



# IIT Madras

## ONLINE DEGREE

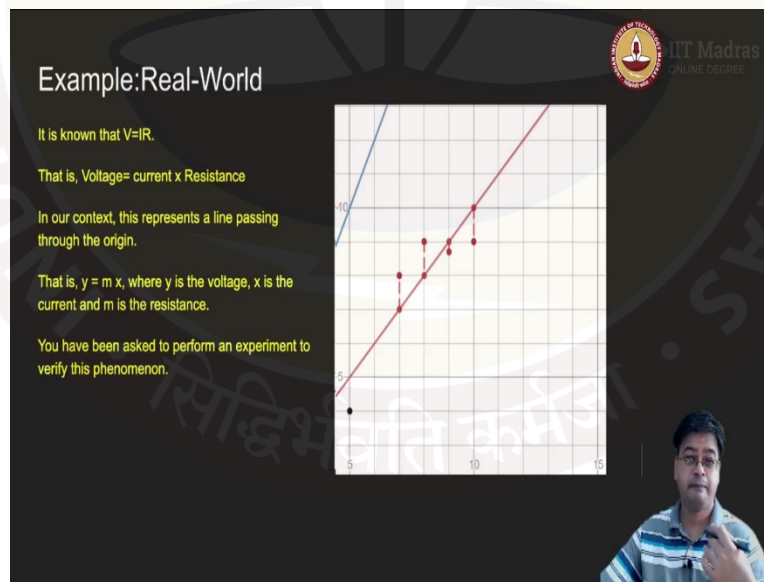
**Mathematics for Data Science 1**  
**Prof. Neelesh S Upadhye**  
**Department of Mathematics**  
**Indian Institute of Technology Madras – Chennai**

**Lecture-24**  
**Straight Line Fit**

Welcome friends welcome back so far what we have seen is a distance of a line from a point distance between two parallel lines. But the question now we can ask is, is that the only distance that we can seek as a distance of a point from the line. To demonstrate this let me give you one example where the paradigm will change as we will compare several points set of points and we will compare the distance from those set of points to the line and the paradigm change that I want to say is you will think differently how the distance will change from a line.

So, let us take one simple example this example is related to a small experiment that you might have conducted in your lab.

**(Refer Slide Time: 01:16)**



It is a physics experiment which says  $V=IR$  that is voltage is equal to current times the resistance. Voltage is equal to current times the resistance you all know this is a law this is the law of physics where voltage is measured in volts current is measured in amperes and resistance

in ohms. Now the experiment that a physics teacher asked you to conduct is you have to verify this law or using this law can you compute the resistance of a particular equipment.

So now what you will do is you will actually relate this with our equation of a straight line. So, if I want to relate this with the equation of a straight line then what will happen you see  $V$  is voltage so on the right hand side you can replace this voltage by  $y$  then the current that is delivered to the circuit or the equipment you can denote it by say  $x$  and you want to determine the resistance which is an unknown so you can put it as  $m$ .

And what is the constant? The constant is 0, so you can relate this with the equation  $y=mx$ , where  $y$  is the voltage,  $x$  is the current and  $m$  is the resistance and the whole purpose is to determine this resistance over here  $m$ . So, the setup is ready the lab technician has arranged a set up and you just have to go and perform the experiment and verify this phenomenon. So, the catch over here is you want to determine what is a resistance.

So, the lab technician was very kind he has given you a priori information that there are only two kinds of resistors our lab has one has a resistance of 1 ohm another one has a resistance of 2 ohms. This is the information that is given to you. Also notice the fact that this line is passing through the origin that means  $(0, 0)$  is one point why  $(0, 0)$  should be a one point because there is no current then there is no voltage this is our assumption.

So  $(0, 0)$  is one point and this line is passing through the origin so if I look at a mathematical theory that I have studied so far I can safely assume if I get one reading if I get one reading from that circuit that will help me in understanding the behavior and I can safely go and tell what is the resistance of this particular equipment. Let us try to see how this assumption works out over here.

