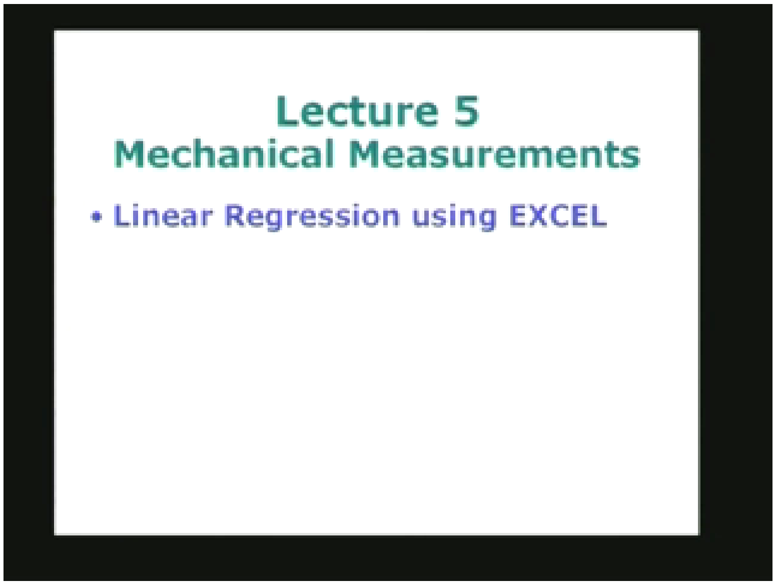
Mechanical Measurements and Metrology
Prof. S. P. Venkateshan
Department of Mechanical Engineering
Indian Institute of Technology, Madras
Module - 1
Lecture - 5

Regression Analysis

This will be lecture number 5 in Mechanical Measurements. Towards the end of the last lecture we talked about linear regression. We took an example and quickly went through the example and what I am going to do in the present lecture is to continue from there. The slide shows the plan of action for this particular lecture.

(Refer Slide Time 1:21)



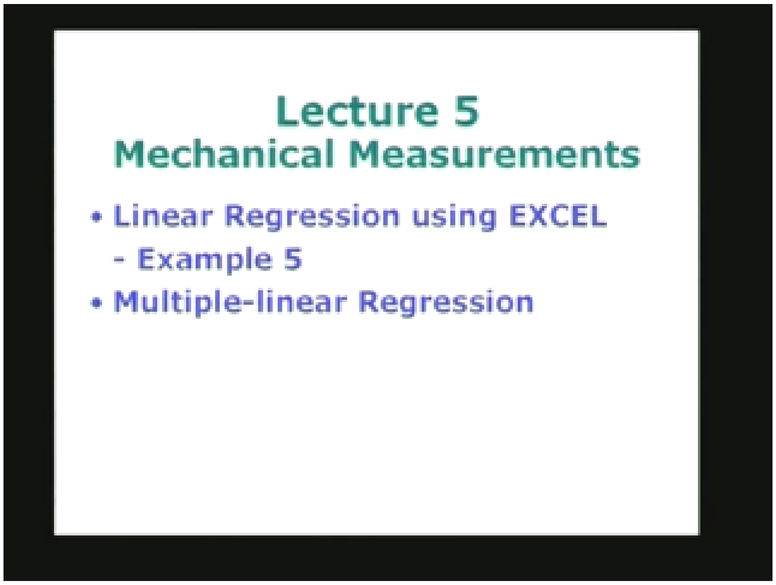
Lecture 5
Mechanical Measurements

- Linear Regression using EXCEL

So what I am planning to do is to do linear regression using EXCEL, which is the program available in Microsoft Windows Office suite and is very useful for regression analysis. By that we can avoid lot of computations which have to be done by using the calculator. The Windows environment in EXCEL makes it very simple and very easy. We will look at that as we go along.

The next topic of course will be through example 5 and the next topic will be multiple linear regression.

