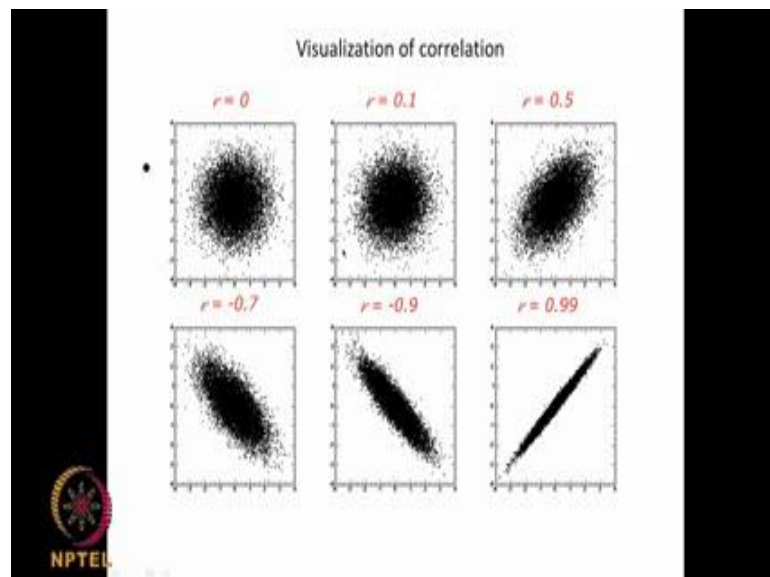


**Biostatistics and Design of Experiments**  
**Prof. Mukesh Doble**  
**Department of Biotechnology**  
**Indian Institute of Technology, Madras**

**Lecture - 39**  
**Regression Analysis**

Hello and welcome to the course on Biostatistics and Design of Experiments. Today, we will continue on the topic of regression analysis. I introduced what is the regression analysis in the previous class. Once you have collected sufficient data, you would have changed many input parameters, or factors, or variables, or axis, or independent variables as they are called, like temperature, or pH, or oxygen, or agitation, and then, you are measuring your output like biomass, or product yield. You would like to fit a mathematical relation, and that is what is called regression analysis. You can fit 1 parameter as the independent, fitting the observed data, that is dependent, or you can have 2, 2 independent variables, or 3 and so on, actually.

(Refer Slide Time: 01:13)



So, initially, if you look at how they seem to be related, or correlated as we call it, if I plot  $x$ , that is the independent variable here, and I plot the dependent variable on the  $y$  axis, and if the data appears like this, then obviously, they are not correlated.

So, when we calculate something called a correlation coefficient, it will become almost 0. As it improves, you know, when you have a good correlation, for example, as you can see, as x increases, y also increases. In such situation, (Refer Time: 1:44) we will have a correlation coefficient tending towards 1. It is, x increases, y also increases. You can, analogous to that, you can have a negative 1 also; that means, as x increases, y decreases. So, this type of figures, in, indicate that the correlation between x and y are extremely good, whereas, this type of figure indicates that the correlation between x and y are extremely poor. So, between 0 and 1, the correlation coefficient values will lie, and pictorially, you can immediately find out if there is, there is a relation. Of course, this is valid, only if you have 1 independent variable at a time, you can plot, actually; 1 independent variable at a time, we can plot and get a pictorial visualization. And, there are certain mathematical relationship, one is called the covariance; other is called the correlation coefficient, which mathematically tells you how strong is the relation between the x and the y.

(Refer Slide Time: 02:45)

Covariance is a measure of the strength of the correlation between two or more sets of random variables

$$COV(X,Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$

$$\text{Correlation} = \rho = \frac{\text{cov}(X,Y)}{\sigma_x \sigma_y}$$

Correlation is a scaled version of covariance

NPTEL

The first one is called covariance. It is a measure of the strength of the correlation between 2 or more set of random variables. So, if we have x and y, I want to see what is the covariance. Then, we do a summation of  $i = 1$  to  $n$ ,

$$\text{COV}(X,Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

that is called covariance;

$$(X_i - \bar{X})(Y_i - \bar{Y})$$

where  $\bar{x}$  is the averages of all the  $x$ 's,  $\bar{y}$  is average of all the  $y$ 's; divided by  $n - 1$ . And, you sum it up to all the  $n$  data points. Now, a correlation coefficient, correlation coefficient, sometimes, it is called Pearson's correlation coefficient and so on, is nothing but

$$\text{Correlation} = \rho = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$$

covariance  $xy$  divided by  $\sigma_x \sigma_y$ ;  $\sigma_x$  is the standard deviation for the  $x$ , and  $\sigma_y$  is the standard deviation for the  $y$ ; that means,  $\sigma_x^2$  is the variance for the  $x$ ;  $\sigma_y^2$  is the variance for the  $y$ . So, they are both related.

So, correlation is the scaled version of covariance; because, you will, you can have, covariance can be a bigger number, whereas, when you do this, correlation will always lie between 0 to 1. So, it is very convenient. Otherwise, you can have very large number; as you can see,

$$(X_i - \bar{X})(Y_i - \bar{Y})$$

, we can have number practically huge; but once you divide it by this, you are sort of making it lie between 0 and 1 only. So, generally, correlation is what is looked at and as I showed, showed in this picture, we can see the correlation coefficient is varying between 0 to, right up to 1, between the x and the y. Now, so, this correlation coefficient as I said, is also called Pearson's correlation coefficient.

(Refer Slide Time: 04:42)


**Definition**

The statistic:

$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

is called *Pearsons correlation coefficient*

**EXCEL:** pearson(array1,array2)  
 Square of correlation coefficient: RSQ(array1,array2)  
 correl(array1,array2)



So, it is nothing but covariance of x y is this, you know. And, that is also represented as  $S_{xy}$ ;  $S_{xy}$  means

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

;  $\bar{x}$  is the averages of x;  $\bar{y}$  is averages of y. Now, as I said, this is the standard deviation for x, standard deviation for y. So, this is nothing but

$$\sqrt{S_{xx}} \sqrt{S_{yy}}$$

$\sqrt{S_{xx}}$ ;  $S_{xx}$  means, instead of a y, you put all x's. Why y means, instead of x, you put all y. So, this is nothing but

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

standard deviation, and then, there will be, ok... This is called the Pearson's correlation coefficient.