Now this is the data you have conducted some experiments you have observed some data so it is like you have passed a current of 1 ampere and you received the output of 2 volts here you can say you have passed the current of 5 amperes and you have received output which is 4 volts and

so on and so forth. So, this is how it is working on. Now we want to identify what is the correct line that will fit because I know from theory that this is a line passing through the origin.

So in particular if I tell you this line which is (1, 2) and (0, 0) then I will get the equation of line using a slope point form or point - point form we also know that the intercept is (0, 0) so slope intercept form  $y=mx+c$ , where  $c$  is 0 you can easily see the line that passes through this point is  $y=2x$ . But with the same register you also got these readings. So, let us see based on the lab technicians' knowledge if we draw two lines, they will be seen they will be visible like this, interesting.

So, if I take only one observation and stop my experiment, I will get the line  $y=2x$ . But if I go for more experimentation then I am getting a line which seems to be similar to  $y=x$ . Now what is it that is happening here, which line is a better fit. So, I need to answer this question because this line actually passes through the point (0, 0) and  $y=2x$ , this line is not passing through any of the points.

So which line is better that is a natural question that comes to our mind? So, we will try to answer this question mathematically. So, how will I answer this question mathematically? Let us zoom in and consider our notion of a perpendicular distance. What is a perpendicular distance? You will actually drop a perpendicular from this point to this point and you will compute the distance of a line. Is that distance a correct distance? Geometrically it is a correct distance that is a distance of a line.

But in this context that we are taking real-world context what is happening here is if I pass a current of let us say 7 amperes this particular line is saying I should get a voltage of 7 volts but actually I got a voltage of 8 volts. So, now I may not be interested if I drop a perpendicular from this point to this point because this line is  $y=x$  it may cross this line at point 7.5. I am not interested what is the value of  $y$  at point 7.5.

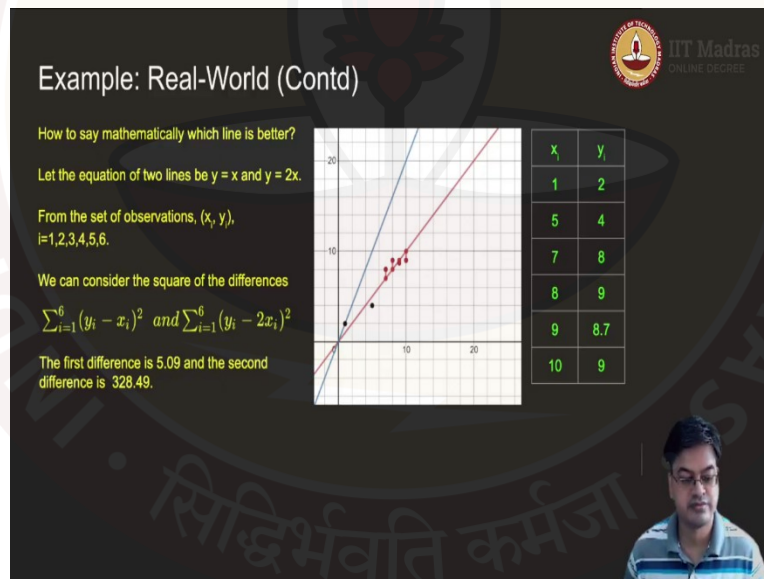
I should be interested in what is the value of  $y$  at point 7 because I have passed the current of 7 amperes not 7.5 amperes. So, the perspective of distance changes here because I want to find the

distance for this particular value of  $x$  from the line and the point how to go about then we will not consider a perpendicular distance. This is a paradigm shift that I was talking about at the beginning of the video.

So now I will not consider this thing but I will consider this distance that is a vertical distance the distance that is parallel to  $y$  axis that is what I will consider. So, once I consider the distance that is parallel to  $y$  axis, I have to consider these distances. So, again coming back to the question which line is the best-fit line I can consider similar distances over here. And I can consider similar distances over the blue line.

So which line is the best fit? We will try to answer this question mathematically. So, mathematically we have seen that perpendicular distance will not fetch me any result directly. So, I need to consider the distances that are parallel to  $y$  axis.

