

Statistical Learning for Reliability Analysis
Prof. Monalisa Sarma
Subir Chowdhury School of Quality and Reliability
Indian Institute of Technology-Kharagpur

Lecture-41
Regression Analysis (Part-II)

Hello guys. So, in continuation of our lecture on relation analysis, today we will be studying regression analysis. Regression analysis we have already done one lecture and regression analysis will be doing a couple of other more lectures as well.

(Refer Slide Time: 00:42)



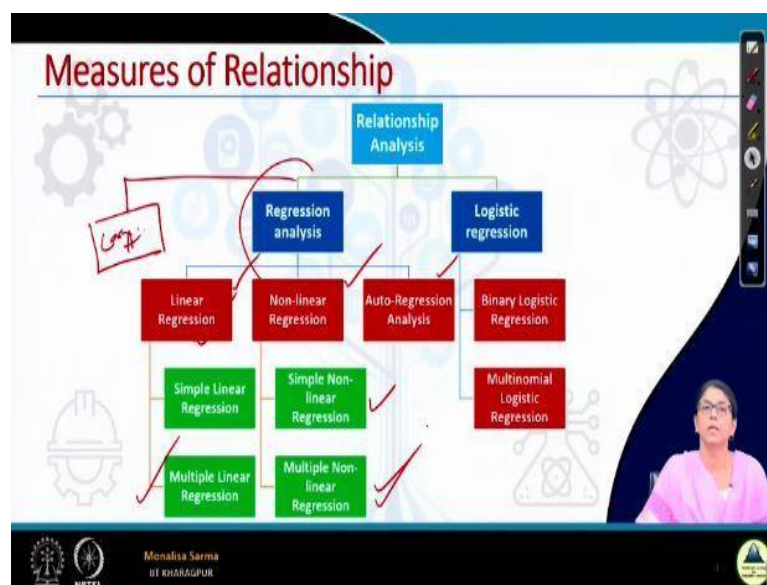
So, in today's lecture, we will basically learn simple linear regression. Simple linear regression as I already mentioned in the last lecture it is like, when there is one independent variable and one dependent variable. That means, the dependent variable we call it the response, independent variable we call it a regressor. So, when one dependent variable, when one response is dependent on the basically the regressor.

And they are basically linearly related and we can find a function that the relation basically, if we want to find out the how these two variables are related they say dependent variable and the dependent variable, how they are related if you can find the function to it. A linear functions basically, which graphs a straight line. So, that those will be discussed in details and that is in simple linear regression will be discussing those.

And now the function that we will find out, for simple linear regression, which will give an independent variable to find out the value of the dependent variable, the function that will do that. So, now this function how good this function is, how the means is this function is properly fitting the relations. Now, there are different ways to check that way. So, now one such measure is called quality of fit.

So, we will be discussing one such measure of finding out the how good the function is the function that we have found out for the simple linear regression. This two will be discussing in this lecture.

(Refer Slide Time: 02:14)



So, now coming to regression analysis well, as we have already mentioned in my last class there is a relation analysis, there are basically two type regression analysis and logistic regression. Well, I have remembered in one of my lecture I told you relation analysis. They are basically two categories, one is correlation analysis and another is regression analysis. So, here basically I can put one more thing, this is correlation analysis, regression analysis and logistic regression.

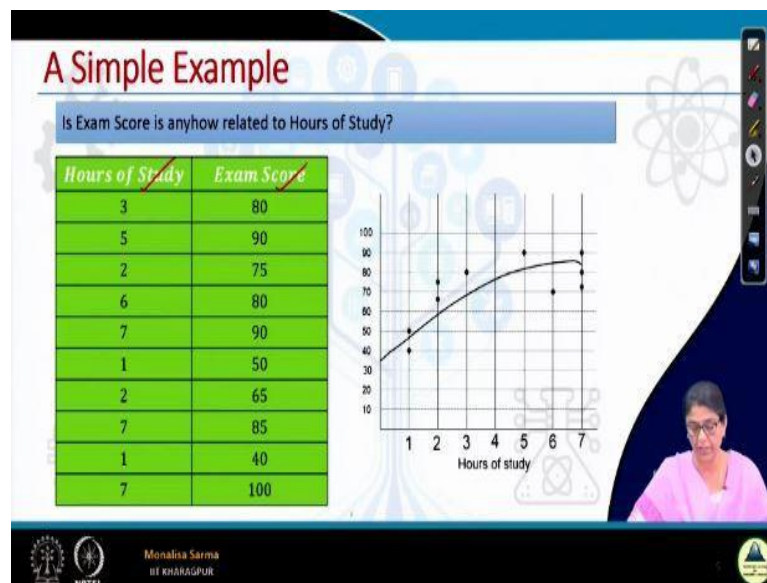
So, correlation analysis we have discussed I am just leaving that for the time being. So, now coming to this regression analysis, so in the regression analysis again there are 3 different types of regression analysis. That is linear regression, non-linear regression and auto regression analysis. So, we will be discussing this 3 regression analysis in 3 different lectures and again coming to linear regression and nonlinear regressions.

