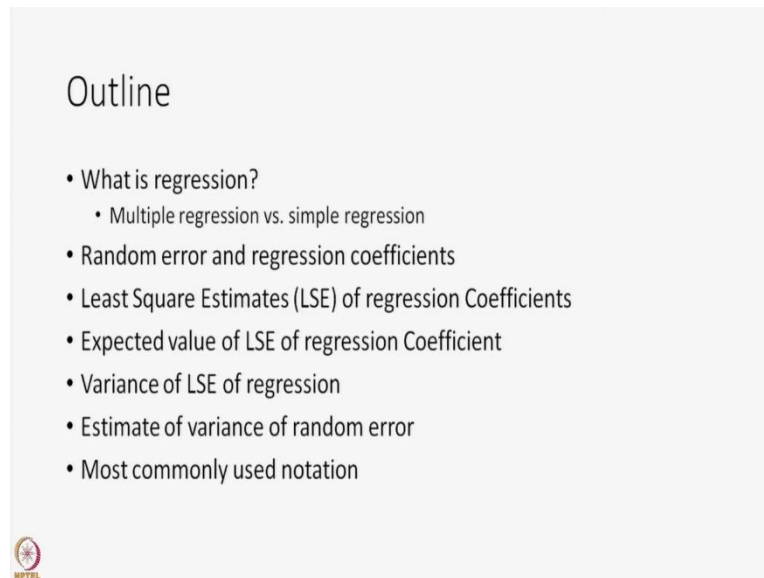


Dealing with Materials Data: Collection, Analysis and Interpretation
Professor. Hina A. Gokhale
Department of Metallurgical Engineering and Materials Science
Indian Institute of Technology Bombay
Lecture 81
Regression Analysis - 1


Hello and welcome to the course on Dealing with Materials Data. Today, we are going to start a very important and a fresh subject called Regression Analysis. Regression analysis plays an important role in data analysis, where you have two sets of data, one is a response value, while the others are the independent values and the most common analysis that is being used or it should be used to begin with, is a regression analysis and now we will go through this session on various aspects of regression analysis

(Refer Slide Time: 01:11)



Outline

- What is regression?
 - Multiple regression vs. simple regression
- Random error and regression coefficients
- Least Square Estimates (LSE) of regression Coefficients
- Expected value of LSE of regression Coefficient
- Variance of LSE of regression
- Estimate of variance of random error
- Most commonly used notation




So, the outline is for this particular session is going to be, first we will define, what is regression? There are two kinds multiple regression versus simple regression. We will talk about the random error and the regression coefficients. The least squares estimates of regression coefficients, the expected value of for least squares estimates of regression coefficient, it is variance. The estimate of variance for random error and we will, at the end we will give a slide on most commonly used notation. This slide will be useful in future for reference.

(Refer Slide Time: 01:49)

Introduction

- Consider Y a response variable and x_1, x_2, \dots, x_r as independent variables.
- Simplest relationship that can exist between Y and x_1, x_2, \dots, x_r is linear
 - $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_r x_r + \epsilon$
 - where $\beta_0, \beta_1, \dots, \beta_r$ are some constants and ϵ represents random error in this relationship
- If $r = 1$, then $Y = \beta_0 + \beta_1 x_1 + \epsilon$ is called simple linear regression
- In general with r independent variables it is called Multiple Regression
- $\beta_0, \beta_1, \beta_2, \dots, \beta_r$ are called regression coefficients...and usually need to be estimated from the data



So let us start, as I said suppose we have a response variable Y and some independent variable X_1, X_2, X_r and you know that there is some relationship between Y and X_1, X_2, X_r or you at least suspect that there might be a relationship between Y and X_1, X_2, X_3, X_r . The simplest relationship that can exist between them is a linear relationship which can be expressed as

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_r x_r + \epsilon$$


This Epsilon is a very important point. In the reality when we get the data, we cannot be always sure that the relationship will be exact like this. If you do not consider epsilon this relationship is a mathematical relationship. When you add an epsilon quantity, which is called a random error. Epsilon expresses represents a random error in this relationship. This is where the relationship becomes random or statistical in nature.