So, excel also has the function called **Pearson RSQ**. There is square of the correlation coefficient or correl, that gives you array 1 and array 2; if you give the array 1 and array 2, it will calculate the correlation coefficient. So, all of them have a, this type of functions. For example, if you look at excel... So, I put in some numbers here; 1, 2, 3, 4, 5, and I get some y values; just randomly putting it. So, we can say, what is a correlation coefficient between these; correl it is called; this comma this, sorry, equal to this, comma this. So, the correlation coefficient is 0.95; that means, it is a good correlation, almost tending towards 1, like in this graph, you know, it is very high. So, even when you plot them also, you will able to see them. I hope you all know how to do plotting in excel. We can use a scatter plot. So, you can see this, right. So, data, **as x increases, y increases**; it is like a straight line. So, this number gives you the whole thing and then, **r sq, RSQ square of the Pearson correlation coefficient**; that is, square of 0.95; square of 0.95; that is, 0.95 into 0.95; that gives you 0.91. **This is called R square**. Generally, we use this as, when you are fitting data, use this. This, the top one, correl we use when we are trying to just see the correlation coefficient. So, both are interrelated with each other. So, this is called the Pearson's correlation coefficient.

(Refer Slide Time: 07:53)

**Regression Models**

- Types of equations
  - System specific
    - Based on specific knowledge of the system
    - Exponential decay  $y_t = b_0 + b_1 e^{-t/\tau}$
    - Sine-waves  $y_t = b_1 \cos(\omega t) + b_2 \sin(\omega t)$
  - Generic
    - a polynomial equation
    - Examples:
      - simple linear regression  $y_t = b_0 + b_1 x$   $y = b_0 + b_1 x_1 + b_2 x_2$
      - single-parameter cubic equation  $y_t = b_0 + b_1 x + b_2 x^2 + b_3 x^3$
      - two-parameter quadratic  $y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_1 x_2 + b_4 x_1^2 + b_5 x_2^2$
      - many-parameter quadratic 
$$y = b_0 + \sum_{i=1}^n b_i x_i + \sum_{i=1}^{n-1} \sum_{j=i+1}^n b_{ij} x_i x_j + \sum_{i=1}^n b_i x_i^2$$

Excel also has these things. Type of regression models; like I said, we can have any type of regression models. If you know the physics, you can have, say some sort of an exponential decay, drug release from a, from a drug carrier may follow exponential decay model, **sine** waves, we can have **cos and sine**; or you can fit this type of binomial relationship **simple regression**

$$y_x = b_0 + b_1 x$$

$$y = b_0 + b_1 x_1 + b_2 x_2$$

, or you can have single parameter cubic  $x^2$ ,  $x^3$  like this, or you can have 2 parameters  $x_1 \times x_2$ , then  $x_1 \times x_2$ ,  $x_1$  square; or, you can have large number of parameters. So, different types of models are possible and finally, the idea is to get all these constants, these bs are all called constants, actually.

(Refer Slide Time: 08:42)

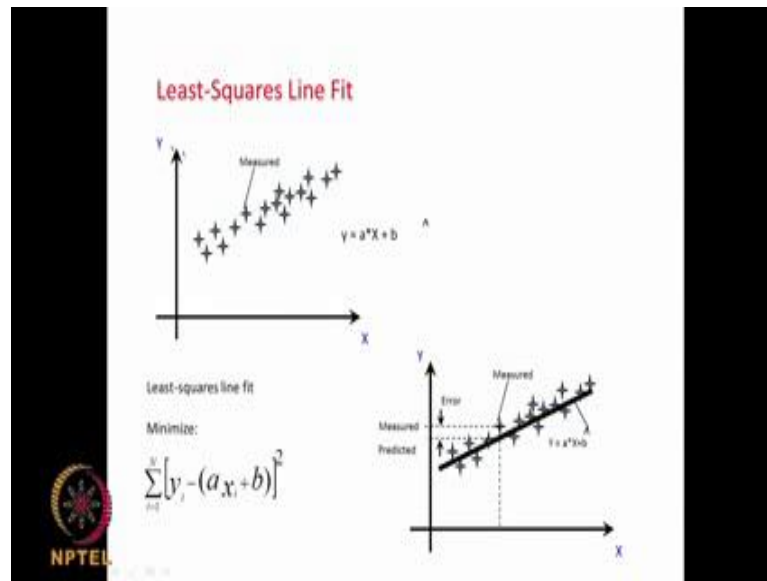
The slide is titled "Data fit" in red. Below the title, it says "minimize the sum-of-squares error:". The formula for SSE is given as  $SSE = \sum_{i=1}^n (y_{\text{meas},i} - y_{\text{model},i})^2$ . A callout box with an arrow pointing to the term  $(y_{\text{meas},i} - y_{\text{model},i})$  contains the text: "The  $(y_{\text{meas}} - y_{\text{model}})$  term is called an error or a residual." In the bottom left corner, there is a circular logo with "NPTEL" written below it.

So, what you are doing is, you are trying to reduce, minimize the sum of squares of the error. What is sum of squares of the error? The data measured and what the model predicts. So, model predicts, suppose I write this as a model, for any given value  $x$ , it will predict a  $y$ . So, that is the model predict; this is the measured value; take a square, and **if you summation**, that is called a error, and you want to minimize the error. So, this is like an optimization problem, where you are trying to minimize

$$SSE = \sum_{i=1}^n (y_{\text{meas},i} - y_{\text{model},i})^2$$

**the sum of squares of this error. This is called the error**, because this is  $y$  measured, this is  $y$  model, and the difference squared summation is called either error or residual; these are the two names for that, actually.

(Refer Slide Time: 09:27)



So, if you have x and y, you are trying to fit a straight line,

$$y = a \cdot X + b$$

, and that means, you are trying to find a and b, so that,

$$\sum_{i=1}^N [y_i - (ax_i + b)]^2$$

$i$  is equal to 1 to 10, is minimized. So, this is basically, it is almost like a optimization problem; that is what you are doing actually. You calculate a and b, so that, this is equal to 0. Now, excel has 2 commands; one is called slope; other is called intercept. This is the slope. This is the intercept. To find out the slope and intercept, for the x and y data, suppose so, let us take the same data; I have a x and have the y. So, if I want to calculate slope, so, I just say, slope. This is the x data, and this is the y data. So, this is the slope of this line; and if you want to get the intercept, so, this is the x data; this is the y data; that is the intercept. So, it is having an negative sort of intercept, because it goes right down, sorry, intercept c 2, c 2, slope and intercept.