There are again simple linear regression and multiple linear regressions. Similarly for nonlinear also simple non-linear and multiple non-linear. So, simple means, when the independent variable is just one independent variable that is just one factor. And when I tell it is in many cases, there is more than one factor. like one good example is remember, when I have given an example price of a house in a particular locality others of the same square size, square feet area.

So, we know the price of a house in a particular area as per the square foot area. So, that is one factor. One area is the square foot area that is the one variable. And again the price of the house may vary based on the age of the house as well is not it? It can vary based on the square feet area, as well as it can vary based on the age also. So, there are 2 factors. When there are 2 factors involved and basically when there is more than one factor single.

Then we call if it is a linear regression, then we call it a multiple linear regression. If it is non-linear then we call it a multiple non-linear regressions.

(Refer Slide Time: 04:29)



So, in this lecture basically we will be discussing simple linear regression. That is one dependent variable, one independent variable. So, let us see the example, it is a very simple example. Here, there are basically 2 variables. One is the hours of study and another is the exam score. So, this is hours of study and the exam score. Now here can you tell me, which is the dependent variable, which is the independent variable?

Definitely hours of study cannot be a dependent variable, hours of study now does not depends on the exam score. Basically exam score may depend on the hours of study. So, here my

independent variable is the hours of study that is my regressor. Independent variable is also called a regressor. So, my hours of study are the regressors and exam score is the response or the dependent variable.

So, here this data is given. Basically how many students say 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 for 10 students, hours they put into studies and consequently the score they get in an exam. So, I want to find out is exam score anyhow related to the hours of study, is there any relations between the exam score and hours of study? So, if I just want to find out is there any relation between exam score and hours of study.

If I just want to find that is there any relation then what I will do? Then I will go to correlation analysis. So, basically correlation analysis will give me is there any relation plus as well if there is relation what is the strength of the relation as well. So, now here that is not my objective. Here my objective is that, of course it has to be related. Once it is related then only we can go to the next part like.

Now my next part is that if I want to find out if I put in say 4 hours of study what will be my likely exam score? So, if I want to answer such type of question definitely it is not a question of correlation analysis it is a question of regression. So, first what I did is I just try to plot the data. It is just simple, it is a scatter plot and I just try to draw a straight line just what to say.

So, that straight line, which more maximum of the points falls in the straight line, if not if the points does not fall in the same plane as at least it moves near by the points what is given. So, from this you can see, it has a linear sort of relationship.

(Refer Slide Time: 07:07)

Regression Analysis

Definition

The regression analysis is a statistical method to deal with the formulation of mathematical model depicting relationship amongst variables, which can be used for the purpose of prediction of the values of dependent variable, given the values of independent variable(s).

$y = f(x)$

Monalisa Sarma
IIT KHARAGPUR

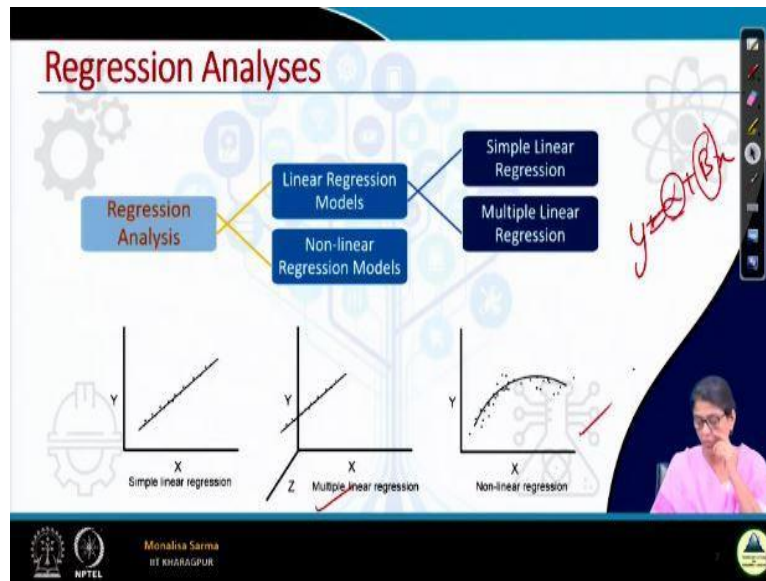
So, now from that I will now try to find out what is the relationship between this x and y . That is the hours of study and exam score. Now, first let us define regression analysis formally. So, what is this? A regression analysis is a statistical model to deal with the formulation of the mathematical model. It is a statistical model, what is this? It formulates a mathematical model.

