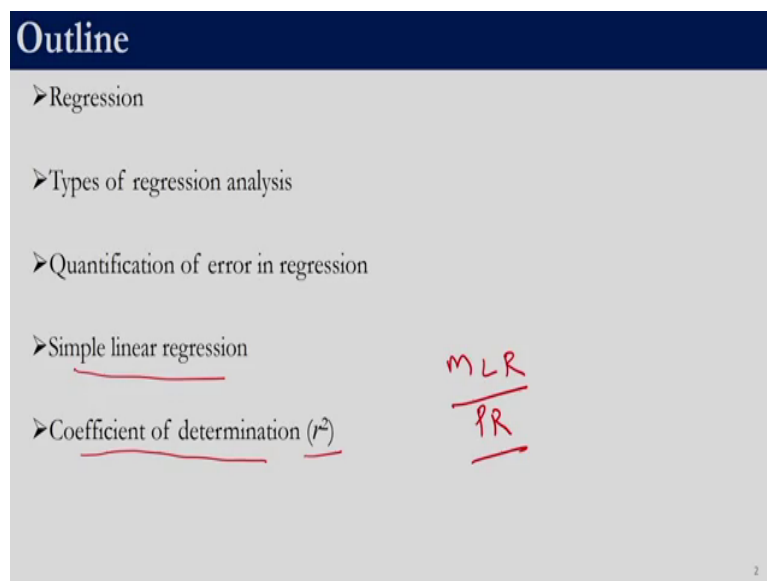


**Computer Aided Applied Single Objective Optimization**  
**Dr. Prakash Kotecha**  
**Department of Chemical Engineering**  
**Indian Institute of Technology, Guwahati**

**Lecture - 02**  
**Curve fitting: Regression**

You welcome back in this session we will be looking at Regression which is a type of Curve fitting technique. In regression we have we are given a set of data points and our objective is to fit a model, the model coefficients are not known. So, we are expected to determine the coefficients of the model, such that the model best represents the data points.

(Refer Slide Time: 00:53)



**Outline**

- Regression
- Types of regression analysis
- Quantification of error in regression
- Simple linear regression
- Coefficient of determination ( $r^2$ )

MLR  
—  
RR

2

The outline of today's session is going to be first we look into what is regression and then we will talk about the types of regression analysis. And, then we will discuss about how do we quantify error in regression, because that is what is going to be our objective right.

We need to have a measure as to how good our model is right. So, that is why we will be looking into Quantification of error for regression, then we will be discussing. In this session only about Simple linear regression, the other types of regression multiple linear regression and polynomial regression we will be looking in the next session. After that we in order to quantify how good our model is we will be looking into something that is popularly known as r square, which is at which is the coefficient of determination right. So, that is going to be the outline of today's session.

(Refer Slide Time: 01:45)

## Regression

- Fits a selected function to the general trend of data.
- An underlying mathematical model is selected, based on physical situation.
- Coefficients of the model are determined such that the error between model values and the given data is minimum.
- Applied when substantial error is associated with the data

Let the approximating function be  $y = f(x) = a_0x + a_1$

Value of the  $k^{\text{th}}$  point can be obtained from  $f(x_k) = y_k + e_k$

Error/deviation/residuals can be determined as  $e_k = f(x_k) - y_k$

Average error:  $E_e = \frac{\sum_{k=1}^n (f(x_k) - y_k)}{n}$

x	0	2	4	5	8	9
y	2	3.2	3.8	4.6	6.2	6.8

In regression we are given the data and we are required to fit a model. So, the form of the model is known. So for example, here if we see  $x$  and  $y$  data are given that is plotted over here and here we are required to fit a straight line right. So, for example, let us assume that we are required to fit this line  $f$  of  $x$  is equal to  $a$   $x$  plus  $b$  right. So, as you can see it is a linear equation right. So, in this points if we can see there are many lines which we can fit right.

So, there are infinite combinations of lines which can be drawn to these points right. So, our task is to find out the line, our task is to find the coefficients of the model such that the error between model values and the given data is minimum. So, let us say we come up with our particular value of  $a$  and  $b$  right. So, once we have a particular value of  $a$  and  $b$  for all these  $x$  values we can actually calculate  $y$  let us say that is  $y$  model. So, the difference between this  $y$  and  $y$  model has to be minimum, so that is what we will be doing in regression.

So, the task is to come up with this  $a$  and  $b$  such that it best fits the data right. So, for a given point  $x_k$  the value from the model can be determined using this expression. So, for a given point  $x_k$  the value of  $y$  can be determined from this right. So, that would be  $y$  that is  $f$  of  $x_k$  and  $y_k$  is what is actually measured, so  $y_k$  is what is given. So, the difference between  $f$  of  $x_k$  and  $y_k$  is the error. So, the difference between what our model would predict the value to be and the value which has been observed. So, the difference between those two is the error  $e_k$ .

So, here as you know  $x$  is the independent variable  $y$  is the dependent variable. So, as we change  $x$  we have a system which provides us the  $y$  value. So,  $x$  is the independent variable and  $y$  is the dependent variable. So, this error is also known as deviation or residuals and the average error can be calculated by this is the error associated with one data point. So, data associated with the error associated with  $n$  points is the summation of all the individual errors and if we divide that by the total number of points.

So, in this case n would be 6 because there are 6 data points right, so that will give us average error. We can also calculate maximum error, we can also calculate root mean square error.