(Refer Slide Time 2:08)

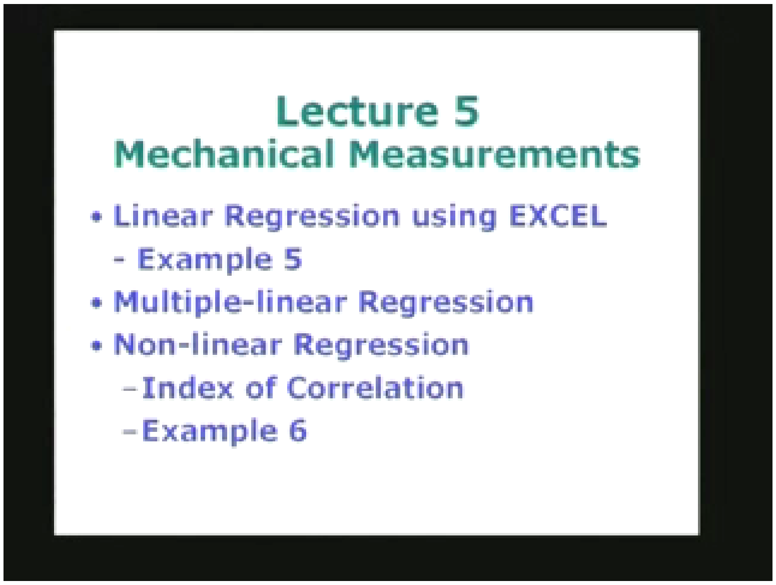


Lecture 5
Mechanical Measurements

- Linear Regression using EXCEL
  - Example 5
- Multiple-linear Regression

I am going to give some hints as to how it is done, and I will use mostly the board to describe the procedure. Then I will take the topic of nonlinear regression which is really what is required in many cases of interest to us.

(Refer Slide Time 2:23)



Lecture 5
Mechanical Measurements

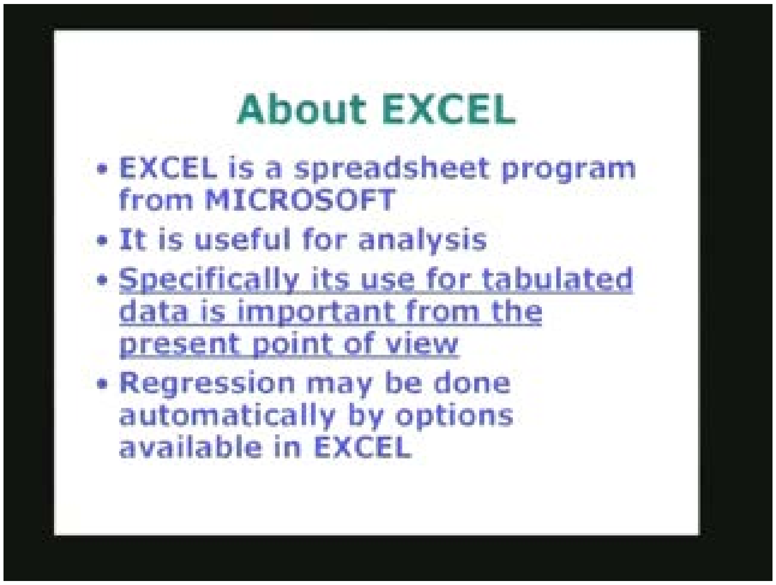
- Linear Regression using EXCEL
  - Example 5
- Multiple-linear Regression
- Non-linear Regression
  - Index of Correlation
  - Example 6

The nonlinear regression is more general, and in fact, it encompasses whatever we have done earlier on linear regression. But a few new concepts are required, and one such concept which determines the goodness of fit in the case of nonlinear fit is the index of correlation. So we will look at the index of correlation, how it is obtained and what is the basis for it. Again that will be worked on the board and I will follow it up with example

6, which will be using EXCEL again. So before we proceed with the actual content of the lecture, let me just digress a little bit and we will talk a little bit about EXCEL, which may be familiar to many students but it is just nice to find out what its capabilities are, what it can do and how it works.