What mathematical model, depicting relationship amongst variables what variables independent variable and dependent variable, which can be used for the purpose of prediction of the values of the dependent variable, given the values of independent variable. So, this is regression analysis is a mathematical model, which depict relationship between 2 variables. That is dependent and independent variables.

And it is used for the purpose of predicting the value of the dependent variable given the values of the independent variable. So, given the values of the independent variable in the last example what I have seen. Given the value suppose my hours of study, I put is 4 hours then what may be my likely exam score. So, I am trying to predict on the dependent variable. Dependent variable is the exam score.

So, I basically need a mathematical model, which gives the relationship between these two. The model is nothing, it is nothing but just that $y = f x$. What do I mean by model? Model is nothing, but just $y = f x$. So, my x is the independent variable, y is the response variable the dependent variable. So, I just need this function. What is this function? How y is related to x .

(Refer Slide Time: 08:42)



So, here is just a picture, how the different regression models look like. If it is simple linear regression then it will just graph a straight line. Maybe in the form of $y = mx + c$ or $y = \alpha + \beta x$. So, if I write y is equals to, this is very rating $y = \alpha + \beta x$ if I write. So, α is the intercept, β is the slope. Slope of the straight line, α is the intercept. Basically at what rate it moves corresponding to x or y moves corresponding to x that gives the slope β .

So, that is a linear regression model. Now if there are more than one factor as I have given the example, that if the property is not only dependent on the square foot it also depends on the age. So, then if it is more than one factor again it may be linear or non-linear. If it is a linear and more than one factor then what happens then again we may graph a straight line, but it will be in a multi-dimensional plane.

So, here you see there is a multiple linear regression. So, if it is a non-linear regression, non-linear regression it will graph a curve sort of pattern. So, this is an example of the figure for non-linear regression.

(Refer Slide Time: 10:02)

Simple Linear Regression Model

In simple linear regression, we have only two variables:

Dependent variable:
Also called *Response*, usually denoted as Y

Independent variable:
Also called *Regressor*, usually denoted as x

Linear regression

A reasonable form of a relationship between the Response Y and the Regressor x is the linear relationship, that is in the form $Y = \alpha + \beta x$

Monalisa Sarma
IIT KHARAGPUR

So, coming to some simple linear regression, as I already mentioned. We have two variables one is called the response or the dependent variable another is called the regressor or the independent variable. It is easy to remember. Response is the output. So, what is the output? Output is basically we are doing something to get the output. So, output will be definitely the dependent variable.

So, that is the response. One is response, other is the regressor. So, a reasonable form of relationship between response y and regressor x is a linear relationship that is a form of $y = \alpha + \beta x$ where, α is the intercept; β , is the slope.

(Refer Slide Time: 10:52)

Simple Linear Regression Model

In simple linear regression, we have only two variables:

A reasonable form of a relationship between the Response Y and the Regressor x is the linear relationship, that is in the form $Y = \alpha + \beta x$

Note

- There are infinite number of lines (and hence α_s and β_s)
- The concept of regression analysis deal with finding the best relationship between Y and x (and hence best fitted values of α and β) quantifying the strength of that relationship.

Monalisa Sarma
IIT KHARAGPUR

So, now this $y = \alpha + \beta x$ that is a mathematical model that we have considered. But then based on the points, if you just plot the points whatever data we have, based on the data we have

plotted the point. And we have tried to find out the line that crosses through this point. But you may see that there may be infinite lines that may cross to this point. Because it is very rarely that that is the case that all the points will fall on a straight line that we have made, that we have modeled.

So, when that is not the case there may be infinite lines that will cross through these points. So, when we have infinite lines means, there will be different values of α , there will be different values of β . So, now the concept of regression analysis what it does is. It deals with finding the best relationship between y and x , there infinite lines, we have to find out the best line. Best line that connects y and x that gives the relationship between y and x .

And hence the best fitted value of α and β . Best lines means, the best value of α and β and that also quantifies the strength of that relationship. So, that is the main objective of the regression analysis which we have to find out the best relationship between this y and x . That means we have to find out the value of the α and β that is the parameters; α and β are the parameters.