**(Refer Slide Time: 08:57)**



So, let us formalize this in a real term, this is the data that was shown in the picture. So, for 1 ampere you have got 2 volts current. For 5 ampere you got 4 volts current, for 7 you got 8, 8 you got 9, 9 you got 8.7 and 10 you got 9. So, there is no direct relation between  $y$  and  $x$ ; you cannot figure out the  $y = x$  is visible over here but something is there which is making that line pass very close to all these points.

This is the this is the demonstration, so  $y=2x$  is way apart and we are assuming that the hypothesis given by the lab technician is correct. So, I want to mathematically formulate this problem. There are two lines  $y=x$  and  $y=2x$  both pass through the origin, so current 0 voltage 0 hypothesis is correct. Now you have the set of observations  $x_i$ 's and  $y_i$ 's. I want to compute which line is better.

So, let us try to see if I consider the sum of the differences, what do I mean by some of the differences? If I consider  $y=x$  is a valid equation of line then I will consider  $y_i - x_i$ , that is the distance between the line  $y$  and  $x$  because here  $y$  is equal to  $x$  if I input  $x_i$  my point that I will get is also  $x_i$  because  $y=x_i$  and the actual output that I have got is  $y_i$  so I will consider  $y_i - x_i$  as one coordinate and  $y_i - 2x_i$  as another difference that will be a point over here  $y_i - 2x_i$  it will be a point over here.

But if I just consider the differences the problem is the differences may cancel each other some differences may be positive some differences may be negative. so, I do not want those differences to cancel out each other so what I will do is I will take square of them. So, in

particular we can define the sum square difference that is  $\sum_{i=1}^6 (y_i - x_i)^2$  and  $\sum_{i=1}^6 (y_i - 2x_i)^2$ .

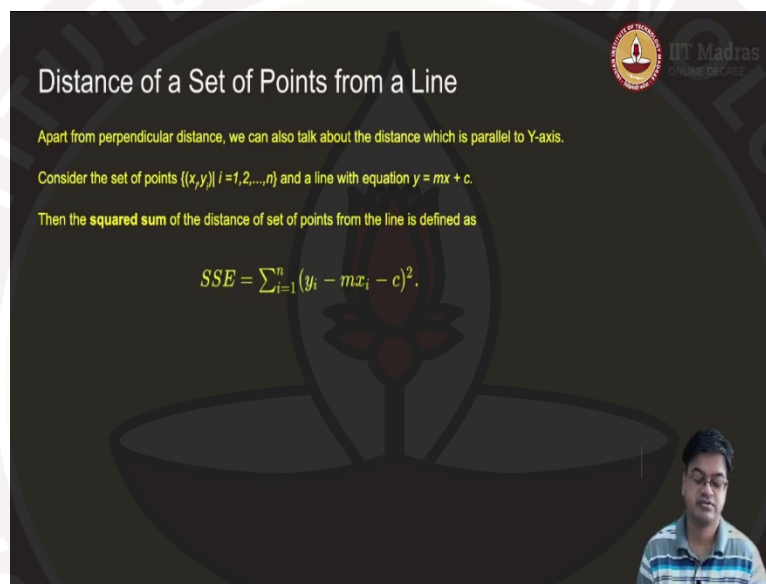
Now what this difference is calculating? It is calculating the difference between  $y_i$  and  $x_i$  in the first case and  $y_i$  and  $2x_i$  in the second case that is the error that we have made when we actually saw the output on the error the equipment has made or in our recording the error which is made in whatever way that is the error made. So,  $(y_i - x_i)^2$  and  $(y_i - 2x_i)^2$  right. Now what do you think which one will be better the one that will be better which will have a least difference.

So, you can actually put in these values and compute these differences and square them sum over them you will get the first difference is 5.09 and the second difference is 328.49. In this situation what should be our conclusion? Our conclusion should be that the difference where the difference is least that is 5.09 this must be a better line as compared to this line that essentially

reduces to a conclusion that  $y$  is equal to  $x$  is a better line as compared to  $y$  is equal to  $2x$  which is pretty evident intuitive from the figure as well.