So the EXCEL, which is a spreadsheet program, actually (Refer Slide Time 3:30) is a program environment. It provides us an excellent tool for doing data analysis and especially for those who are going to do experiments either during their undergraduate program or later on will find it very useful. So I will talk to some extent on EXCEL. Actually when I take up an example, automatically it will become clear as to what we are talking about. The point to note is that EXCEL is a tool for analysis; it is very useful and as far as we are concerned, we are specifically interested in its use as shown in the slide. Can we go to the slide please?

(Refer Slide Time 4:34)

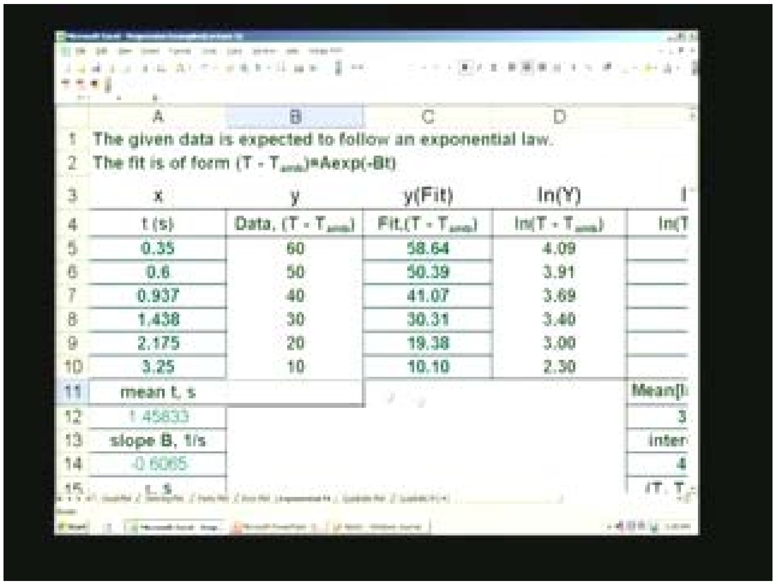


About EXCEL

- EXCEL is a spreadsheet program from MICROSOFT
- It is useful for analysis
- Specifically its use for tabulated data is important from the present point of view
- Regression may be done automatically by options available in EXCEL

Specifically it is used for tabulated data. It is important from the point of view of our present discussion. The advantage of the EXCEL spreadsheet program is that regression may be done automatically by using the options available in EXCEL. And what I am going to do is to show how EXCEL can be used for doing or performing a linear regression by taking a specific example. I will indicate the steps involved in doing this analysis. I will also indicate how all the statistical parameters can be automatically obtained using EXCEL. The Windows EXCEL, as I was mentioning, is a useful thing which provides us an easy way of doing the analysis, which is always to be done after performing an experiment. The analysis of data is also called post processing in the terminology of what we engineers use. So what I am going to do is to first introduce the EXCEL as a program environment, indicate how it can be used, and then immediately follow it up with an example, which is going to be linear regression example, and I will indicate how it is done in a very easy fashion.

(Refer Slide Time 5:58)



	A	B	C	D	
1	The given data is expected to follow an exponential law.				
2	The fit is of form (T - T_{amb})=Aexp(-Bt)				
3	x	y	y(Fit)	ln(Y)	l
4	t (s)	Data, (T - T_{amb})	Fit,(T - T_{amb})	ln(T - T_{amb})	ln(T
5	0.35	60	58.64	4.09	
6	0.6	50	50.39	3.91	
7	0.937	40	41.07	3.69	
8	1.438	30	30.31	3.40	
9	2.175	20	19.38	3.00	
10	3.25	10	10.10	2.30	
11	mean t, s				Mean[l
12	1.45833				3
13	slope B, 1/s				inter
14	-0.6065				4
15	t, s				(T, T