(Refer Slide Time: 12:17)

Regression Analysis

Given: The set $\{(x_i, y_i), i = 1, 2, 3, \dots, n\}$ of data involving n pairs of (x, y) values

Objective: To find "true" or population regression line, such that $Y = \alpha + \beta x + \epsilon$

Here, ϵ is a random variable with $E(\epsilon) = 0$ and $var(\epsilon) = \sigma^2$. The quantity σ^2 is often called the **error variance**.

Population regression line $Y = \alpha + \beta x$

$y = \alpha + \beta x$

Monalisa Sarma of KHARAGPUR

So, now actually, here there is one thing like if I just write $y = \alpha + \beta x$. That means, if I know α , if I know β that means given x I can find y . So, that means it is a deterministic equation. So, for given any value of x I will be able to find out the value of y , if α and β is if known. Now see the example like in the first class, when I was talking about regression analysis I have given few example like.

If the amount of money spent on the promotion of a movie depends that will help me in predicting, box of his performance. If I spend more on the more money and the promotion that my box of his performance will be more, means I will be able to predict based on the amount of money, which I spent in the promotion. But sometimes we see if many movies, where we have spent a lot of money under promotion.

But still it tanks in the box office. Again like same is the case to the advertisement, sometimes usually based on the advertisement expense we can predict the revenue of an item. But sometimes though the advertisement is we have spent quite a lot on advertisement, but still the revenue is not less. Still the product would not be sold again. Another one example, what I have given is the engine volume.

Different automobile engine volume is same, but then the mileage is different. Again this price of the house in the same locality, same square feet but, still some house are sold at a higher price, some houses are sold at a lower price. So, you see that this is not a deterministic equation. This is a deterministic equation.

So, basically from this example, what I have just told and there are many such scientific example, engineering applications, where we will see that this relationship it is not a deterministic relationship. There is another some random component to it. That random component I am calling it as epsilon. That is the error basically this random component the engineers or the statistician who calculate this.

Either they do not know what this random component is, why the price of home in the same locality with the same square foot, why it is different? This random component, this random portion why it is so either they do not know or they do not use it for calculation, there might be some other factors which are beyond their control. So, they do not use it for the calculation. So, there is a random component.

Similarly for all the other examples, so this random component makes the regression line random, it is not a deterministic line. So, our basically the regression line is this $y = \alpha + \beta x + \text{epsilon}$. So, now here x is very much a deterministic value, x we just take it the value, which you calculate, which is already given to us. Like the number of hours, which I will put in study that is 4 hours.

There is no randomness to it that is a deterministic value. So, just there may be measurement, there may be some negligible error in measurement, other than that there is no in-deterministic pathway, no randomness to it. But for the independent and dependent variable there is a random component attached to it because of which the equations that maps the dependent variable to the independent variable is no longer in deterministic equations, because of this epsilon.

Now here this epsilon is a random variable, where expected value is equals to 0. This is the assumption we take for linear regression to be valid; our epsilon should be our error component. This error should be such that this mean of this error should be 0. If the mean of this error is not zero basically the model which you come out that model is not a proper model actually.

So, if we assume that a mean of this expected value of epsilon is 0 and it has a constant variance, homogeneous variance, that is equals to σ^2 , a constant variance among all the epsilons, among all the errors. Error means, the difference of the point from the straight line from the line which we have got the regression line. The quantity σ^2 is often called the error variance.

(Refer Slide Time: 16:55)

Regression Analysis

Population regression line (hypothesis)

Note

- $E(\epsilon) = 0$ implies that at a specific x , the y values are distributed around the "true" regression line $Y = \alpha + \beta x$ (i.e., the positive and negative errors around the true line is reasonable).
- The values of the regression coefficients α and β to be estimated from data

Monalisa Sarma
BY KHARAGPUR

Now what does this expected value epsilon = 0 what does it implies actually? This implies that, that x and y are distributed around a true regression line. What does that expected below epsilon = 0 we have assumed that your epsilon is a random variable with expected value of epsilon = 0. So, what does this imply? This implies that for a specific x, the y values are distributed around the true regression line.

True regression line this is $Y = \alpha + \beta x$ and that is the positive and negative errors around the true line is reasonable. What does this expected value of $\epsilon = 0$? It implies that for a specific x of my y values, which are distributed around a regression line. It will be above the regression line or down the regression line it will be distributed around the true regression line in such a way. That means the positive errors.

Positive errors means, if it is above, negative error means it is down. If the positive errors and the negative error around the true line is reasonable, it is not very high. And when it is reasonable then we can assume that e of $\epsilon = 0$. And that is a necessary condition actually if we can do not assume that expected below $\epsilon = 0$ then that means our fit is not proper. That means our regression line the model, which have come up is not proper line.