So, you can see this figure you can see this chunk of points that are located around  $y=x$  and therefore the resistance of the equipment that is given to us must be 1 ohm that should be our conclusion. So, I want to introduce a notion of this kind to handle the real-world problems. So, let us see what is that notion? In this case you were very lucky the lab technician has given you the set of points or the resistance values there are only two resistance values.

**(Refer Slide Time: 13:49)**



**Distance of a Set of Points from a Line**

Apart from perpendicular distance, we can also talk about the distance which is parallel to Y-axis.

Consider the set of points  $\{(x_i, y_i) \mid i=1, 2, \dots, n\}$  and a line with equation  $y = mx + c$ .

Then the **squared sum** of the distance of set of points from the line is defined as

$$SSE = \sum_{i=1}^n (y_i - mx_i - c)^2.$$

But real life is not that lucky, so there they may not give you the set of values, and you want to find out what is the best line that is passing through these set of points. In that case this notion of a distance of a set of points from a line may help. So, what is this notion? First of all, we know one notion is perpendicular distance but that perpendicular distance may not be of much use when we are coming to the real-world perspective.

In that case we talk about the distances that are parallel to  $y$  axis from the distance of a points that are parallel to  $y$  axis. So, in particular if you have been given  $n$  points  $\{(x_i, y_i) \mid i=1, 2, \dots, n\}$ . You just plot this equation  $y=mx+c$ . Now remember here this equation is valid when it is not a vertical line. If it is a vertical line this equation is not valid. And if it is a vertical line you do not need such a complicated procedure to estimate it.



So  $y = mx + c$  is our standard equation of line which is a slope point form or slope-intercept form to be precise and then as in the previous case we have defined the squared sum of the distance of the set of points from the line. So, in the previous case  $y_i - x_i$  but in this case what should it be

$(y_i - mx_i - c)^2$  and you have to sum over all of them. So,  $\sum_{i=1}^n (y_i - mx_i - c)^2$ .

So, we call this as sum squared error or some squared distance, sum squared error so the abbreviation is SSE.

$$SSE = \sum_{i=1}^n (y_i - mx_i - c)^2$$

Now the fact is when we are handling a general problem, we do not know what will be  $m$  and what will be  $c$ .

**(Refer Slide Time: 16:08)**

**Least Squares Motivation**

- In general, this raises the following question
- Given a set of points, how to find the line that fits the given set of points?
- In other words, what is the equation of the best fit line for given set of points?

In other words, if I need to find the equation of line  $y = mx + c$ , then the question can be reframed into two questions.

- What is the value of  $m$  and  $c$  that best fits the given set of points.
- What is a meaning of best fit?

**Best Fit:** Given a set of  $n$  points,  $\{(x_i, y_i) | i = 1, 2, \dots, n\}$ , define

$$SSE = \sum_{i=1}^n (y_i - mx_i - c)^2.$$

Find the value of  $m$  and  $c$  that minimizes SSE.

So, our goal should be if I want to find the best line, I want to find the best line passing through this point what should be my goal. So, these raises two questions if I have some square, I want to know the value of  $m$  I want to know the value of  $c$ ? So, given the set of points how to find a line that fits the given set of points remember now I am not uniquely determining the line I am saying but that fits the given set of points.



The line may not pass through any of the points in this particular case in other words  $y=mx+c$  so what is the equation of the line that best fits the given set of points. This will mean I need to find an equation of a line  $y=mx+c$  and then the question can be reframed into two questions that is what do I mean by the value of  $m$  and  $c$  that best fits the line and then I have to define what is the best fit according to me.

Obviously, the best fit according to me will be the sum squared error minimization. And so, if I define SSE in this manner then I want to find the values of  $m$  and  $c$  that minimize SSE but this is right now beyond our scope as so far, we have handled only linear terms. But if you look at these terms, they appear to be in the form of squares of something. So, we need to divide some strategies in order to find this minimization for  $m$  and  $c$  so with that we will see in few upcoming videos of the course, thank you.