Now, If $r = 1$, then $Y = \beta_0 + \beta_1 x_1 + \epsilon$ is called simple linear regression

In general, with r independent variables, it is called multiple regression and the notation wise $\beta_0, \beta_1, \beta_2, \dots, \beta_r$ are called regression coefficients and they are generally unknown they need to be estimated from the data.

(Refer Slide Time: 04:25)

- The random error ϵ is assumed to have expected value 0, then
- $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_r x_r + \epsilon$ can also be written as
- $E(Y|x_1, x_2, \dots, x_r) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_r x_r$
- So the regression coefficients $\beta_0, \beta_1, \beta_2 \dots \beta_r$ need to be estimated given the values of x_1, x_2, \dots, x_r .
- First take the case of simple regression model:
 - $Y = \beta_0 + \beta_1 x_1 + \epsilon, E(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2$



So, then you have a random error. Generally, it is assumed that the random error has an expected value of 0. There is a random error epsilon, which has an expected value 0. What it means is that on the average Y actually equal to this value on the average. But otherwise there is a plus or minus error in it. This plus or minus error is represented by epsilon and that we can say that its plus or minus average is by saying that its expected value is 0.

So, in other words, if you recall our previous sessions, we can say that the expected value of Y given

- $E(Y|x_1, x_2, \dots, x_r) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_r x_r$

Because the expected value of epsilon is 0 there is no epsilon here. So, the regression coefficients beta 0, beta 1, beta 2 and x_r need to be estimated given the values of X_1, X_2, X_3, X_r .


It means that the independent variable will be given fixed random fixed values for us not a random variable in this particular case and Y is going to be random variable. Because of the randomness of epsilon. First we will discuss in detail, the estimation procedure for beta 0, beta 1 and sigma square, which is a variance of epsilon, variance of error through the case of simple regression that is

- $Y = \beta_0 + \beta_1 x_1 + \epsilon, E(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2$

(Refer Slide Time: 06:41)

Least Square Estimator

- Want to estimate β_0, β_1 in the regression relation $Y = \beta_0 + \beta_1 x_1 + \epsilon$
- Want to find β_0, β_1 by minimizing the squared error between values of Y and its estimator $\beta_0 + \beta_1 x_1$
- Let us denote estimated value of β_0, β_1 as A and B respectively
- Want to minimize
- $SS = \sum_{i=1}^n (Y_i - A - Bx_i)^2$
Sum of squares



What estimation we could have done? The most commonly use estimator is called least squares estimator. What are we really trying to do? Let us take the case. This is X and this is Y axis. Suppose we have a few data points, which go like this and you have to fit a line through it, you have already done this exercise in algebra and this line we fit in such a way that the distance between the actual, the line and the actual value is minimize. This is called least squared estimator.

So, we would like to find estimate β_0, β_1 by minimizing the squared error between values of Y and its estimator $\beta_0 + \beta_1 x_1$. So, let us denote the estimator to differentiate between the actual values beta 0 and beta 1 and its estimator we called the estimators A and B.

So, what we are trying to do is, we want to minimize the sums of squares of

$$SS = \sum_{i=1}^n (Y_i - A - Bx_i)^2$$

I guess you already know, why do we take a square? Because if we do not take a square, the sum of the distances that you calculate that is without the square if you take Y_i minus A minus B xi the best is when it becomes 0. If you take a mean value.

So, the idea is that we square the distance, so that we remove the sign of the difference between Y and A plus B xi and then we take a square of it and now we try to find A and B which would minimize, which would minimize the sums of squares SS, SS is called SS because it is sum of squares, it is a sum of squares.

(Refer Slide Time: 09:36)

• $\frac{\partial SS}{\partial A} = -2 \sum_{i=1}^n (Y_i - A - Bx_i) = 0$

• $\frac{\partial SS}{\partial B} = -2 \sum_{i=1}^n x_i (Y_i - A - Bx_i) = 0$

• Further simplification we will get

• $A = \bar{Y} - B\bar{x}$

• $B = \frac{\sum_{i=1}^n x_i Y_i - n\bar{x}\bar{Y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$

NPTEL

So, the easiest way of doing it is by taking a partial differential, partial differentiation with respect to A.