So, because there are lots of error to it. So, now the next what we need to do is that. We found that so we will basically as we know there may be infinite lines. So, we have to find out the best line that fits the data points, that crosses the data points in such a way that expected value of the random component, that is expected value of $\epsilon = 0$. And once we find this line, so basically now the next step will be we have to find out the coefficient.

Finding the line means underline unless we know the coefficient that is the coefficient that is the α and β . The line is not complete our mathematical model is not complete. So, in our next step the values of the regression α and β are to be estimated. So, value of α and β are not given to us, we will have to estimate this value of α and β from the data.

(Refer Slide Time: 19:20)

The slide is titled "True versus Fitted Regression Line". It features a scatter plot of data points with two regression lines. The upper line is labeled $\hat{Y} = a + bx$ and the lower line is labeled $Y = \alpha + \beta x$. Handwritten red text next to the lines indicates $\alpha = a$ and $\beta = b$. A blue text box on the left contains the following text:

- The task in regression analysis is to estimate the regression coefficients α and β .
- Suppose, we denote the estimates a for α and b for β . Then the fitted regression line is $\hat{Y} = a + bx$ where, \hat{Y} is the predicted or fitted value

The slide also includes a video feed of a presenter in the bottom right corner and logos for NPTEL and Monalisa Sarma at Kharagpur in the bottom left corner.

So, suppose so what we will do, we have drawn an arbitrary line. Now the line is $Y = \alpha + \beta x$. Now that means we have to find out until and unless this line carries no meaning until and unless we know the value of α and β . So, basically we have to estimate the value of α and β . So, suppose we denote the estimate for $\alpha = a$ and for $\beta = b$.

Suppose we got an estimate, suppose we take $\alpha = a$ and $\beta = b$. Suppose this is the estimate we got. So, if this is the case then what is our line now? Our line is now $a + bx$. Now for that instead of Y I am writing is \hat{Y} , because this is my assumption. I have assumed that value of α is a , value of $\beta = b$. With this assumption the line I got $a + b x$. Suppose I got is \hat{Y} I am using a different variable name that is \hat{y} , where \hat{y} is the predicted or fitted value.

Suppose this is my actual line, this is my actual line $Y = \alpha + \beta x$. I do not know what is α , what is β . I do not know there is in the space there is a line. Basically there is a line, that line is $y + \beta x$, $Y = \alpha + \beta$ that is my actual line. I do not know. I have estimated the value, say I have estimated a value of α is a and estimated the value of β is b and with that I got a line that is $a + bx$. So, suppose that is my $a + bx$ gives a value \hat{Y} , this is my predicted fitted value.

(Refer Slide Time: 21:19)

Least Square Method to estimate α and β

Concept of Residuals

This method uses the concept of residual. A residual is essentially an error in the fit of the model $\hat{Y} = a + bx$. Thus, i^{th} residual is

$$e_i = y_i - \hat{y}_i, i = 1, 2, 3, \dots, n$$

The slide features a graph with a vertical axis labeled 'Y' and a horizontal axis labeled 'X'. Two lines are plotted: a solid line representing the actual model $Y = \alpha + \beta x$ and a dashed line representing the fitted model $Y = a + bx$. A point on the solid line is connected to the dashed line by a vertical dashed line, and the distance between them is labeled as the residual e_i . The slide also includes a small video inset of a woman in the bottom right corner and logos for IIT Kharagpur and NPTEL at the bottom.

So, now we need to find out, we have estimated that value a and b . Estimated that $\alpha = a$, β is b , we have to find out what is that a , what is that b . For that first we will have to know the term there is a term called residual. So, what is this residual? So, basically the thing is that to find out the values of α and β one of the most used methods is they call the least square method.

So, we will be learning this least square method, least square methods to estimate α and β . So, in this least square method uses a concept called residual. What is residual now? The residual is essentially an error in the fit of the model. My predicted line is this, is not it. What is my actual line? Actual line is equal to $Y = \alpha + \beta x$. So, from these two lines, I get a residual.

What is that and that is the i th residual is e of i is Y of i - \hat{Y} of i cap. That means the difference between these two. This is my actual line, this is my predicted line. So, residual is difference between these two, this portion, this is my residual. So, e of i . So, when I can say that the value whatever predicted is a good fit. So, when my e of i is as small as possible is not it. So, my residual both negative residual and positive residual when it is as small as possible then I can say that is a good fit.