(Refer Slide Time: 04:23)

### Types of regression analysis

➤ Linear regression: linear model is used to fit the data

- Simple: linear model with one independent variable.  $y = a_0 + a_1x$
- Multiple: linear model with two or more independent variables.  
 Let number of independent variables be two, then  $y = a_0 + a_1x_1 + a_2x_2$   
 For  $m$  independent variables:  $y = a_0 + a_1x_1 + a_2x_2 + \dots + a_mx_m$
- Polynomial: model to fit the data is a higher order polynomial  
 Let the order of the polynomial be two, then  $y = a_0 + a_1x + a_2x^2$   
 For an  $m^{\text{th}}$  order polynomial:  $y = a_0 + a_1x + a_2x^2 + \dots + a_mx^m$

Handwritten notes and diagrams include:  
 - A matrix representation for linear regression: 
$$\begin{array}{c|c} n_1 & n_2 & y_1 \\ \hline n_{11} & n_{21} & y_1 \\ n_{12} & n_{22} & y_2 \\ n_{1s} & n_{2s} & y_s \end{array}$$
  
 - A boxed list of coefficients:  $a_0, a_1, a_2, \dots, a_m$   
 - A matrix representation for polynomial regression: 
$$\begin{array}{c|c} n & y \\ \hline n_1 & y_1 \\ n_2 & y_2 \\ n_3 & y_3 \\ \vdots & \vdots \\ n_n & y_n \end{array}$$

The different types of regression are simple in which we, in which there is one independent variable and one dependent variable. So, in all these cases will have only one dependent variable and we can have more than one independent variable. When we have only one independent variable, so for example here y is the dependent variable and x is the independent variable right. So, this is a, this is what we will call it a simple regression right. So, the task here is to find out the values of a naught and a 1, so that it best fits the data.

Now, when we say it best fits the data we will have to define what is best right, so that we will do in subsequent in subsequent discussion. In multi linear regression we have two or more independent variables, remember the dependent variable is still y. So for example, here

if we see  $y$  is equal to a naught plus  $a_1 x_1$  plus  $a_2 x_2$ , remember  $x_1$  and  $x_2$  are not the data for variable  $x$  right. So, there are two variables. So, the data is going to be something similar to  $x_1 x_2$  and then  $y$ .

So, for a particular value of  $x_1$  for a particular value of  $x_2$  we are going to have  $y$  right, so because we have so for example  $x_1 = 1, x_2 = 2$  and so we can have  $x_1$ . The first data point  $x_2$  the variable the first data point and  $y_1$ , similarly the variable  $x_1$  the second data point the variable  $x_2$ . The second data point and  $y_2$  variable  $x_1$  the third data point variable  $x_2$  the third data point and the  $y$  value. So, this is how our data is going to be right.

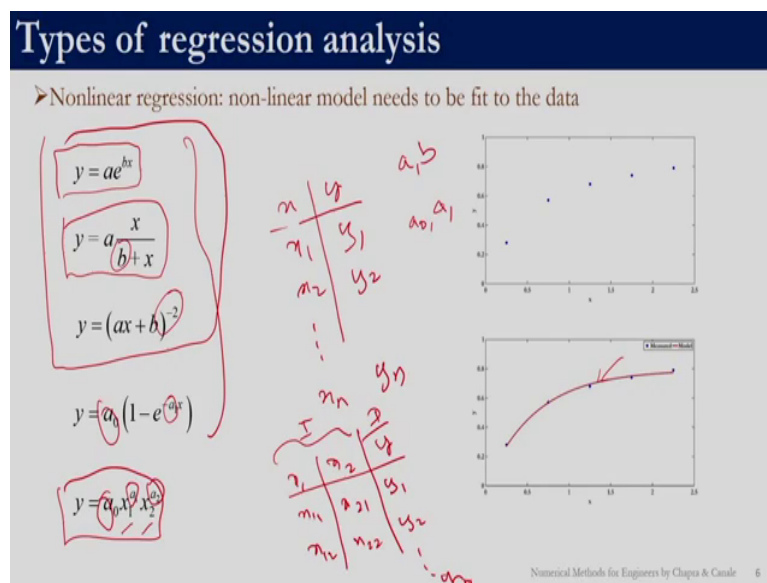
So, the task is to fit this model right is to fit this model, so that is to determine the values of a naught  $a_1$  and  $a_2$ . So,  $x_1$  and  $x_2$  are known right our task is to find out these three coefficients. So, this is if we have two independent variables. So, the independent variables are  $x_1$  and  $x_2$ , similarly we can have  $m$  independent variable. So, when we have an independent variable this is the constant coefficient a naught  $a_1 x_1 a_2 x_2$  and all the way up to a  $m \times m$  right. So, remember this  $y$  is nothing but the value predicted by the model, there is going to be there might be some error between the value predicted by the model and the value which we have observed right.

So, we need to come up with these coefficient such that the difference is minimum, so that is our multiple regression. In polynomial regression we have a polynomial that is to be fit. So for example, if we have a second order if we want to fit a second order polynomial then we have this data  $x$  versus  $y$  right. So,  $x_1 x_2 x_3$  these data points are known  $x_1 y_1 x_2 y_2$  and  $x_3 y_3$  and so on up to  $x_n y_n$ .

So, the task is to come up with the best values of a naught  $a_1 a_2$ , such that the difference between the value predicted by the model and the actual value which was observed is minimum for all the points right. So, for if you want to fit an  $n$ th of the polynomial this is what the model would look like a naught plus  $a_1 x_1$  plus  $a_2 x_2^2$  plus  $a_3 x_3^3$  all the way up to a power  $m$  all the way up to a  $m \times m$ .

Here, if we see all these three cases simple regression multiple regression and polynomial regression, the coefficients of a naught the power of a naught a 1 a 2 and all the way up to a m, whatever the coefficients which are unknown. So, in this case these are the coefficients which are unknown right. So, their power was one, so that is why this is linear regression right. So for example, here we have x square so that does not make it non-linear regression, because the data points x are known what is not known are the model coefficients.