- $\frac{\partial SS}{\partial A} = -2 \sum_{i=1}^n (Y_i - A - Bx_i) = 0$
- $\frac{\partial SS}{\partial B} = -2 \sum_{i=1}^n x_i (Y_i - A - Bx_i) = 0$

And this is a very simple simplification, you will come to know that when you do this little algebra A turns out to be

$$A = \bar{Y} - B\bar{x}$$

B turns out to be which looks a little bit complicated, but as we go on you will recognize this term,

$$B = \frac{\sum_{i=1}^n x_i Y_i - n\bar{x}\bar{Y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

If you look at it very carefully this comes very close to correlation coefficient. But how to derive it. I leave it to you, I think it is a good exercise to simplify this to get to this equation.

Now, we come to what is the distribution of A and B? Remember that now you are Y_i is a this whole thing is estimated using the value, let us start the pen. This whole thing is estimated using value Y_i and \bar{Y} . Remember that x_i and \bar{x} are given values. So, they are not random variables.

It is the Y_i which is a random variable and therefore A and B now are random variable and we must know, what is distribution like as we are done in the past. While working out the estimation theory and the hypothesis testing, we need to know the distribution of this random quantity, which we are going to use as estimator.

(Refer Slide Time: 12:11)

Distribution of A and B

$Y_i = \beta_0 + \beta_1 x_i + \epsilon \quad i=1, \dots, n$
 $E(\epsilon) = 0 \quad \text{Var}(\epsilon) = \sigma^2$
 No distribution assumed for ϵ so far.

- Assume that $\epsilon \sim N(0, \sigma^2)$, then for $i = 1, 2, \dots, n$
- $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2) \quad i=1, 2, \dots, n$
- $E(B) = \frac{\sum (x_i - \bar{x}) E(Y_i)}{\sum x_i^2 - n\bar{x}^2} = \beta_1$
- $E(A) = \sum_{i=1}^n \frac{Y_i}{n} - \bar{x} E(B) = \beta_0$
 $= \beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x} = \beta_0$

Handwritten derivation for $E(B)$:

$$\begin{aligned} E(B) &= \frac{\sum (x_i - \bar{x})(\beta_0 + \beta_1 x_i)}{\sum x_i^2 - n\bar{x}^2} \\ &= \frac{\sum (x_i \beta_0 - \bar{x} \beta_0 + \beta_1 x_i^2 - \beta_1 x_i \bar{x})}{\sum x_i^2 - n\bar{x}^2} \\ &= \frac{\sum (x_i - \bar{x}) \beta_1}{\sum x_i^2 - n\bar{x}^2} = \beta_1 \end{aligned}$$

- $\frac{\partial SS}{\partial A} = -2 \sum_{i=1}^n (Y_i - A - Bx_i) = 0$
- $\frac{\partial SS}{\partial B} = -2 \sum_{i=1}^n x_i (Y_i - A - Bx_i) = 0$
- Further simplification we will get
- $A = \bar{Y} - B\bar{x}$
- $B = \frac{\sum_{i=1}^n x_i Y_i - n\bar{x}\bar{Y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$

So, how to find a distribution of A and B? First we make an assumption, on the distribution of Y. Remember that Y is defined as

$$Y = \beta_0 + \beta_1 x_1 + \epsilon$$

So far our assumption were only

$$E(\epsilon) = 0 \text{ and } \text{Var}(\epsilon) = \sigma^2$$

Remember just to remind you that this sigma square does not depend on the data value i .

So, this is the relationship with respect to i is equal to 1 to n . But this sigma square is not dependent on i . So, in this relationship, now only we are adding an assumption. There is no distributional assumption made so far, no distribution assumed for epsilon only so far. So, far we have not made any assumption that is being made now.

Now we are saying that suppose epsilon is distributed as a normal distribution with mean value 0 and the variance common variance sigma square for all i is equal to 1, 2, 3 etc n . Why this is a normal? Because it is very common to and it is very well known fact. That by enlarge the errors are distributed as normal distribution.

It is a very old story, that it was the Galileo who made so many observations of stars and when he found that every time when he makes an observation there is a minute error and that error. After 200 years, it was Gauss who found that this error behaves in a very perfect bell shaped curve and it was called the Gaussian distribution therefore it has become a normal distribution. But that is the side story.