So, my residual should be as small as possible. Now this is I have found out e i just for one value of x i , similarly I will take different values of x i , x 1, x 2, x 3, here like for the example I have taken different hours of study. Hours of study 1, 2, 3, 5, 7 times and many I took different X i . For different X i 's I got a Y i value is not it. So, similarly I will find out the residual for all such X i 's.

(Refer Slide Time: 23:20)

The slide is titled "Least Square Method to estimate α and β ". It features a green header "Sum of Squares Error (SSE)". The main content is on a blue background and includes the following text and formulas:

The residual sum of squares is often called the sum of squares of the errors about the fitted line and is denoted as

$$SSE = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^n (y_i - a - bx_i)^2$$

A graph on the right shows a coordinate system with X and Y axes. A line is plotted, and a point (x_i, y_i) is shown. The vertical distance between the point and the line is labeled e_i . The line is labeled $\hat{Y} = a + bx$. A text box below the graph states: "We need to minimize the value of SSE and hence to determine the parameters of a and b ."

At the bottom of the slide, there are logos for NPTEL and the presenter's name, Monalisa Sarma, of Kharagpur.

So, this is to find out so what is the best? I told you the best is when my this residual is as small as possible. So, what will I do with this? Now this residual can be positive also as well as it can be negative also. So, I will take the square. This method is called the sum of squares error. So, what will I do is that this is the residual sum of square is often called the sum of squares of the errors, about the fitted line and is denoted as SSE.

So, sum of squares of the errors, it is called sum of squares error SSE. So, how it is? SSE is nothing but summation of all e_i squares. What is e_i ? e_i is $Y_i - \hat{Y}_i$ actual value minus predicted value. So, similarly for each X_i I will get actual value and predicted value. So, SSE is summation of all such Y_i squares. So, this is e_i^2 is nothing but this is the thing and what is \hat{Y}_i ?

\hat{Y}_i is $a - bx_i$ is not it, that is what we have predicted the value of α and β we got this. So, this is my SSE is this? So, I need my SSE to be minimum. So, under what situation my SSE will be minimum under for what value of x my X and Y and SSE will be minimum. So, to find it out we need to minimize the value of SSE and hence to determine the parameters of a and b .

So, under what situation my this SSE will be minimum, for what value of a and for what value for this difference because a and b these are the parameters which have assumed. So, for assuming this a and b I have got a predicted value. So, now this difference of predicted and original actual value is my e_i . So, sum of all this e_i that I am calling it SSE. Now I need to minimize this SSE.

So, and for what value of my parameter estimation, because of this different parameter estimation only I am getting a different line is not it. So, for what value of this parameters a and b I will get a minimum SSE? So, how to do when I want to find a value of the variable, so that the function is a minimum.

(Refer Slide Time: 25:39)

Least Square Method to estimate α and β

Minimizing the Sum of Squares Error (SSE)

Step 1: Differentiation
Differentiating SSE with respect to a and b , we have

$$\frac{\partial(SSE)}{\partial a} = -2 \sum_{i=1}^n (y_i - a - bx_i) = 0$$

$$\frac{\partial(SSE)}{\partial b} = -2 \sum_{i=1}^n (y_i - a - bx_i) \cdot x_i = 0$$

Step 2: Equating the partial derivatives to zero
For minimum value of SSE, $\frac{\partial(SSE)}{\partial a} = 0$, and $\frac{\partial(SSE)}{\partial b} = 0$

So, it is just partial differentiation and then we equalize it to 0. So, differentiating SSE with respect to a and b. This is a simple function concept we all of you know. So, I will do partial differentiation SSE with respect to a then I will do differentiation of SSE with respect to b. When I differentiate SSE with respect to a I got this and with respect to b I got this. Now this I will equate it to 0.

This is equated to 0, I got 2 equations. From this 2 equation, I can find out the value of a and b. I want to find out the value of my parameters a and b is not it? So, that is all equating the partial derivatives to zero.

(Refer Slide Time: 26:21)

Least Square Method to estimate α and β

Minimizing the Sum of Squares Error (SSE)

Step 2: Equating the partial derivatives to zero
 For minimum value of SSE, $\frac{\partial(SSE)}{\partial a} = 0$,
 and $\frac{\partial(SSE)}{\partial b} = 0$
 Thus we get,

$$na + b \sum_{i=1}^n x_i = \sum_{i=1}^n y_i$$

$$a \sum_{i=1}^n x_i + b \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i$$

Step 3: Solving for a and b
 These two equations on the left can be solved to determine the values of a and b, and it can be calculated that

$$b = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$a = \bar{y} - b\bar{x}$$

The slide also features a small table with columns X and Y, and a video inset of a presenter in the bottom right corner.