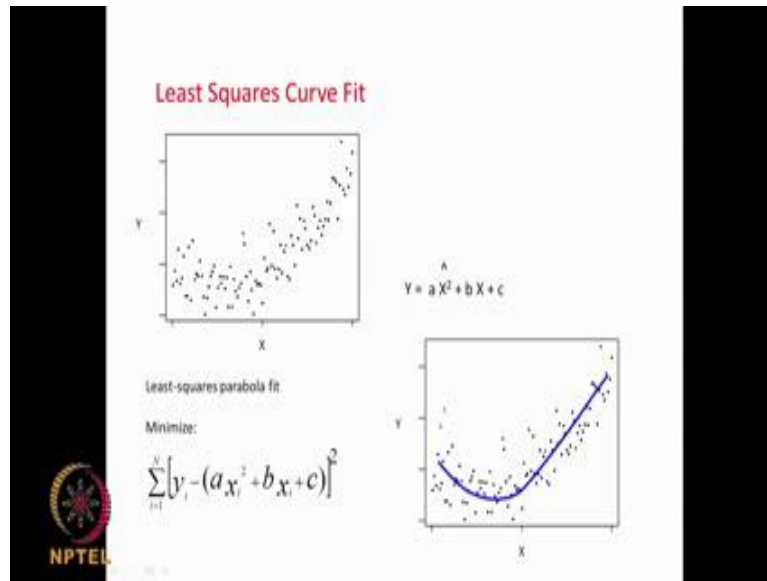
So, we can also fit it using add a trend line command, and, we can calculate display equation R square and close. So, it gives you slope and intercept. So, I have made a mistake here, I think. So, the slope command, through the given data set. So, I have to give the **y's** first, that is the mistake I made. I have to give the y s first; I have to give the **x's** later. That gives you 2.4, intercept. Please remember, I have to give the **y first, and the x later**. So, 17.6. So, this is the slope; this is the intercept. So, this is the slope, and this is the intercept. So, we can use the excel command slope and intercept to calculate, you can use the slope and intercept to calculate the slope and intercept of a data point. So, do not forget, you need to give the **y's first, and then the x**; whereas, for the correlation it does not matter, whether you give the x first or y later. But interestingly, I do not know why we want the **y's first, and then the x's later**. So, if it is a linear regression, that means, if you are trying to fit a linear data, it is very simple; we can use slope command, or intercept command, or even I showed you with the graphics, we can just draw a graph and then, we can say, draw trend line, and find our equation. So, excel can very well do. Of course, there are many softwares, hundreds of softwares, commercial softwares on payment basis, which can do all these fancy things, actually; it is not a big deal.

(Refer Slide Time: 13:00)



So, you have to give slope, the y first, then the x later; intercept also, we give that.

(Refer Slide Time: 13:11)



So, if it is a least, if it is a second order, like a

$$Y = aX^2 + bX + c$$

so then, you are minimizing

$$\sum_{i=1}^N [y_i - (ax_i^2 + bx_i + c)]^2$$

I think you can do that also in excel. We can once draw this, but, we cannot calculate it as slope and intercept, but we can draw it, and we can try to fit a second order or higher order type of a polynomial in excel also. So, what are the steps in regression?

(Refer Slide Time: 13:42)

**Steps in Regression**

- Specify the form of the regression model
  - Linear or quadratic?
  - Which two-factor interactions ( $X_i \times X_j$ ) should be included?
- Perform the regression analysis
  - Compute the regression coefficients – mean & variance
- Compute statistics ( $R^2$  and  $R^2_{adj}$ )
- Compare predicted values vs actual values
  - How well does the regression model fit the data? Check the residuals.
- Determine if a response transformation improves the fit
  - Would a regression model for  $\ln(Y)$  or  $\sqrt{\text{qt}}(Y)$  or some other function of  $Y$  improve the fit?
- Identify significant terms
  - Does inclusion of a term in the regression model significantly improve the fit of the model to the data?
- Remove insignificant terms
- Final assessment of the regression model

NPTEL

We have to specify what should be the regression model, linear or quadratic, perform the regression analysis and then, compute the statistics. There are many statistics  $R^2$  and  $R^2_{adj}$ ,  $R^2_{pred}$  we will come to that, and compare predicted value with the actual value; for example, like I said, it gives you an idea about the error. What is error? **Summation of  $y$  calculated by  $y$  actual square**. So, error should be minimum. So, it gives you an idea about that. And, if the fit is not good, we can transform it. We can take logarithm of the data; especially, if you look at drug discovery, normally, the concentrations you are using is in millimole or micromole, and the response could be some activity. So, normally, they take a logarithm of the concentration, because, millimole or micromole may be very small number.

So, when you take logarithm, you can bring it down to 1, 2, 3, 4, that sort of thing. So, most of the drug response curves will have logarithm, for the drug concentration. Then, identify which terms are important. Suppose, if you are fitting a multi linear relationship, suppose, you are fitting a multi linear relationship, some terms may be very important, some terms may not be important. So, you can identify which are very important, and you can also plan whether there are any insignificant terms which can be omitted. There could be some terms which may have very very small absolute value, then, you can remove that, and then, again redo the regression model until you get a good regression

model. Good means, the error should be minimal. This statistic should be quite high. All these generally vary between 0 to 1. So, higher the number, better is supposed to be your fit.

(Refer Slide Time: 15:36)

Suppose we want to fit an equation of the form

$$\hat{y} = a + bx + cx^2 + pe^x + q \sin x$$

to our data points.


We have N data points  $(x_i, y_i)$

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \\ \vdots \\ \hat{y}_N \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 & e^{x_1} & \sin x_1 \\ 1 & x_2 & x_2^2 & e^{x_2} & \sin x_2 \\ 1 & x_3 & x_3^2 & e^{x_3} & \sin x_3 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_N & x_N^2 & e^{x_N} & \sin x_N \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ p \\ q \end{bmatrix}$$

Or, in matrix form:

$$\hat{\mathbf{y}} = \mathbf{X} \mathbf{b}$$

where  $\mathbf{b} = (a, b, c, p, q)$



So, in matrix term, this is what it is all about. Let us not go too much into that.

(Refer Slide Time: 15:41).