So, any error to be assumed as a normal distribution is a natural process. So, here we assume it as a normal distribution with mean 0 and variance sigma square. and therefore it implies that our Y_i for i is equal to 1 to up to n is also distributed as a normal distribution.

Assume that $\epsilon \sim N(0, \sigma^2)$, then for $i = 1, 2, \dots, n$

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

Now we can find an estimated value of B . Because B you please recall the previous slide the estimator of A involves B . Therefore, first we must try to find the expected value of B and use it in the estimation expected value of A and therefore we come here and we find that expected value of B can be found by, you remember that these are all the constant values given values to us.

$$E(B) = \frac{\sum (x_i - \bar{x}) E(Y_i)}{\sum x_i^2 - n\bar{x}^2} = \frac{\sum (x_i - \bar{x}) (\beta_0 + \beta_1 x_i)}{\sum x_i^2 - n\bar{x}^2} = \frac{\sum (x_i - \bar{x})^2}{\sum x_i^2 - n\bar{x}^2} = \beta_1$$

$$E(A) = \sum_{i=1}^n \frac{Y_i}{n} - \bar{x} E(B) = \beta_0$$

Therefore, it is only the Y_i which is a random variable. Therefore, this becomes this now if you replace Y_i by $\beta_0 + \beta_1 x_i$, it will reduce down to the same thing. Shall we do it here?

So, this quantity cancels and this quantity brings out the beta 1 and the xi square minus summation x bar square. So, it will bring you nx bar square and therefore this will become beta1. But this is the quantity in which you have to realize, so we have the, this quantity will cancel out and this quantity results into this value and therefore it is beta 1 and once you put this into it, this is very simple.

Because this basically gives you the expected value of Y bar, which is nothing but beta 0 plus beta 1 x1 bar and then you will again make minus beta 1 x bar and therefore this will become beta 0. So, this is how the distribution in the distribution of A and B, we find that expected value of A. So, expected value of B is beta 1 and expected value of A is beta 0.

Just go back and think a little bit. Because epsilon is assumed to be normal with 0 mean and variance sigma square Yi becomes normal with a expected value of beta 0 plus beta 1 xi and variance sigma square and you can see that, the estimate value of B is also a function of Y with certain constant and estimate or the, the expected or the estimate of beta0 A is also a function of Y only rest of it is a constant.

You will find that this two are also distributed that is A and B random variables are also distributed as normal. So, all we know to know is its expected value and its variance. So, in the next case, we will go, we are going to find out the variance of B and variance of A.

(Refer Slide Time: 20:56)

Variance of A and B

$$\bullet \text{Var}(B) = \frac{\text{Var}[\sum_{i=1}^n (x_i - \bar{x}) Y_i]}{[\sum_{i=1}^n x_i^2 - n\bar{x}^2]^2} = \frac{\sigma^2}{S_{xx}}, \text{ where}$$

$$\bullet S_{xx} = \sum_{i=1}^n x_i^2 - n\bar{x}^2 = \sum_{i=1}^n (x_i - \bar{x})^2$$

$$\bullet \text{Var}(A) = \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n(\sum_{i=1}^n x_i^2 - n\bar{x}^2)} = \frac{\sigma^2 \sum x_i^2}{n S_{xx}}$$

$\text{Var}(z) = \sigma^2$
 $\text{Var}(az) = a^2 \sigma^2$
 $\text{Var}(X) = \sigma_x^2, \text{Var}(Y) = \sigma_y^2$
 X and Y are indep.
 not corr.
 $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$
 $= \sum (x_i - \bar{x})^2 \underbrace{\text{Var}(Y_i)}_{\sigma^2}$

So, the variance of A and B again you have to follow the same formula. Variance of A and B is

$$\text{Var}(B) = \frac{\text{Var}[\sum_{i=1}^n (x_i - \bar{x})Y_i]}{[\sum_{i=1}^n x_i^2 - n\bar{x}^2]^2} = \frac{\sigma^2}{S_{xx}}, \text{ where}$$

$$S_{xx} = \sum_{i=1}^n x_i^2 - n\bar{x}^2$$

$$\text{Var}(A) = \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n(\sum_{i=1}^n x_i^2 - n\bar{x}^2)}$$

Because when you do the variance, the constant is square what formula we have used here, if you recall is that if a variance of a random variable x is sigma square. Then variance of random variable ax is a square sigma square.