(Refer Slide Time: 08:29)



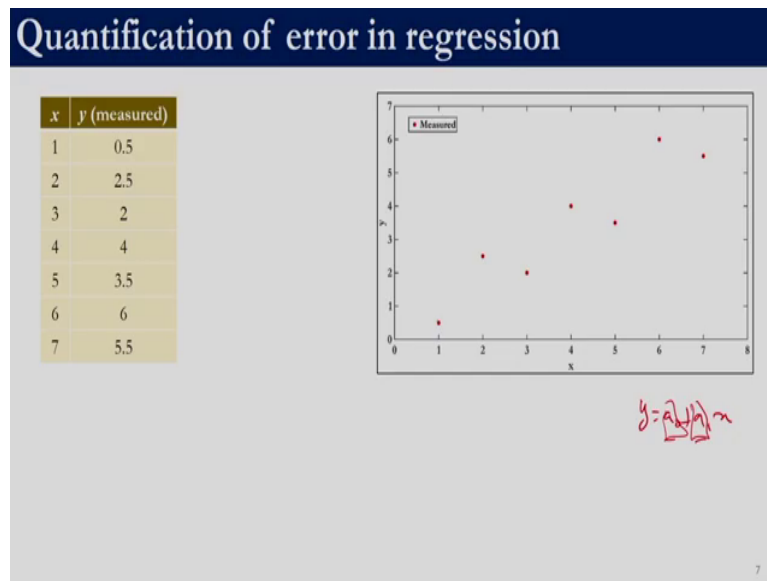
So, sometimes we are required to fit non-linear models. So for example, assume this is a data assume this data set right x versus y right. In this case we want to fit a model similar to this. So, this model is now if you see it is a non-linear model. So, some other some other few examples of non-linear model is y is equal to a e power b x because of b this model is non-linear right. So, over here also if you see b occurs in this denominator so it is a non-linear

again, because of this minus 2, it becomes non-linear again over here a 1 is not known and we need to find out. So, this is a non-linear model right. So, this is also a non-linear model.

So, all these three all these first four models have only one dependent variable right. So,  $x$  versus  $y$  is known  $x_1 y_1 x_2 y_2$  all the way up to  $x_n y_n$  all the data points are known. The task is to find out the coefficients  $a$  to  $b$  in this first three models, in the fourth model we are interested to find  $a$  naught and  $a_1$  right and in this for this particular model we have two independent variables. Just like in linear regression we had multiple independent variables, here also we can have that multiple independent variables. So,  $x_1$  is a variable  $x_2$  is a variable and  $y$  is the dependent variable.

So, these two are independent variables  $y$  is a dependent variable. So, we have these points  $x_1 y_1 x_2 y_2$  all the way up to  $n$  points we have  $n$  points. So, the task is to come up with the coefficients  $a$  naught  $a_1$  and  $a_2$  right. So, this is non-linear regression. So, first we look into linear regression right and then we will move on to non-linear regression. So, non-linear regression can be solved in two ways, in certain cases we can transform the model into a linear model. In certain cases the model is not transformable, so in that case we will have to handle it as non-linear model itself.

(Refer Slide Time: 10:45)



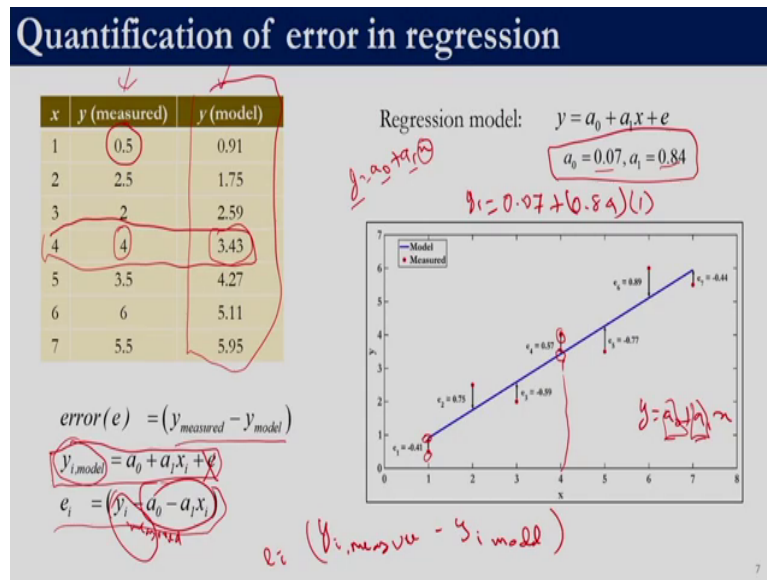
So, we will so we will discuss that as we come to non-linear regression. So, consider this data set x and y right. So, this data set is available the values of x are 1 2 3 4 5 6 7 and the measured values of y are 0.5 2.5 2 4 3.56 and 5.5 let us say. So, over here if we see there are multiple ways to fit this, there are multiple models you can fit multiple models in the sense multiple straight lines we can draw right.

So, here the task is to fit a straight line let us say y is equal to a naught plus a 1 x right. So, here if we see by changing a naught and a 1, we can draw as many lines as we want right. But not all of those lines are will best reflect the data points right. So, that question is what is the best value of a naught and a 1 right. So, before finding before going into how to find the best values of a naught and a 1, let us understand how are we going to define the measure right. When we say best we also need to define best with respect best in what sense right, you want



the error to be minimum or you want to be you want the square of the error to be minimum right.

(Refer Slide Time: 11:55)



So, that is what we will first look into it right. So, let us assume that some way we are able to come up with this model this a naught and a 1 remember the task was to find out a naught and a 1, right now I am saying let us say there was some way with which we found this a naught to be 0.07 and this a 1 to be 0.84 right. So now, since I have this model coefficients right. So, with these model coefficients we can find out the values of y right.

So for example, for 0.91 what would have been done is y 1 is equal to. So, y 1 from model 1 this y 1 is what we have measured right. So, but we are now trying to now we are trying to calculate what is the value of y predicted by the model, if the model coefficients are 0.07 and

0.84 right. So, in this case we will calculate  $y_1$  is equal to  $0.07$  plus a  $1$  is  $0.84$  into  $1$  because that is the  $x$  value right. So, we will get this  $0.91$ .

