, in this case it is aligning more towards our positive side, but it is evenly distributed across. So, this is also acceptable.

(Refer Slide Time: 31:14)



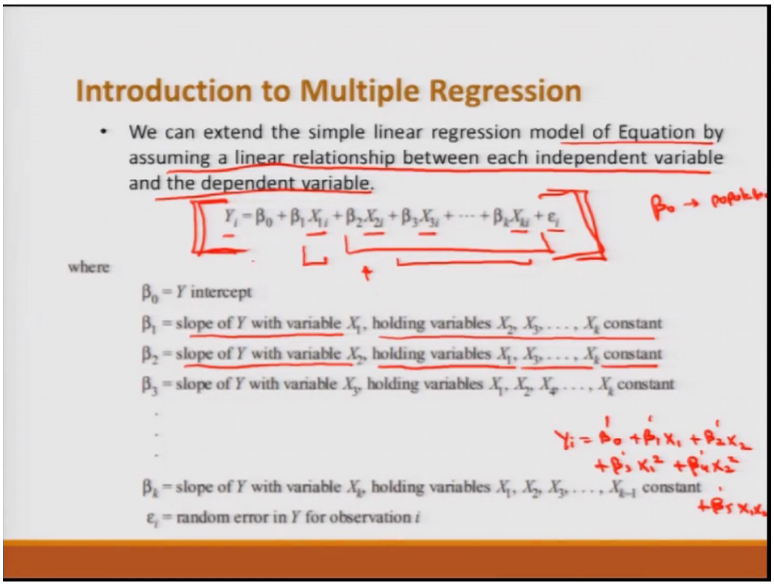
So, there are certain pitfalls or drawbacks in regression and certain ethical issues are there, that is, in regression there is lack of awareness of assumptions of least square regression because we just used the least square phenomena in regression. So, there are certain assumptions here which we have just discussed. Generally, the researchers do not know do and they are not aware of this assumptions and the regression only cannot just predict the final output. This just gives an overview discussed give broadly gives that what would be the behaviour of our product based upon the independent variables. So, the lack of awareness is there.

Number-2; not knowing how to evaluate the assumptions is one pitfall not knowing about the evaluation of assumptions, then not knowing what the alternatives to least square regression r if particular assumption is violated if some assumption is violated. What do we do? We apply the diagnostics. We will see these the forthcoming session. So, what diagnostics are to be apply for assumptions are not met this is not known generally.

So, using a regression model without knowledge of a subject matter is a big issue extrapolating outside the relevant range can be one issue. So, concluding that a significant relationship identified in an observation studies due to cause and effect relationship.

So, to learn regression we need to know what are the diagnostics, what are the various plots which will see in the r session of this regression here. What are the various statistics of parameters which check the validity of the regression model those are very important.

(Refer Slide Time: 34:00)

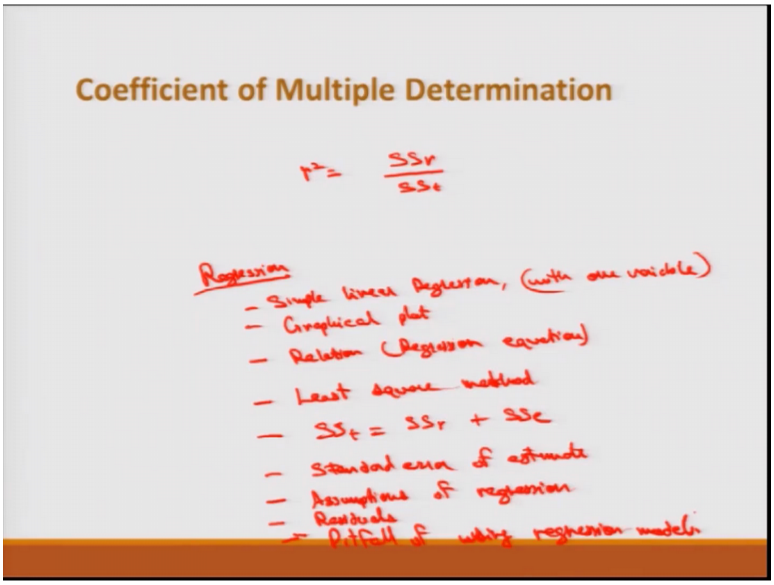


So, next comes the multiple regression we can extend this simple linear regression model of equation by assuming a linear relationship between each independent variable and then dependent variable. They say that Y i is depending upon X 1, X 2, X 3, so on up to X k, these much of independent variables and there is some error there is a linear relationship. So, this was one of the assumption here, where beta naught is the intercept for population. So, mind it this is beta, this is for population and beta 1, 2, 3 are the slope of Y with variable X 1 holding variable X 2, X 3 so on, up to X k constant, when these are all constant this slope is then calculated.

So, when X 1 and X 3 to X square constant X 2 is calculated again beta 2 is slope of XY with variable X 2 holding the variables X 1, X 3 and so on up to X k constant. So, this is multiple regression, but this is linear regression this is not non-linear regression. Even if the independent variables are of higher degree the regression is still linear regression if I say is the relation like this one Y i is equal to beta naught plus beta 1 X 1 plus beta 2 X 2 plus beta 3 X 1 square plus beta 4 X 2 square and if there is some interaction; interaction means both of the independent variables are interacting with each other that I can put beta 5 into X 1, X 2, is this a linear equation? No.

This is a polynomial equation on in second degree polynomial equation the equation is polynomial, but the regression is again linear here please note this thing generally people think that this is a non-linear regression non-linear regression is only when these coefficients beta naught beta 1, beta 2, beta 3, beta 4 are related to each other in a non-linear manner. So, that is the only non-linear regression we already talked about the linear regression and this equation here is the linear regression or linear multiple regression.

(Refer Slide Time: 36:58)



So, in this case for multiple regression also the coefficient of determination, coefficient of multiple determination is again r square that is the regression sum of squares by total sum of squares. So, this was regression. So, what we discussed here, we discussed simple linear regression with one variable. Then we saw the graphical plot of regression model then we saw the relation or regression equation in which we had the dependent variable and independent variable and intercept and error term.

Then, we saw the least square method which actually defined the relations here least square method of regression in which the predicted value of the sample was seen and how the predicted value is related to the observed and mean value. We saw that the total sum of squares is equal to the sum of the regression sum of squares and error sum of squares. Here, we saw that the error sum of square is intended to be minimised and regression sum of square value should be high. Then we saw standard error of estimate and we put a quick glance on the assumptions of regression, then we had a little information on the residuals that the variance should be constant and we saw the pitfalls of using regression models.

So, with this I would like to stop here. Let us meet in the next lecture.

Thank you.