So, this formula is used being here. Therefore, the denominator which is only a multiplier remember xi is a given value, so it is a constant value. We understand it is not a random variable. So, it is only seats, comes out as a squared of it 1 divide by that as a square of it and then you have to take the variance of Yi with its multiplier and therefore that is also going to be and another formulae if you know that variance of x is sigma1 square and variance Y is sigma 2 square. Then and X and Y are independent or to be very clear not co-related.

In that case, variants of X plus Y is variance of X plus variance of Y. So, using that formula we can simplify this. By stating that, this is equal to summation of Xi minus X bar whole square. This part comes out because of this and then you have a variance of Yi and you know that variance of Yi this part is equal to sigma square.

So, when you simplify it, it comes to sigma square divided by Sxx and this is a notation we would like to introduce here. Sigma xx is equal to summation of xi square minus or it is the same as summation of xi minus x bar whole square i rise from 1 to n. This is a new notation we are including here and therefore this becomes a variance. The variance of A can also be derived in a similar way and the variance of A can be found to be the same thing.

$$\text{Var}(A) = \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n(\sum_{i=1}^n x_i^2 - n\bar{x}^2)}$$

Please note that this can also be written as sigma square summation xi square divided by n Sxx.


(Refer Slide Time: 24:43)

Sum of Squares of Residual (SSR)

- Y_i is an i^{th} observed value and $A + Bx_i$ is estimated value i^{th} observation.
- Therefore residual $R_i = Y_i - A - Bx_i$, then
- $SSR = \sum_{i=1}^n R_i^2$
- $\frac{SSR}{\sigma^2} \sim \chi_{n-2}^2$, Hence, $E\left(\frac{SSR}{\sigma^2}\right) = n - 2$
- Or
- $E\left(\frac{SSR}{n-2}\right) = \sigma^2$

3 unknown parameters
 $\beta_0, \beta_1, \sigma^2$
 $\downarrow \quad \downarrow \quad \downarrow$
 $A \quad B \quad ?$

n data points
 β_0, β_1 two para
d.f. = $n-2$



Sum of squares of residual. Now one thing is important is we have, there are actually three unknown parameters. There are three unknown parameters beta0 beta1 and sigma square. We estimated this by A we estimated this by B. What about this? This is the question we want to answer and that we are going to do that if Y_i is an observed value and $A + Bx_i$ is an estimated value. Then we define residual R_i as

$$R_i = Y_i - A - Bx_i$$

Then sum of square of residual is defined as summation of R_i square and you can make out that Y is a normal random variable.

$$SSR = \sum_{i=1}^n R_i^2$$


If you take $A + Bx_i$, this is also a normal random variable. Therefore the difference should also follow normal random variable with a mean 0 and therefore summation of R_i square will follow Chi square distribution and the degrees of freedom will be $n - 2$ because we had n data points.

We had n data points and we have already estimated beta0 and beta1, two parameters estimated. Therefore, degrees of freedom come to $n - 2$. So, this follows Chi square $n - 2$, and then the expected value of sums of squares of residual divided by $n - 2$ is sigma square.

(Refer Slide Time: 26:53)

Notation

- $S_{XY} = \sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y}) = \sum_{i=1}^n (x_i Y_i) - n\bar{x}\bar{Y}$
- $S_{XX} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n (x_i^2) - n\bar{x}^2$
- $S_{YY} = \sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (Y_i^2) - n\bar{Y}^2$
- Then

$$B = \frac{S_{XY}}{S_{XX}}, \quad A = \bar{Y} - B\bar{x} \text{ and } SSR = \frac{S_{XX}S_{YY} - S_{XY}^2}{S_{XX}}$$


Finally, we introduced some of the notations

$$S_{XY} = \sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y}) = \sum_{i=1}^n (x_i Y_i) - n\bar{x}\bar{Y}$$

$$S_{XX} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n (x_i^2) - n\bar{x}^2$$

$$S_{YY} = \sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (Y_i^2) - n\bar{Y}^2$$