Similarly for  $y_4$  this  $4$  is observed value that is already given. So, to get this  $3.43$  we will have to plug the value of  $4$  in this expression in this equation  $y$  is equal to a naught plus a  $1$   $x$  right. So, a naught is  $0.07$  a  $1$  is  $0.84$ . So, we will have to calculate with  $x$  as  $4$  we will have to calculate the  $y$  value, so that would be  $3.43$ .

So, now we have the measured values and now we also have the values predicted by the model, so for each of this point we can calculate what is the error right. So, for each of this point what we can calculate what is the error. So, the error is shown over here. So, for the first point the model is providing this value right and the data point is this so the error is minus  $0.41$  right. So, the error over here is minus  $0.4$  here for  $1$  right. Similarly, for  $0.4$  the actual data point is this and what is predicted by model is over here right for  $x$  is equal to  $4$ , whatever that is predicted by the model. So, we can calculate the individual error right.

So now, individual error which is nothing but  $y$  measured minus  $y$  model and the model value itself is nothing but substituting the data value right. So, this  $e$  should not be here over here, the model value is what is given by the model and not the error. So, the error is not known right. So, when we are talking about measured it is  $y$  model plus error. So for example, if I want to write the relation between  $y$  measure and  $y$  model it is  $y_i$  what is measured minus  $y_i$  what is what we get by model right. So, that is the error associated with the  $i$ th point.

So, since we are talking about model over here then there is this  $e$  term should not be here, it is just a naught plus a  $1$   $x_i$  and the error is this one right where  $y_i$  is nothing but  $y_i$  measured. So the measured value minus the model value that will give us the error. So now we have the individual errors right and we want the error to be minimum right. So, one obvious thing is that we sum this errors right. So, if we sum this error what is going to happen is the positive and negative errors are going to get cancelled out right. So, we do not we cannot add the error as such.

(Refer Slide Time: 15:35)

### Linear regression ( $y = a_0 + a_1x$ )

➤ Sum of squares of residuals for  $n$  data points

$$Min S_r = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_{i,measured} - y_{i,model})^2$$

$$S_r = \sum_{i=1}^n (y_i - a_0 - a_1x_i)^2 \quad y_i = (a_0 + a_1x_i)$$

➤ Differentiating  $S_r$  equation with respect to each unknown coefficients of polynomial

$$\frac{\partial S_r}{\partial a_0} = \sum_{i=1}^n (y_i - a_0 - a_1x_i) = 0$$

$$-\sum_{i=1}^n y_i + \sum_{i=1}^n a_0 + \sum_{i=1}^n a_1x_i = 0$$

$$\sum_{i=1}^n a_0 + \sum_{i=1}^n a_1x_i = \sum_{i=1}^n y_i$$

$$na_0 + a_1 \sum_{i=1}^n x_i = \sum_{i=1}^n y_i$$

$$\frac{\partial S_r}{\partial a_1} = -2 \sum_{i=1}^n x_i (y_i - a_0 - a_1x_i) = 0$$

$$-\sum_{i=1}^n x_i y_i + \sum_{i=1}^n x_i a_0 + \sum_{i=1}^n a_1 x_i^2 = 0$$

$$a_0 \sum_{i=1}^n x_i + a_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i$$

$f(n)$

$$J = \begin{bmatrix} \frac{\partial}{\partial a_0} \\ \frac{\partial}{\partial a_1} \end{bmatrix} = 0$$

$(5x-3)^2$

$2(5x-3) \cdot 5$

$x^2$

$2x \cdot 2x$

So, what we will do is we will add the square of the errors right. So, we will say this is our measure right, when we say a line is a best fit it is going to be with respect to this measure, that we want the error the square of error. This is the square of the error associated with each point right, this across all the points summed across all the points. So, this is what is called as sum of square of errors right or sum of square of residuals, because error is also known as residuals right.

So, this is going to be indicated by  $S_r$  right. So, this we want to minimize right, we want to minimize  $S_r$  and  $S_r$  is nothing but Sum of square of all the errors right. So, if you substitute the expression for if we substitute the expression measured minus model for error will have this expression right and then to make it compact if we get rid of this measure. So, this is actually measured, but we are not writing it to keep it compact right. So,  $y_i$  minus it is  $y_i$

minus a naught plus a 1 xi right. So, that is why you have the two minus over here when you expand this you will get that right. So, this is what we have.

So, this is what we have now going back to the basics which we have studied. So, if you have a function  $f$  of  $x$  right, we know that the minima will occur at the stationary point right. So, the stationary point can be determined by equating the gradient to be 0 right. So, once we get the stationary point we will have to evaluate the second derivative and see if it is greater than 0 or less than 0 and then we will call it will establish whether the stationary point corresponds to a minima or a maximum right. So, that is for a single variable problem.

For a multi variable problem we will have to equate the Jacobian to be 0 right, as we discussed in the previous session. So, if we have two variables  $x_1$  and  $x_2$  then  $\text{d}f/\text{d}x_1$  and  $\text{d}f/\text{d}x_2$  has to be equated to 0. So, the Jacobian has to be equated to 0 that will give us the stationary point right. In this case remember the points  $x_1$  are known right. So, what is unknown as a naught and a 1. So, when we say we want to determine the stationary points, we need to differentiate  $S_r$  right with respect to a naught and a 1. So, that is what we will do now right.

So, we will differentiate  $S_r$  with respect to each unknown coefficients in the model right. So, here we have linear model. So, we need to differentiate  $\text{d}S_r$  by we need to differentiate  $S_r$  and find out what is  $\text{d}S_r/\text{d}a_{naught}$  and we need to find out what is  $\text{d}S_r/\text{d}a_1$  right. So, just like we have an expression, if you have an expression if you want to differentiate  $(5x - 3)^2$ , then we do 2 into  $(5x - 3)$  into 5 right. So, our  $x^2$  differentiation of  $x^2$  is  $2x \text{ d}x$ . So, same thing we will use over here so initially we will call this as  $x$  right so  $x^2$ .