### Computing the Regression Coefficients


Minimize  $\Phi = (\hat{y} - y) \cdot (\hat{y} - y) \leftarrow$  Sum of squares of the distances between the predicted and actual y-values.

$$\frac{\partial \Phi}{\partial b} = 0 \quad \text{is the necessary condition that } \hat{y} \text{ be a minimum}$$

$$\Phi = (\mathbf{X}\hat{b} - y) \cdot (\mathbf{X}\hat{b} - y)$$

$$\Phi = \mathbf{X}\hat{b} \cdot \mathbf{X}\hat{b} - 2\mathbf{X}\hat{b} \cdot y + y \cdot y = \hat{b} \cdot \mathbf{X}^T \mathbf{X} \hat{b} - 2\hat{b} \cdot \mathbf{X}^T y + y \cdot y$$

$$\frac{\partial \Phi}{\partial \hat{b}} = 2\mathbf{X}^T \mathbf{X} \hat{b} - 2\mathbf{X}^T y = 0$$

$$\hat{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T y$$


So, if you want, if you are trying to calculate the a and b, that is, suppose, the slope and the intercept, if you look at it from calculus point of view, we are integrating with respect to, say b, or integrating with respect to a, and then, you are equating it to 0. So, when you do that, we will get an equation for b; and for a, it is very simple, because  $a + x b = y$ . So, it is very simple.

(Refer Slide Time: 16:11)

The Regression Coefficients are Random Variables

mean values given by:


$$b_i = [(X^T X)^{-1} X^T y]_i$$

variances given by:

$$\text{Var}(b_i) = S^2 (X^T X)^{-1}_{ii}$$

where S is the standard error.

This standard error estimate provides an estimate of the average variance in the predicted values in the design space.



(Refer Slide Time: 16:14)

Suppose that we have 4 factors: X1, X2, X3, and X4, and that we want a quadratic fit. We measure the response value at 15 points.

How many terms will the response surface equation have?

- 1 constant
- 4 linear (Xi)
- 6 linear interaction (Xi\*Xj)
- 4 quadratic (Xi^2)


---

Total = 15 terms

The regression equation has 15 unknown coefficients, and we have 15 data points to use to compute them.

Error df = 0

We'll get a perfect fit, with  $R^2 = 100\%$



Suppose, we have 4 factors X1, X2, X3, and X4 and we want to fit a quadratic fit, that means, it will have a + b X1 + c X2 + c X3 + c X4 + d, and then, plus, that is, a + b X1 + c X2 + d X3 + e X4, and then, f X1 X2 g X1 X3 h X1 X4 i X2 X4 j X3 X4 k X3 X4, and then, like that, if we, and then, you do that X1<sup>2</sup> X2<sup>2</sup> X3<sup>2</sup> X4<sup>2</sup>, what will happen is, you will have 1 for constant, 4 for the X1 X2 X3, the main effect, 6 for the, the 2 level interactions, and then, 4 constants for the quadratic; these all adds up to 15. 1 plus 4, 5, 6, for 15. And, you have got only 15 data points. So, there are no error, degrees of freedom at all; it will become 0. So, if you try to fit that sort of data, then of course, you will get very good fit, because that is no degree of freedom at all. So, you better be watching out. You need to have enough degree of freedom for the error; do not forget that. So, depending upon how many constants we have, and the dependants, so, subtract that from the total data points you have, and check whether the degrees of freedom is reasonably good; otherwise, your fit, you cannot rely on the fit. You may end up having 100 % fit, but in fact, the error does not have a degree of freedom.

(Refer Slide Time: 18:02)

**Goodness of Fit**

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

$\sum_{i=1}^n (y_i - \hat{y}_i)^2$  ← Sum of squares of the residuals  
 $\sum_{i=1}^n (y_i - \bar{y})^2$  ← Variance

*R<sup>2</sup> is the fraction of the total variance that is explained by the regression equation.*

Will lie between 0 to 1

NPTEL

So, there are many terms such as R square, R square adjusted, R square predicted, which give you an idea about the quality of the fit between the x and y. The first one is called R square. And, this is equal to

$$R^2 = 1 - \frac{SSE}{SST}$$

. This is equal to  $1 - \frac{SSE}{SST}$ , what is the sum of squares of the error, I told you, the, this is the predicted; this is the actual; this is the predicted, actual. So, this is the error, squaring up summation and the denominator has the variance for y,  $(y_i - \bar{y})^2$ ;  $\bar{y}$  is the mean of y. This is called  $R^2$ . This is like a, this is as predicted by the model; this is the total variance; this is the total variance, as predicted by the model.

So, if it becomes 1, that means, the entire variance can be predicted by the model. So, closer to 1, that means, large proportion of it could be predicted by the model; you understand. So,  $1 - \frac{SSE}{SST}$  - sum of squares of the error, that is given by y predicted, sorry, y predicted and y observed square, and this is the sum of squares of the total variance,  $(y_i - \bar{y})^2$ ; and, 1 minus of that, actually. So generally, it will lie between 0 and 1; closer to 1, it means, the fit is extremely good; closer to 0, that means, it is very poor.

(Refer Slide Time: 19:41)

*Degrees of Freedom*

# Degrees of Freedom = # Model Parameters - 1

The number of points should be at least ~25% more than the number of degrees of freedom in the regression model.

NPTEL

So, the degrees of freedom is number of parameters **minus 1**, as I mentioned sometime back also. So, you need to have, you cannot have parameters equal to the number of experiments, or the number of experiments just matching with the parameters; in that

case, you will, you will not have any degrees of freedom. So, if I am going to fit a linear relation like  $y = ax + b$ , you have 2 parameters; if you do only 2 experiments, that is not good; because your d f will be 0. So, you should have done at least, minimum 3 experiments; do not forget that. So, but then, there is a rule of thumb; number of point should be at least 25 % more than the number of degrees of freedom. So, if you take  $y = ax + b$ , I have 2 parameters, and so, number of points should be at least 25 % more than number of the degrees of freedom in the regression model. So, I have 2. So, I should, even if I do one, at least 3 or 4, then its good actually.

(Refer Slide Time: 20:45)

**Adjusted  $R^2$  Statistic**

$$R^2_{adj} = 1 - \frac{SSE(n-1)}{SST(n-p)} = 1 - \left(\frac{n-1}{n-p}\right)(1 - R^2)$$

where:  $n$  = number of points  
 $p$  = number of terms in the regression model

$R^2_{adj}$  is a better indicator of the ability of the regression model to accurately explain the observed variance in the data.

NPTEL

Next,  $R^2$  is,  $R^2_{ad}$ .