So, I got these two expressions. This simple nothing to explain here, all of you know how to find out the minimum of a function and what value the function will have a minimum value that is all nothing else. So, these are the two expressions. So, once we solve this equation. This is my value of b and this is my value of a. By solving this two equation I got my value of a and b.

So, this is the least square method to estimate α and β . This method is called least square method to estimate the α and β . So, what α value I have estimated a and β value is estimated as b. What is my a value? This is my a value, this is my b value. So, a value and b value all I am getting it in terms of x i's. See x i's and y i's basically, so this is my b value, I got all my terms in x i, y i, x bar, y bar all from the data itself. I have a set of data. Similarly I got a value also.

So, now from this we found out the value of the parameters a and b. So, I got the line $y = a + bx$ that is my line which basically gives the relationship between x and y given a value of x I can be able to predict the value of y using this line $Y = a + b x$ I got the line. Now the question is that, is this line the properly fits the data that is actually given or this line maybe it is not a proper fit. So, for that we need to find out one measure that is called measure of quality of fit.

(Refer Slide Time: 28:08)

R^2 : Measure of Quality of Fit

Coefficient of Determination

A quantity R^2 , is called **coefficient of determination** is used to measure the proportion of variability of the fitted model.

Total corrected sum of squares:

We have $SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$. It signifies the **variability due to error**.

Now, the **total corrected sum of squares** is defined as $SST = \sum_{i=1}^n (y_i - \bar{y})^2$

R^2 :

SST represents the variation in the response values. The R^2 is: $R^2 = 1 - \frac{SSE}{SST} = D$

Monalisa Sarma
of KHARAGPUR

So, the one measure of quality of fit is R^2 measure. It is a very commonly used R^2 measure to find out the quality of the fit of the regression line, quality of the fit of the mathematical model that we got. So, this is nothing but a mathematical model is not it $Y = a + bx$. So, how good the line is? So, this quantity R^2 is called coefficient of determination.

This is used to measure the proportion of variability of the fitted model. The fitted model that I have fitted that means the straight line, the straight line that I have fitted as how much variability is there from the actual. Means actual value, which I have from x_i and y_i what value it gives and what value this line gives, how much variability is there? So, basically that is the quantity R^2 .

The quantity R^2 is the coefficient of determination is used to measure the proportion of variability of the fitted model. So, R^2 . To define R^2 basically we need two terms, one is SSE that is that we have just already seen some of this. This one this SSE what is this SSE? Sum of squares error, how did you find out the sum of squares that are fitting the what the predicted value and the actual value.

So, for R^2 calculation we need two values, one is SSE and one other is SST. So, this is SSE, SSE is nothing it signifies the variability due to error. SSE what does it signify? This is the formula for SSE. We have just now seen is not it. What is the variability due to the error? Means it is basically how much is the error between the predicted line and the actual line. So, that is my sum of squares due to error.

And another one term that we need is total corrected sum of squares. Total corrected sum of squares you are not new to this term we have used this term by learning anova remember. So, total corrected sum of squares it is basically again it is also variability of the data from the mean of the data. So, that all the data that we have used for regression analysis how much this data is variable from the mean of that data.

See SST is nothing but $y_i - \bar{y}$. We have taken some data x_i and y_i is not it. This y_i that we have already some data as we have based on this data only we are finding the model. Without the data we cannot find a model. So, now this data how much this data there is variability in this response data. So, that is SST. Now so what does R^2 . See SST represent the variation in the response values.

How much this response why I add a response values, how much variation is there in the response value? That has nothing to do with the error or something like that. We have taken different x value, for different x value what is the y value. So, that is this different y values that we get this what is the variability of this data. That is the SST total corrected sum of squares. So, my R^2 is nothing but $1 - SSE / SST$.

Now if my $SSE = 0$ suppose, under what situation my SSE will be 0? If my SSE is 0 that means my line is a very good fit is not it, my line is a very good fit then only my sum of squares due to errors is equals to 0. In that case what I will get $R^2 = 1$. That means it is a very good fit, quality of fit is very good. Even if suppose my SSE is quite less than SST. If it is quite less than SST it is not zero but quite less than SST then also I will get a very less term here.

If I divide this SSE / SST and my R^2 will be closer to 1 that also indicates a good fit. Of course 1 is a very good fit and values which are closer to 1 also good fit. Now if SSE is almost equal to SST. That means the sum of variability due to error as well as the variation in the response;

variable due to its error is so high that we are getting almost equal to the variation in the response value.