So, what we will have it we will have it is this  $1/2$  times  $x$  right and then we need to differentiate  $y_i$  with respect to a naught. So, that is 0 we need to differentiate this a naught with respect to a naught, so that will give us minus 1. So, that is why we have this minus sign over here and then when we differentiate  $a_1 x_i$  with respect to a naught again we will get 0

right. So, this is the differentiation of this expression  $S_r$  with respect to  $a_0$  that has to be equated to 0.

So, here if we expand this equation right, so this 2 can be removed because the right hand side is 0 right. So, if we expand this equation we have  $-\sum y_i$  plus  $\sum a_0$  plus  $\sum a_1 x_i$ , because remember we have a minus sign over here. So, minus into minus that is why this is plus, this minus into this minus we have a plus over here right so equal to 0. So now, if we see for when we are working with this problem  $x$  and  $y$  are given to us right.

So, since  $y$  is given to us we can actually calculate what is  $\sum y$  right. So, all this  $\sum$  we have removed this index just to keep it compact, but all the  $\sum$  are going from 1 to  $n$  right. So,  $\sum y_i$  can be determined, so that is why this term is being taken to the right hand side over here and the  $\sum a_0$  is nothing but  $n$  times  $a_0$  right. So, this  $a_0$  out of this  $a_1 x_i$  we can take this  $a_1$  outside right. So, this is the expression we will have after differentiating  $S_r$  with respect to  $a_0$  right.

So, now if we see in this in this case  $n$  is known  $n$  is the number of data points right  $a_0$  is not known  $a_1$  is not known,  $\sum x_i$  can again be calculated it is nothing but the summation of these values and  $\sum y_i$  can be calculated. So, there are two unknowns in this right  $a_0$  and  $a_1$ . Similarly, if we differentiate over here right; so, two times the entire expression and if we differentiate  $y_i$  with respect to  $a_1$  we will get 0,  $a_0$  with respect to  $a_1$  we will get 0. When we differentiate this term  $a_1$  into  $x_i$  with respect to  $a_1$  we will get  $x_i$  right.

So, that is why this  $S_i$  appears over here and now if we expand this, then this  $x_i$  will be associated with each term. Again over here if we see  $x_i$  is known  $y_i$  is known. So, this point is known this point is known, so we can calculate  $x y$  right. So, this is multiplication of these two. So, all these terms are known. So, this is going to be completely known.

So, we will take it to the right hand side and the negative sign would disappear. So, we will have  $\sum x_i y_i$ , similarly over here if we see  $a_0$  is a constant which we do not know  $a_0$  is the model coefficient which we do not know, so  $a_0 x_i$  plus  $a_1 x_i^2$ . So,

we know x will calculate x square for will square each point and then sum it up. So, this term is also known this is also known right.

So, now if we see this is also expression which involves the unknowns are a naught and a 1 and both of these equations are linear. So, here we had a non-linear optimization problem right. So, over here if we see we have this square term, we had a non-linear optimization problem. We applied the stationary condition and the stationary condition in this case happens to be two unknowns in two equation and both the equations are linear right. So, it has a unique solution, so we will get only one stationary point.

(Refer Slide Time: 22:39)

**Linear regression ( $y = a_0 + a_1x$ )**

➤ Sum of squares of residuals for  $n$  data points

$$Min S_r = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_{i,measured} - y_{i,model})^2$$

$$S_r = \sum_{i=1}^n (y_i - a_0 - a_1x_i)^2$$

$y = a_0 + a_1x$

➤ Differentiating  $S_r$  equation with respect to each unknown coefficients of polynomial

$$\frac{\partial S_r}{\partial a_0} = -2 \sum_{i=1}^n (y_i - a_0 - a_1x_i) = 0$$

$$\frac{\partial S_r}{\partial a_1} = -2 \sum_{i=1}^n x_i (y_i - a_0 - a_1x_i) = 0$$

$$\sum y_i + \sum x_i a_0 + \sum a_1 x_i^2 = 0$$

$$a_0 \sum x_i + a_1 \sum x_i^2 = \sum x_i y_i$$

$$n a_0 + a_1 \sum x_i = \sum y_i$$

$$a_0 = \bar{y} - a_1 \bar{x}$$

$\bar{x}, \bar{y}$ : means of  $x, y$

➤ Simultaneous linear equations:

$$\begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

NORMAL EQ

So, now we have this we can arrange it in. So now, we can put these two equation in the standard format right. So, the two equations can be put in standard format  $n a_0 + a_1 \sum x_i = \sum y_i$  the second equation is  $a_0 \sum x_i + a_1 \sum x_i^2 = \sum x_i y_i$

is equal to  $\sum x_i y_i$  right. So, these are two equations in two unknowns this can be put in the conventional form right. So, wherein we say the unknowns are  $a$  and  $b$ , so we write it as a vector and if this is a vector the constants related to these two equations are  $n$  plus  $\sum x_i$ .

So, if I expand this first equation if I expand this it will be  $n$  into  $a$  plus  $\sum x_i$  into  $b$  is equal to  $\sum y_i$  which is nothing but our first equation. The second equation is  $a$  plus  $\sum x_i^2$  into  $b$  is equal to  $\sum x_i y_i$ . So, that is our second equation. So, this is nothing but in the conventional form this is the coefficient matrix this is the  $x$  vector and this is the right hand side values of the linear expression right.

So, this can be completely calculated this can be completely calculated given the data points right. So, by solving these two equations in two unknowns we can get the values of  $A$  and  $b$  or this can be rearranged right. So, for example I can expand this right. So, the right hand side for example, I can write I can expand this and I can get this expression right. So, if you are interested you can either use this or you can use this one, so these are known as normal equations. So, these two equations are known as normal equations. So now, we know how to solve a linear regression problem. So, let us take a problem so here we have 9 data points these are the  $x$  values and the  $y$  values right.