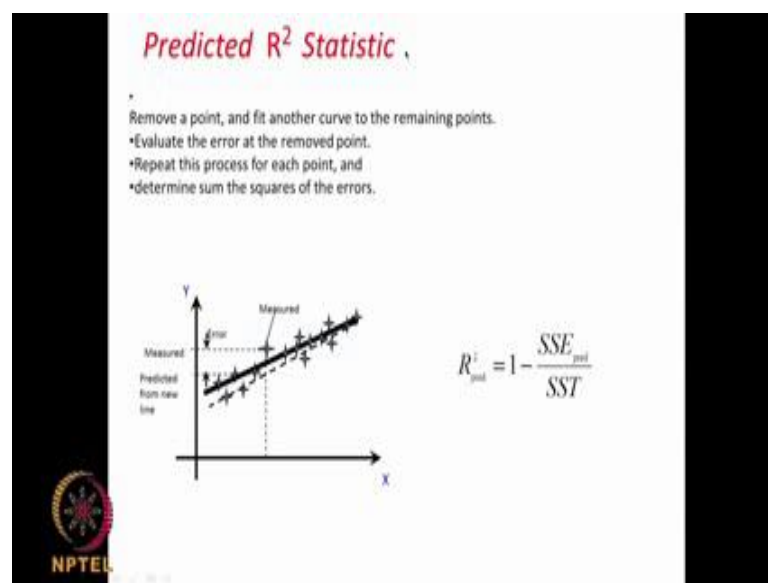
$$1 - \frac{SSE(n-1)}{SST(n-p)}$$

This is 1 minus sum of squares of the error multiplied by n minus 1, / sum of squares, total sum of squares multiplied by n minus p; n is the number of points, data points, and p is the number of terms in the regression model. So, if you have  $y = ax$  plus b, p will be 2; if you have  $y = ax^2 + bx + c$ , then, p = 3. So, this can be rearranged also;  $1 - n - 1 / n - p, 1 - R^2$ . So, this model tells you something interesting, because you can always add terms and try to improve your  $R^2$ . But then, if you calculate  $R^2$ , if you add more terms,



p will become large. So, this will become small. So, this whole term becomes large. So,  $1 -$  this whole term will become small. So,  $R^2_{adj}$  will keep going down, although  $R^2$  will go up. So, p is an indication of, whether you are doing an over fitting of your data. So,  $R^2$  gives you an first level of a data fitting, but,  $R^2_{adj}$  takes care of the number of parameters you have in your model; so, it adjusts for that here. So, that is a better indication than  $R^2$ ; and if I keep increasing p, we will always keep increasing your fit, but then,  $r$ ,  $R^2$  will increase, but  $R^2_{adj}$  will decrease, because of p being here; that is called adjusted  $R^2$  statistics.

(Refer Slide Time: 22:33)



There is something called **predicted  $R^2$** . So, this **predicted  $R^2$**  is an indication of the predictive capability of your model. Ultimately, I told you long time back, once I have a model, I can use it for predicting, at different conditions. For example, if I have a model, yield of the bio, of the metabolite as a function of pH and temperature, a plus b p H plus c temperature, I do lot of experiments, and then, I get the model for a b c. Now, I can use this model to predict what will be my yield of the metabolite at different pH values and different temperature. So, ultimately, I am going to use this model for predicting. So, this  **$R^2_{predicted}$**  gives you an indication of how good the model is able to predict; whereas the remaining, , previous  **$R^2$ ,  $R^2$  and  $R^2_{adjusted}$** , are talking about how good the fit is, whereas the R square predicted tells you how good the predictive capability of the

model. So, how do you do this? What you do is, suppose, you have 10 points, what you do is, you remove 1 point, and fit your model, fit your relationship, say  $y = x + b$ , to, to the 9 points. And then, using that, you try to predict for the 10<sup>th</sup> point. So, of course, the prediction will not be exact. So, there will be some difference; you square it, that we call it error. Then, you put that point back, you remove another point, and then, with the new set of 9 points, you fit a, fit an equation, and then, try to predict the  $y$  value for the data point which was missed out. Again, you will get some error; you square it. So, you keep on doing that for each of the 10 points, and then, you square those errors, add up, that is called the sum of squares of the predicted; understand; that is called the sum of squares of the predicted. So,  $R^2$  square predicted

$$R_{\text{pred}}^2 = 1 - \frac{SSE_{\text{pred}}}{SST}$$

What is this? This is summation of  $(y_i - \bar{y})^2$ . Do you understand how it is done? So, if you have 10 data points, what you do is, you remove the 10th point, and then, fit an equation for the 9 points, try to predict the  $y$  for the 10th point using the equation you have developed. Now, of course, it will not be exact; there will be some difference; that is, error is there. You square it up. Now, you put the 10th point back; remove the 9th point, and then, fit in equation for the remaining 9 points, and try to predict the  $y$  value for the 9th point. Again, you will get some error; square it up. Like that, you keep on removing 1 point at a time. Of course, you put the other points back; then you find out, try to predict the  $y$  at that point using the model. And then, error, you square it up; you add all these errors; sum of square of the error, that is called the sum of squares of the error of the predicted; divide it by the sum of squares of the total, that is  $(y_i - \bar{y})^2$ ; subtract from 1; you will get the  $R_{\text{pred}}^2$ . This is very stringent  $R^2$ . It gives you a better understanding of the predictive capability of the model.

So, if you fit a data and generally,  $R^2$  will be high;  $R^2$  adjusted will be lower and  $R^2$  predicted will be still lower. So ideally, if I am fitting a model, I expect all these 3  $R^2$  squares to be above 0.6. You know, that is a reasonable model, especially with the experimental system. If you might, you will never get 0.9, or so on, for experimental data study. So,  $R^2$  will always be high, and  $R^2$  adjusted will be lower, and  $R^2$  predicted will

be still lower. And ideally, all these **3R square** should be above 0.6. Then, you can be reasonably sure that, your model is reasonably good. So, these are the statistics, statistics which we need to keep in mind to understand the predictive capability of the model, the fitting of the data and effect of parameters on the fitting of the data, ok.