Variation in a response value is different thing is not it; my variable to do error is so high that I am getting it is almost equals to SST. Then what will happen? My R^2 will be equal to 0. That means my fit is not at all a proper fit. It is not a good fit $R^2 = 0$ means it is not a good fit at all. And theoretically in fact my R^2 value can be less than 1 also. Under what situation my R^2 value will be less than 1.

When my SSE value is bigger than SST see pay attention please when my SSE value is bigger than SST sum of squares due to error is bigger than SST you can imagine that means that the line I have got that is not at all it is showing totally wrong trend. That is not at all the line that actually these two independent variables and dependent variables are fitted. The line that the regression line I have got that is totally in the opposite trend.

It is not showing the proper trend that is why I am getting such a high error that it is bigger than SST. So, theoretically R^2 less than 0 is also possible, negative also is also possible. In that case it indicates my quality for fit is equation of quality of it does not come basically, it is totally then my regression line that I got it is totally showing a wrong trend.

(Refer Slide Time: 34:11)

R^2 : Measure of Quality of Fit

Coefficient of Determination

Note

- If fit is perfect, all residuals are zero and thus $R^2 = 1.0$ (very good fit)
- If SSE is only slightly smaller than SST, then $R^2 \approx 0$ (very poor fit)

$R^2 = 1.0$ (Very good fit)

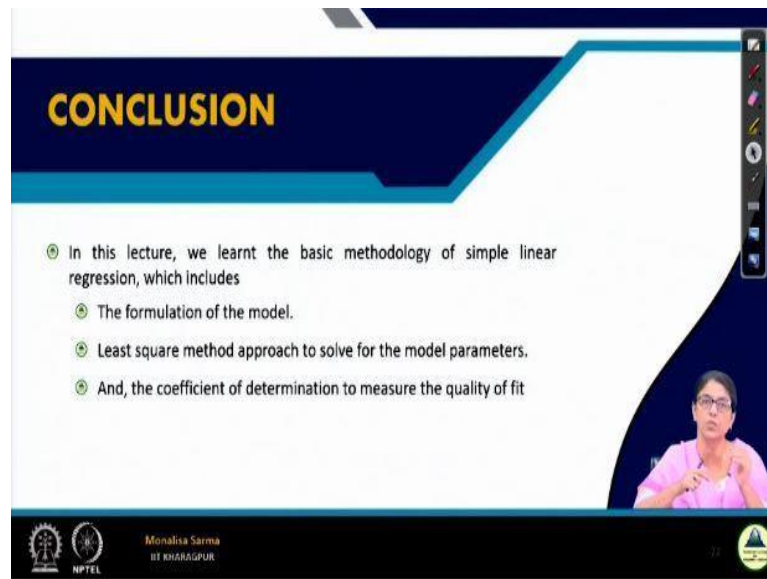
(Very poor fit)

Mona Lisa Sarma
IIT BHARAGUR

So, see if the fit is perfect, all residuals are 0 and $R^2 = 1$ as I told you SSE = 0 then all fit is perfect, all residuals are fit, that is of course it is very impractical but getting a value, which is closer to 1 is also very good fit. So, if SSE is only slightly smaller than SST, then $R^2 = 0$ almost

equals to 0 that is a very poor fit. So, you see it is a very good fit. Almost all the points are in the line or very near the line and this is example where it is a very poor fit.

(Refer Slide Time: 34:50)



CONCLUSION

- In this lecture, we learnt the basic methodology of simple linear regression, which includes
 - The formulation of the model.
 - Least square method approach to solve for the model parameters.
 - And, the coefficient of determination to measure the quality of fit

Monalisa Sarma
IIT KHARAGPUR

NPTEL

So, in this lecture, we learned the basic methodology of simple linear regressions. What that includes? That includes the formulation of the model, how we basically formulate the model to get the function, between the dependent variable and the independent variable. Then we have seen the least square method that approach to solve the model parameters, that is α and β . That is the intercept and the slope.

And also we have seen the coefficient of determination to measure the quality of fit, coefficient of determination that is the R^2 to measure the quality of fit. So, we have discussed all these three things in this lecture.

(Refer Slide Time: 35:30)

REFERENCES

📖 The Elements of Statistical Learning, Data Mining, Inference, and Prediction (2nd Edn.), Trevor Hastie, Robert Tibshirani, Jerome Friedman, Springer, 2014.



The book cover is orange and black. It features the title 'The Elements of Statistical Learning' in white and black text, with the authors' names 'Trevor Hastie, Robert Tibshirani, Jerome Friedman' above it. The Springer logo is at the bottom.



A small video feed in the bottom right corner shows a woman with glasses wearing a pink shirt, speaking.

  Monalisa Sarma
IIT KHARAGPUR 

And so this is the reference. And thank you guys.