(Refer Slide Time: 24:23)

**Example: Linear regression ( $y = a_0 + a_1x$ )**

	x	y
1	1	4
2	3	5
3	5	6
4	7	5
5	10	8
6	12	7
7	13	6
8	16	9
9	18	12
$\Sigma$	85	

*Handwritten notes:*  $\sum_{i=1}^n x_i^2$  and  $\sum_{i=1}^n x_i y_i$

$$\begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

So, now let us look at an example right, so here we have 9 data points. So, this is just serial number, so 1 2 9 these are the x values and the y values. So now, our task is to come up with this  $a_0$  and  $a_1$  to determine this value of  $a_0$  and  $a_1$ , such that sum of square of errors right for of all the points, sum of square of errors is minimum right. So, there is no other value of  $a_0$  and  $a_1$  for which the sum of square will be minimum right. So, this we have seen this problem is nothing but solving a set of simultaneous linear equations involving two unknowns  $a_0$  and  $a_1$ .

So, this is what we have derived from in the previous slide right. So,  $n$  is the number of data points. So, in this case  $n$  happens to be 9  $\sum x_i$  is nothing but the summation of all this values right. In this case it happens to be 85 right and then we require  $\sum y_i$ , so  $\sum y_i$  is the summation of this  $y$  values.



(Refer Slide Time: 25:41)

**Example: Linear regression ( $y = a_0 + a_1x$ )**

	$x$	$y$	$x^2$	$xy$
1	1	4	1	4
2	3	5	9	15
3	5	6	25	30
4	7	5	49	35
5	10	8	100	80
6	12	7	144	84
7	13	6	169	78
8	16	9	256	144
9	18	12	324	216
$\Sigma$	85	62	1077	686

$$\begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

$$\begin{bmatrix} 9 & 85 \\ 85 & 1077 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} 62 \\ 686 \end{bmatrix}$$

$$\begin{bmatrix} a_0 = 3.43 \\ a_1 = 0.37 \end{bmatrix}$$

$$y = 3.43 + 0.37x$$

*Handwritten notes:  $\frac{dy}{dx} = \frac{a_1}{1} = 0.37$*

And then we will require xi square. So, remember very often students make this mistake that it is not square of 85; it is not 85 square it is just that each value has to be squared. So, 1 square 1 square 3 square 5 square and so on right. So, this each element has to be squared and then it is it is sum has to be taken right. So, it is 1077 and not 85 square right, so that is that will be required over there. Similarly xi yi it is not 85 into 62 right, it is every element of x has to be multiplied with every element of y right. So for example: 1 into 4 3 into 5 6 into 5 into 6 7 into 5 10 into 8 12 into 7 right.

So, this is xi x y and the summation is 686 right, so we can plug if you plug all those values. So now we have two equations in two unknowns. So, you know how to solve this two equations in two unknowns right. So, if we calculate the value of a naught is 3.43 and 0.37 right. So, the model which we have is y is equal 3.43 plus 0.37 x right. So, the advantage of

having this model in this form is now for example, if you ask me what is the value of  $y$  at  $x$  is equal to 8 at which we do not have the value of  $y$  available. So, this model can be used.

So, if we plug in the value of  $x$  as 8 we can estimate we can predict what would be the value of  $y$ . So, that is one benefit of having this model and suppose let us say you have this data and you want to find out how does  $y$  vary with respect to  $x$   $\frac{dy}{dx}$  or  $\frac{d^2y}{dx^2}$ . So now we have this model in compact form, now we have this now since we have this model in the regular form we can actually calculate what is  $\frac{dy}{dx}$  and  $\frac{d^2y}{dx^2}$  right.

So, this is how we fit a simple straight line to a given set of data points. So now we so now the question is how will this model has captured other data points right. So, we have this model one way is to plot it and see right. So, the question is can it be quantified as to how good it is right. So, for that we have something called as coefficient of determination or commonly or commonly known as  $r^2$  right.

(Refer Slide Time: 28:13)

## Coefficient of determination ( $r^2$ )

- Quantifies the 'goodness' of a fit.  $y = ax + b$   $y_m = \bar{y}$
- $S_t$ : Magnitude of the residual error associated with the dependent variable with respect to the mean.  $S_t = \sum_{i=1}^n (y_i - \bar{y})^2$ ,  $\bar{y} = \text{mean}(y)$
- $S_r$ : The sum of the squares of the residuals around the regression line.  $S_r = \sum_{i=1}^n (y_i - y_{i(\text{model})})^2$
- $(S_t - S_r)$ : Quantifies the improvement or error reduction due to describing data in terms of a straight line rather than as an average value.
- Coefficient of determination:

$$r^2 = \frac{S_t - S_r}{S_t}$$

Applied Numerical Methods with MATLAB for Engineers and Scientists by Chapra 10

So, our r square can be said to quantify the goodness of a fit right. So, to calculate r square we require what is known as the total error right  $S_t$  denoted by  $S_t$ . So,  $S_t$  is given by instead of mode, so let us say we were trying to fit this model  $y$  is equal to  $a x$  plus  $b$ . Instead of this let me write just  $y_m$  right, instead of this if we say that my model is not  $a x$  plus  $b$ , but merely the mean of the  $y$  values, the average of the dependent variables value right.

So, in that case what would be the error associated with each point is  $y_i$  minus  $y_i$  minus  $y$  bar right. So, this is  $y_i$  bar because there is only one mean for a given vector so  $y_i$  minus  $y$  bar the whole square just. So, that you do not want to cancel out the positive and negative errors and the summation across all the data points. So, this is called as  $S_t$ . So, instead of regression right instead of regression, if we were to assume the model is nothing but the mean value of  $y$  what would be the total error. So, that is indicated by  $S_t$  right.