(Refer Slide Time: 26:54)

**Residuals**

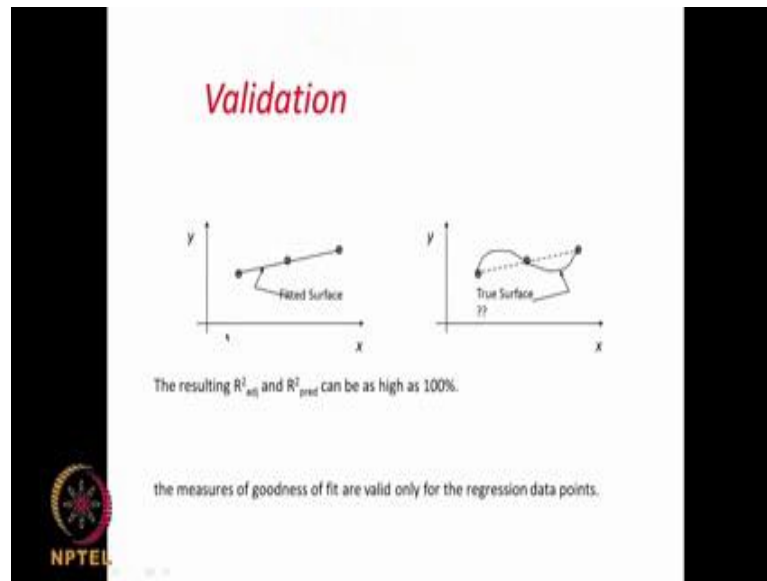
Difference between the actual observation (or data point) and the fitted value (average of the factor level).

- Residuals should be random and normally distributed (we can perform a test for normality)
- A pattern generally implies some sort of model insufficiency
- Unexplained outliers

The slide contains two small residual plots. The left plot shows a scatter of points around a horizontal line, representing a good fit. The right plot shows a clear upward trend in the residuals, indicating a positive bias or a non-linear relationship not captured by the model.

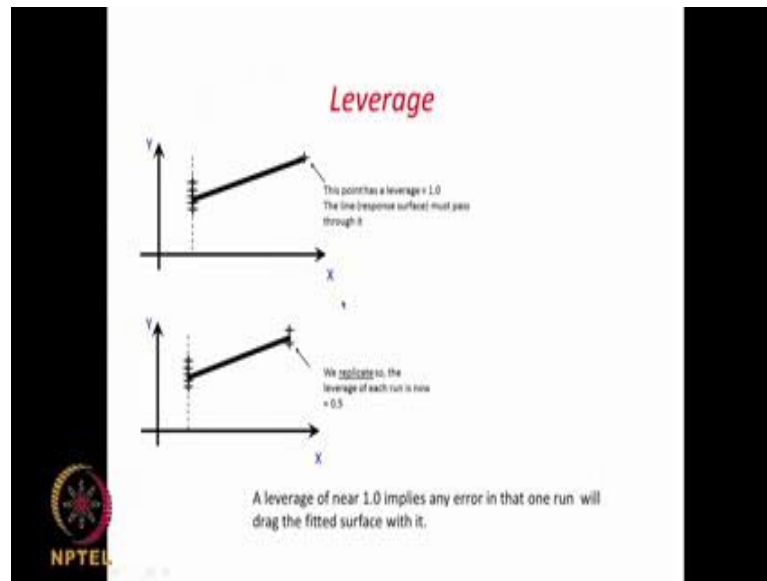
Once you do the fitting, you look at the residuals. That is, residual is nothing, but the difference between the predicted and the actual. Generally, these residual should be random; that means, if there are 10 residuals, some of them should be plus; some of them should be minus; it will not be all pluses, all minus. Then, you should be, you are very sure that there is a mistake; I mean, there is some catch, problem in your fitting. So generally, it should be normally distributed. So, we can even do a test for normality to see. It should not be biased in all pluses all minuses. And also, pattern; you should not get residuals like this; you know, it should generally be like this; you know, almost, some of them small positives, some of them small and negative. And also, you should not have unexplained outliers; suddenly, one of the data, one of the residual is very high. Suppose, you have ten data points, you are fitting it, you are calculating the residuals. So, nine residuals are reasonably small, small, small, some of them positive, ant then, one of them is extremely high. Then obviously, there is some problem. We need to really look at it, look at that data point. It is called unexplained outliers.

(Refer Slide Time: 28:13)



Then, of course, once you have done the model fitting, you need to validate; you need to see, whether it is able to predict for some new axes; that is very very important. Unless you do that, you are not sure your model is good, because, you may be, suppose, you have 3 points. You fit  $x$  and  $y$ ; you think it is a straight line,. But in reality, it could be like this. So, you do not know, although you think by looking at these 3 points, they all fall in a nice line. So, you fit a line like  $y = a x + b$ , straight line; but in reality, it may be going up like this, and coming down like this, right. So, you have to be very careful on that actually. So, we need to have a validation. Do not forget that, we need to have a validation. That is one point. Another point is, we cannot extrapolate, because, that fit you do is based on the experimental data and you cannot extrapolate from this fit, at some other point which is wrong, because, you are not sure how this graph will change. In some region, it could be linear; you may think it is a linear relation, but actually, the graph may be going up; especially if you look at the biomass, exponential growth phase, all those things happen, right; stationary phase. So, whatever region in which you have collected data and fitted, you have to be sure, you have to use the model only in that range; be sure about it; be careful.

(Refer Slide Time: 29:44)



Then, there is something called leverage. This is very important. I need to talk, mention about it. Suppose, you have collected data and suppose, this is your plot; you have lot of x data collected around this place. And, there is one x here. Now, you are fitting a line. What is the danger of this? This particular point is leveraging the entire line, slope of the entire line. Suppose, if the, if I have made a mistake in this data collection, the slope will change because of this; whereas, even if you make some mistakes here, on these points, there will not be much change, because you have many points; whereas, you have only one points. So, this point leverages your entire slope of this line. So, such an experiment is extremely bad. You should have done at least 2 experiments, so, even if one of them is partially wrong, you are going to take a mean, so, the slope will not change too much. So, the leverage here is only 0.5. You have 2 points, leverage is 0.5. A leverage of 1 implies that, any error in collecting data of this point is going to change the slope; exactly; whereas, if you have 2 points, and any error, we will get only half. If you have many points, 3 or 4, then obviously, it will keep going down 1 by 3, 1 by 4. So, that is very important. So, when you collect data, it should be well spread out, and you should not have points which have a very high leverage like this, watch out.

(Refer Slide Time: 31:23)



*Changes to Get a Better Fit*

- Use a higher-order model
- Drop superfluous terms from models
- Use transforms on parameters and/or data
- Use non-dimensional combinations of parameters
  - For example, use Reynold's number, aspect ratio (length/width) instead of length and width separately

NPTEL

So, how do you change the model to get a better fit? Use a higher order model. You can drop superfluous terms. So, if you have  $y = a + bx + cx^2 + dx^2 e^{ex^2}$  and so on, so, you can look at some of those a, b, c, d, e and if some of them are very, very small, you can drop them out. You can use transforms, like I said, in drug discovery, concentrations are always in micromole which is small, 0.0001 mole and so on. So, you can take a logarithm of that. Sometimes, you can take 1 by that, especially, again in drug discovery, we use 1 by. So, transformation. We can use, use even non dimensional combinations. If you are an engineer, you will know things like Reynolds number. You can even take aspect ratio, length by width, that sort of non dimensional combination that will improve your model fit.