Then

$$B = \frac{S_{XY}}{S_{XX}}, \quad A = \bar{Y} - B\bar{x} \text{ and } SSR = \frac{S_{XX}S_{YY} - S_{XY}^2}{S_{XX}}$$

(Refer Slide Time: 27:55)

Distribution of LSE parameters

• Assume that $\epsilon \sim N(\mathbf{0}, \sigma^2)$ then,

• $A \sim N\left(\beta_0, \frac{\sigma^2 \sum x_i^2}{nSxx}\right)$, and

• $B \sim N\left(\beta_1, \frac{\sigma^2}{Sxx}\right)$

• $\frac{SSR}{\sigma^2} = \frac{SxxSYy - SxY^2}{Sxx\sigma^2} \sim \chi^2(n-2)$ $\left[\left(\frac{SSR}{n-2}\right) \sigma^2\right]$



Notation

• $SxY = \sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y}) = \sum_{i=1}^n (x_i Y_i) - n\bar{x}\bar{Y}$

• $Sxx = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n (x_i^2) - n\bar{x}^2$

• $SYy = \sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (Y_i^2) - n\bar{Y}^2$

• Then

$$B = \frac{SxY}{Sxx}, \quad A = \bar{Y} - B\bar{x} \text{ and } SSR = \frac{SxxSYy - SxY^2}{Sxx}$$



The distribution of least square parameter, under the assumption that the errors are distributed as normal with the sigma square.

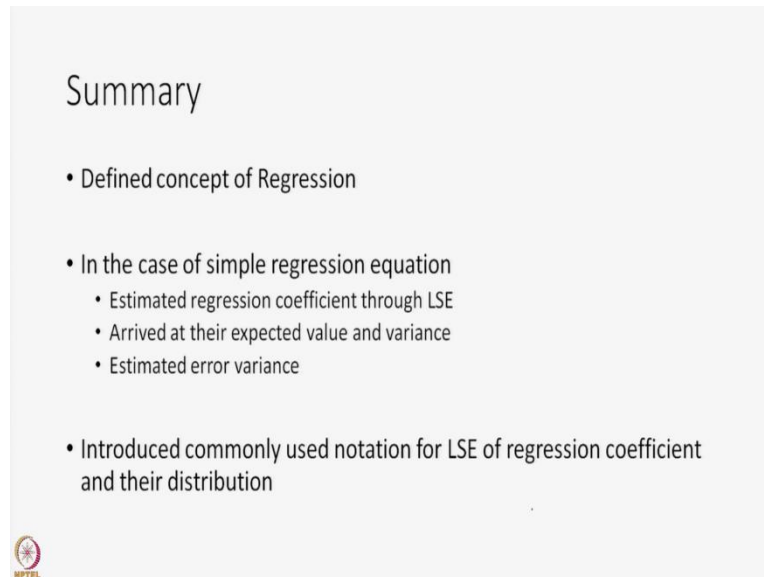
$$A \sim N\left(\beta_0, \frac{\sigma^2 \sum x_i^2}{nSxx}\right), \text{ and}$$

$$B \sim N\left(\beta_1, \frac{\sigma^2}{Sxx}\right)$$

$$\frac{SSR}{\sigma^2} = \frac{SxxSYy - SxY^2}{Sxx\sigma^2} \sim \chi^2(n-2)$$

Therefore you can write that SSR, sums of squares of residual divided by $n - 1$, $n - 1$ minus 2. Its expected value is equal to σ^2 . These are the two tables worth remembering and worth understanding this is the crux of today's lecture.

(Refer Slide Time: 29:23)

A slide titled "Summary" with a list of bullet points. The slide has a light gray background and a thin black border on the right side. At the bottom left, there is a small circular logo with the text "NPTEL" below it.

Summary

- Defined concept of Regression
- In the case of simple regression equation
 - Estimated regression coefficient through LSE
 - Arrived at their expected value and variance
 - Estimated error variance
- Introduced commonly used notation for LSE of regression coefficient and their distribution

So in summary, we defined the concept of regression. In the case of simple regression equation, we estimated the regression coefficients through least squares estimate arrived at their expected value and variance estimated the error variance introduced commonly used notation for least squares estimate of regression coefficient and their distribution.