And then we know this  $S_r$  right in this case  $y_i$  minus  $y_i$  model right. So, model is  $ax_i$  plus  $b$  right this is what we minimize right. So, once we have this  $a$  and  $b$  we can actually calculate what is  $S_r$  right. So,  $r^2$  is given by  $S_t - S_r$  by  $S_t$  right. So,  $S_t - S_r$  quantifies the improvement of error due to describing data in terms of a straight line, rather than as an average value right.

So, this is the error that we get if we assume the model to be  $\bar{y}$  and this is the error which we got which we get after fitting a straight line right. So,  $S_t - S_r$  quantifies the improvement of improvement or error right. So,  $S_t - S_r$  divided by  $S_t$  this coefficient of determination  $r^2$  can be calculated right.

(Refer Slide Time: 30:21)

**Coefficient of determination ( $r^2$ )**

- Case 1: Perfect fit ( $r^2 = 1$ )
  - The line explains 100 percent of the variability of the data.
- Case 2: No improvement ( $r^2 = 0$ )
  - No reduction of error by describing the data in a straight line.
- $r^2$  close to 1 does not mean the fit is necessarily good.

*Handwritten notes on slide:*  
 $r^2 = \frac{S_t - S_r}{S_t}$   
 $= \frac{S_t - 0}{S_t} = 1$

Applied Numerical Methods with MATLAB for Engineers and Scientists by Chapra 11

So,  $r^2$  is  $S_t - S_r$  by  $S_t$  let us say  $S_r$  is 0 right. So, if we say  $S_r$  is 0 what does that mean right. So, if  $S_r$  is 0  $S_t - 0$  by  $S_t$  so  $r^2$  is equal 1, so that is the case

which we are discussing so that is called as perfect fit. Where in the line the passes through all the data points right, because there is no the model passes through all the data points right.

So, there is no error associated with any of the data points only then  $S_r$  would be 0 right. So,  $S_t$  minus 0 by  $S_t$  is equal to 1 right and if  $S_r$  happens to be equal to  $S_t$  right in that case  $r$  square will be equal to 0 right. So that means, that the entire effort to fit a straight line was useless, we could have just taken mean of the dependent variable. So, we could have just taken the mean of the dependent variable. A common misconception is that if  $r$  square is close to 1 it is a very good fit right that may or may not be the case right, with  $r$  square equal to 1. It is definitely a very good fit, because the line will pass through all the data points right.

(Refer Slide Time: 31:47)

### Coefficient of determination ( $r^2$ )

➤ Anscombe's (1973) four data sets ( $n = 11$ )

Dataset I		Dataset II		Dataset III		Dataset IV	
$x_i$	$y_i$	$x_i$	$y_i$	$x_i$	$y_i$	$x_i$	$y_i$
10.00	8.04	10.00	9.14	10.00	7.46	8.00	6.58
8.00	6.95	8.00	8.14	8.00	6.77	8.00	5.76
13.00	7.58	13.00	8.74	13.00	12.74	8.00	7.71
9.00	8.81	9.00	8.77	9.00	7.11	8.00	8.84
11.00	8.33	11.00	9.26	11.00	7.81	8.00	8.47
14.00	9.96	14.00	8.10	14.00	8.84	8.00	7.04
6.00	7.24	6.00	6.13	6.00	6.08	8.00	5.25
4.00	4.26	4.00	3.10	4.00	5.39	19.00	12.50
12.00	10.84	12.00	9.13	12.00	8.15	8.00	5.56
7.00	4.82	7.00	7.26	7.00	6.42	8.00	7.91
5.00	5.68	5.00	4.74	5.00	5.73	8.00	6.89

Property	Value
Mean of $x$	9
Mean of $y$	7.5
Linear regression line	$\hat{y} = 3.00 + 0.500x$
Coefficient of determination of the linear regression	0.67

$$\begin{bmatrix} n & \sum x_i & \sum y_i \\ \sum x_i & \sum x_i^2 & \sum x_i y_i \\ \sum y_i & \sum x_i y_i & \sum y_i^2 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

Applied Numerical Methods with MATLAB for Engineers and Scientists by Chapra 12

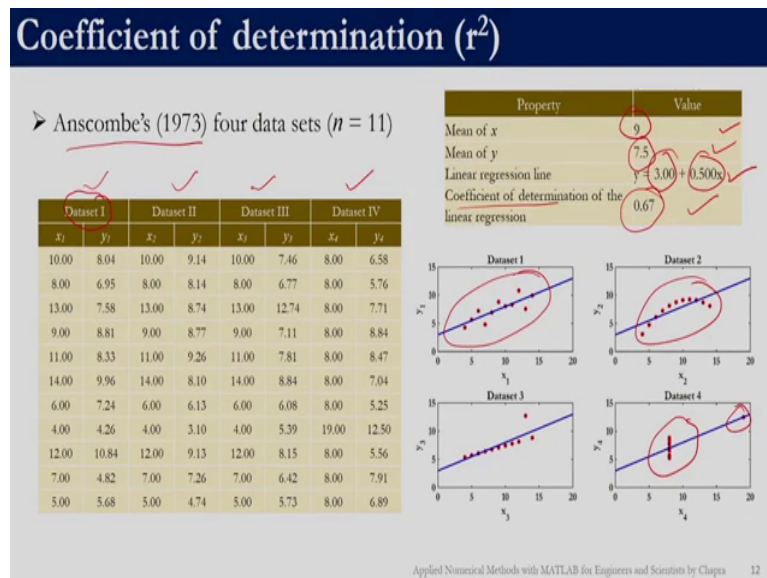
So, we will see an data set to understand this right. So, this is called as Anscombs data set right, it has four data set data set 1 data set 2 data set3 and data set 4 right. So now, let us say

we have this let us let us for a minute let us forget all the three other data, set let us just focus on data set 1. So, we have the x values we have the y values right. So, we can actually do a regression linear regression right. So, we can we can find out that  $n \sum x \sum x \sum x^2$  a naught a 1 and  $\sum y \sum y \sum xy$  right. For these the first data set and we can calculate a naught and a 1.