(Refer Slide Time: 32:24)

**Common Problems**

Reverse Causation:  
Storks are observed near the home of new born babies

Simple Chance: June usually brings a peak in the number of both weddings and suicides. Which is the cause?

Selective Observations / Selective Memory:

- "My spouse never listens to me!" (forgetting 98% of the observations where listening did occur).

Omitted Factors (other X's):

Multicollinearity: (Two or more X's confounded)

NPTEL

So, some of the common problems that one comes across, reverse causation; this is a very interesting problem. So, in Europe and England, they used to see lot of storks. These are birds, migratory birds observed near the home of new born babies. So, did these storks bring the new born babies? June usually brings a peak in the number of both weddings and suicides. Especially, in the western countries, weddings are held during summer time. So, again, suicides are also very high. So, which is the cause? Selective observation, selective memory. My spouse never listens to me, forgetting that 98 % of the observations were listening did occur. Even if the spouse had been listening to 98 % of the time, the 2 % of the time when the spouse does not listen is what hits the person, and that gets embedded in the thought process.

So, the person may make a statement, my spouse never listens to me, but that could be only 2 % of the time. The remaining 98 % of the time, the spouse may be limit, listening. Omitted factors. There could be some other x s, you know; I may think temperature and pH and carbon is important. There could be a magnesium, or some other micro nutrient which may be helping the growth of the organism which is called omitted factors. Multi-collinearity. 2 or more x's are confounded. Sometimes, x's may get confounded. Suppose, I am looking at weight, height, as independent variable, and I am measuring the, the blood glucose level, but then, weight and height may be

confounded. Taller people may be more heavier than shorter people, because height also has some weight involved in it.

So, those 2 x's are confounded. So, unless you adjust for that, then, that could be multicollinearity. For example, the blood pleasure increasing with, say lipid content, but then, even age, increase in age, also may increase the blood pleasure. So, there could be a confounding between the age and the lipid content. So, you need to remove the effect of age, so that, if you want to study only the lipid content, then, that is correct. Otherwise, it is getting confounded. So, the multicollinearity is another problem. So, these are common problems you have to keep in mind, when you are doing a regression fit.

(Refer Slide Time: 35:10)

Example:  
 $Y = b_0 + b_1 x$   
 (5 data points)

Factor	SS	Df	Ms	F(cal)	F(table)
Regression	12.1	1	12.1	0.1	
Residual	326.7	3	108.9		
Total	38.8	4			

Residual ss =  $\sum (Y_{data} - Y_{model})^2$   
 Total ss =  $\sum (Y_{data} - \bar{Y})^2$   
 Regression ss =  $\sum (\bar{Y} - Y_{residual})^2$

NPTEL

So, if you are, suppose, you are fitting like this,  $Y = b_0 + b_1 x$

. So, the, you have the, there is a residual, that is your sum of squares of the error. Suppose, I have 5 data points, and I have 2 constants. So, the regression will have 1 degree of freedom; 5 data points, so, 4 degrees of freedom and the residual will have 3 degrees of freedom. So, sorry, 4, 3, 1. So, the total sum of squares, as I said,



$$\text{Total ss} = \sum (Y_{\text{data}} - \bar{Y})^2$$

; residual sum of squares = y predicted minus y actual square. So, from this difference, we can calculate, we can calculate the regression sum of squares.

$$\text{Residual ss} = \sum (Y_{\text{data}} - Y_{\text{model}})^2$$

. Total sum of squares is

$$\text{Total ss} = \sum (Y_{\text{data}} - \bar{Y})^2$$

. So then, we can calculate the mean sum of squares by dividing, as you know, the degrees of freedom. Then, we can calculate f value, regression divided by residual. And then, from the f value and f table, we can say, whether the, what is the significance, whether we can calculate p. The  $h_{naught}$ , the  $h_{naught}$  is, there is no relationship;  $h_{naught}$ ,  $h_1$ ,  $h_a$  is, there is a relationship.

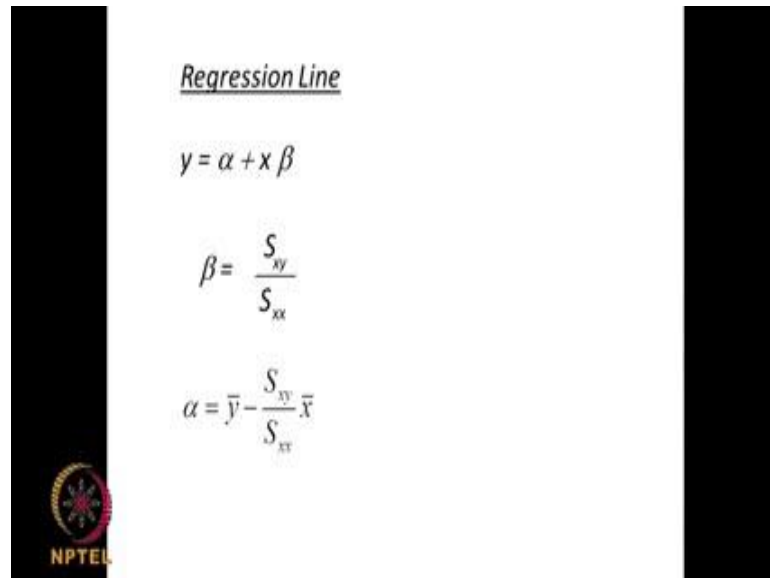
So, if you get p, sorry, if you get f table value lesser than the f calculated value, obviously, we can say, there is a relationship. This is what you do. So this is, there is an ANOVA which is created when you do a regression analysis. There is an ANOVA which is created when you do an regression analysis, and there are 3 sum of squares; one is called the regression sum of squares,

$$\text{Regression ss} = \sum (Y - Y_{\text{model}})^2$$

which is equal to y, y minus y model, that is a residual sum of squares, which is y data - y model prediction; total sum of squares is y data - y bar square; y bar, bar is nothing but average of y. And then, you have degrees of freedom. So, total degrees of freedom will be total data points minus 1 and the regression degrees of freedom, if you have y is equal to linear fit, regression will have one degree of freedom. So, the remaining will go to the residual. And then, you calculate f calculated by dividing the regression by the residual,

and then, see whether this number is larger than the f table value; if it is larger, then, we can say, there is a, the regression relationship is valid; otherwise, we can, we will say that regression relationship is not significant. This is how you do the regression analysis on the corresponding ANOVA table.