So, you can try it if you calculate a naught and a 1 it will be 2 plus 0.5. So, that is the line for this and once we have this we can also calculate the r square value right. So, r square value for this would turn out to be 0.67 right and the mean of y is 7.5 and the mean of x is 9 right. So, let us say we have done regression for this the first data set right. So, similarly we can do the regression for the second data set the third data set and the fourth data set.

So, for all these four data set you will see that the model is same, the mean of x and y is same and the coefficient of determination is same right. So, what we expect is that if I were to plot this, let us say if I were to plot x y and the model right. The straight line because the r square is equal to 0.6 what you might expect is that it fits more or less equally in all the four cases right.

(Refer Slide Time: 33:39)



But, if you see you over here right here the fit is kind of you can say it is good right. In this case definitely a straight line is not something that is capturing it well. Look at this case because of this just one point instead of getting a vertical line we have got this line with a slope right and here also the fit is not really good right.

So, but in all the four cases the regression coefficient is 0.67 right. So, when we are interpreting  $r$  square we need to be extremely careful right, it is always suggested that we actually plot the data and the model and have a look at it rather than just relying on  $r$  square.

(Refer Slide Time: 34:23)

### Linear regression: Coefficient of determination

x	y	y <sub>model</sub>	(y - y <sub>mean</sub> ) <sup>2</sup>	(y - y <sub>model</sub> ) <sup>2</sup>
1	4	3.8	8.35	0.04
3	5	4.54	3.57	0.21
5	6	5.28	0.79	0.52
7	5	6.02	3.57	1.04
10	8	7.13	1.23	0.76
12	7	7.87	0.01	0.76
13	6	8.24	0.79	5.02
16	9	9.35	4.45	0.12
18	12	10.09	26.11	3.65

$\bar{y} = 6.89$

$S_y = 48.87$

$S_r = 12.12$

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

$$S_r = \sum_{i=1}^n (y_i - y_{model})^2$$

$$r^2 = \left[ \frac{S_r - S_t}{S_t} \right]$$

Handwritten notes:  $r^2 = 0.27$ ,  $y = 2.07$

So, let us look into how to exactly calculate the r square right. So, this is the same data set for which we had fit the model right. So, this were the x this was the data set x and y and since we have already the model, we could have we can calculate what is the value predicted by the model right.

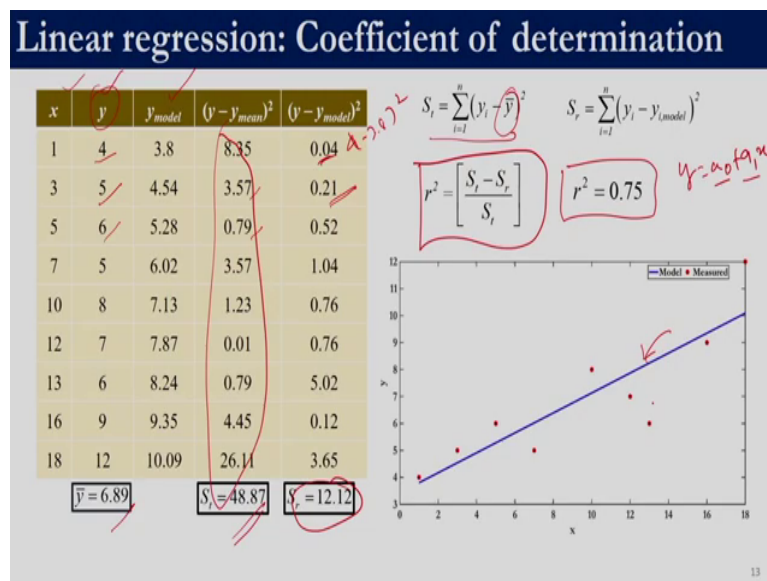
So, now we have this right. So, S t is to calculate S t we require the mean of the dependent variable right. So, the mean happens to be 6.89 right and then we can calculate y minus y mean the whole square. So, y minus y mean the whole square this 8.35 is nothing but 4 minus 6.89 the whole square 5 minus 6.89 the whole square, 6 minus 6.89 the whole square. So, this is y minus y mean the whole square right, those are the summation of this is nothing but our S t value which in this case turns out to be 48.87.



We also have the model now right, so 4 minus 3.8 the w5 minus 4.545 minus 4.5 the whole square, so that will be 0.21. So similarly we can calculate the difference between the error between the observed value or the measured value y and what is predicted by the model. So, model we have previously determined, so the coefficients worth something 3.43 and 0.37 right.

So, this was a naught and this was in a 1 and the model was y is equal to a naught plus a 1 x right. So, this is the S r values, now if we plug in these values into this expression of r square S t minus S r by S t. We can determine the r square value right.

(Refer Slide Time: 36:11)



So, r square happens to be point seven five and the data points we can directly plot right you can just do in MATLAB plot of x comma y, you will get the data points. Since the data points are already available you can plot the data points right and this is since you know the model y

is equal to  $a_0 + a_1 x$  and we have already found out what is the value of  $a_0$  and  $a_1$ . So, this line can be drawn right.

So, this is what we have obtained by regression and these are the data points right. So, here if we can see even the  $r^2$  is 0.75, there is significant error associated with many of the data points. So, that concludes simple linear regression right. So, basically you need to in simple linear regression, what we do is we define what is best fit right. So, in this case we said the best fit is the values of the coefficient  $a_0$  and  $a_1$  of the straight line  $a_0 + a_1 x$  for which the sum of square of error is minimum.

And, to find out that  $a_0$  and  $a_1$  we applied the stationary condition, once we have a not any one we can quantify the fitness using the coefficient of determination right that is  $r^2$ . So, now we know how to determine the model coefficients how to calculate the  $r^2$  value and remember the Anscombs data set that it is not usually safe to merely rely on the  $r^2$  value. At least in the case of linear regression we can plot and see how well the model actually fits the data points.

So, this is a classical application of optimization.