(Refer Slide Time: 38:10)

A slide with a black background. On the left side, there is a vertical black bar containing the NPTEL logo at the bottom. The main content of the slide is centered and consists of the following text and equations:

Regression Line

$$y = \alpha + x \beta$$
$$\beta = \frac{S_{xy}}{S_{xx}}$$
$$\alpha = \bar{y} - \frac{S_{xy}}{S_{xx}} \bar{x}$$

Now, suppose, this is a regression relationship,

$$y = \alpha + x \beta$$

$\alpha$  is the constant term and  $\beta$  is your slope. So,  $\beta$  is given by

$$\beta = \frac{S_{xy}}{S_{xx}}$$

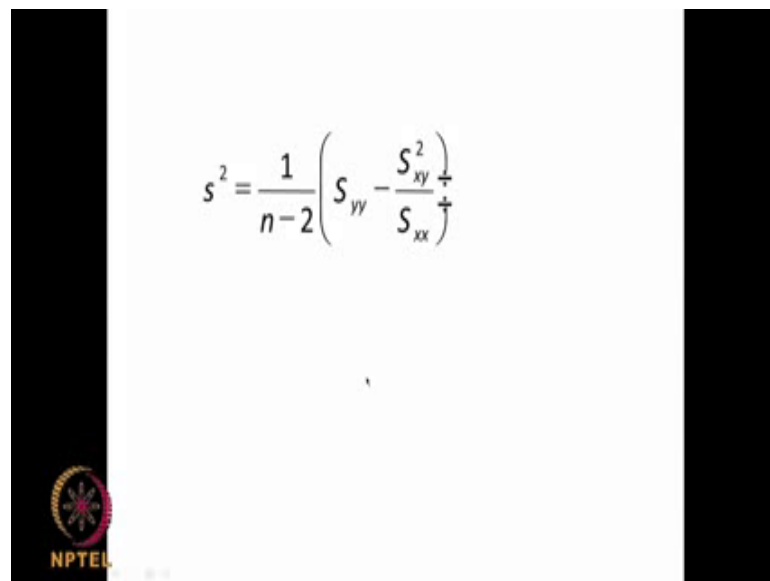
---

. Then, the alpha is given by

$$\alpha = \bar{y} - \frac{S_{xy}}{S_{xx}} \bar{x}$$

. So, this is how we calculate both the  $\alpha$  and  $\beta$ . Of course, I showed you in that excel, for slope, you give, give slope and intercept, we give intercept. So, the slope is nothing but  $S_{xy}$  divided by  $S_{xx}$  and the intercept is nothing, but this. This is how the software also calculates. Of course,  $\alpha$  and  $\beta$  then will have a region of confidence, because it is not single.

(Refer Slide Time: 39:04)


$$s^2 = \frac{1}{n-2} \left( S_{yy} - \frac{S_{xy}^2}{S_{xx}} \right)$$


NPTEL

(Refer Slide Time: 39:07)

$(1 - \alpha)100\%$  Confidence Limits for slope  $\beta$ :

$$\hat{\beta} \pm t_{\alpha/2} S_{\hat{\beta}}$$
$$\hat{\beta} \pm t_{\alpha/2} \frac{S}{\sqrt{S_{xx}}}$$

$t_{\alpha/2}$  critical value for the t-distribution with  $n - 2$  degrees of freedom



So, confidence limits for the slope, this is your slope, sorry, this is your slope;  $S_{xy}$  by  $S_{xx}$ . So, it has got a confidence limit,

$$\hat{\beta} \pm t_{\alpha/2} S_{\hat{\beta}}$$

$$\hat{\beta} \pm t_{\alpha/2} \frac{S}{\sqrt{S_{xx}}}$$


with  $n$  minus 2 degrees of freedom;  $S_{xx}$ , you know;  $S$  is the standard deviation given like this,  $S^2 = 1 / n - 2, S_{yy} - S_{xy}^2 / S_{xx}$ ; understand. So, you understand this, this parameter  $S$  is given like this;  $S_{xx}$  is summation of  $x - \bar{x}$  squared, that is what it is. This is for  $S^2$  and this is  $t_{\alpha, 95\%}$  if you can put, with  $n - 2$  degrees of freedom. Similarly, the intercept also will have a confidence limit that is given by this type of relationship. The intercept calculated  $\pm t_{\alpha/2}$  square root of  $1 / n + \bar{x}^2$  by  $S_{xx}$  with  $n - 2$  degrees of freedom. So, slope and intercept, because, after all the slope and intercept is single value which we have calculated, but then, as we have been doing, we should, it is like an  $\bar{x}$ , you know. You remember, we used to always give a confidence region limit.

(Refer Slide Time: 40:42)

$(1 - \alpha)100\%$  Confidence Limits for a point on the regression line :

$$\hat{\alpha} + \hat{\beta}x_0 \pm t_{\alpha/2}S\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

$t_{\alpha/2}$  critical value for the t-distribution with  $n - 2$  degrees of freedom



The confidence limit for a point on the regression line, so, if I fit a regression line, so at any point  $x_0$ , it will give as

$$\hat{\alpha} + \hat{\beta}x_0 \pm t_{\alpha/2}S\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

$\hat{\alpha} + \hat{\beta}x_0$ , that means, it is predicting the  $y$  naught at a given  $x$  naught, this is your regression line. This also will have a confidence, because it will not be absolute; that is given by

$$t_{\alpha/2}S\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

$n$  minus 2 degrees of freedom, t distribution. So, if I am predicting a  $y$  naught at a value of  $x$  naught, so, it is not just the slope, sorry, the slope and the intercept alone, there is a plus or minus term coming in here, because, there is always a confidence limit; do you understand? 95 % confidence limit is associated with this, and that is given by this. So, you need to consider all these confidence limits for this slope, confidence limits for the

intercept, as well as confidence limit for a point on the regression line as well. So, we have completed the regression which is a very important topic. Once you have a data, you need to fit the data to a linear or a non-linear model with the  $x$ 's as your factors, or independent variables and  $y$  is your dependent variable. We will continue further in a new topic in the next class.

Thank you very much.