As you see from the slide, I have opened the EXCEL worksheet. This is called an EXCEL worksheet and you will see that it is full of tables. For example, this is a cell. In this cell, I can write a statement. For example, I will just say something here; I can just write something, I can erase it and then I can put a number or I can put a formula. For putting a formula, I have to put "equal to." That means whatever is going to follow is going to be a formula. For example, exponential of let us say 2: so if I come out of that cell, the formula has been calculated e to the power of 2 is 7.38. It has calculated; because whatever formula you give here, it will calculate immediately and give the value.

The formula can also use, I will indicate here, for example, I will say "equal to." That means I am going to put a formula there. Let us say there is a cell which is identified as A, column A and row 12. This is the cell I am talking about, so I am going to modify this. For example, I will say A 12. That is the number in the cell number A 12. I will multiply it by a factor of 2 to indicate what we can do. Star 2 and immediately it will give you the value 2.91, which is 1.458, multiplied by 2 is 2.91, etc.

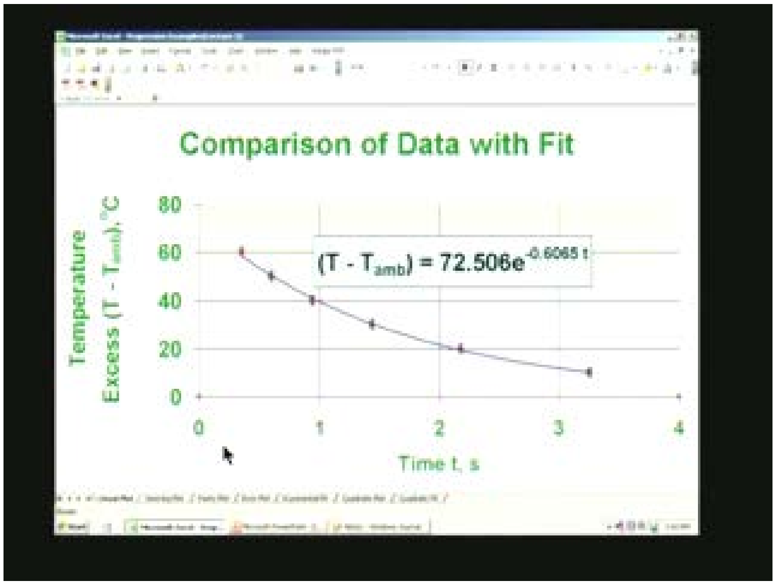
Actually any cell here, I can refer to any other cell in this environment, any number of cells here. From any cell I can take the number in that cell and perform some operation by putting "equal to" here and then invoking that cell number; invoke it and then modify it using some formula. So the advantage of this environment now is very clear. I am able to enter text; actually I have entered text here as you can see this. Here I have entered the text and in fact the text appears in the text box, which is not very clearly seen, but it is here. It appears there at the top; it is going to appear here in the text box. And the text bar contains all the text I have entered here, the given data is expected to follow an exponential law and that is what you see at the top here. And I can just come out the cell and it gives you what is in the cell in which you are currently positioned. In this cell I have simply written X and that's what you see in the formula bar or the text bar. This bar here if you put “equal to,” it becomes a formula bar. If there is no “equal to” sign preceding it, whatever you are going to give is going to be text bar.

With this background, which is just enough to understand what is going on, let us take an example. The example is given or stated by giving the values of the time and the corresponding temperature, which is measured in a certain transient experiment. So the time is given as T in seconds. Data is actually the difference in temperature between whatever we are measuring, the temperature of whatever object we are measuring, compared to the ambient temperature. So the difference is what I am plotting here and the t in seconds is also identified as X in our normal terminology. Data T minus T ambient is identified as Y. You will also notice that in these 2 cells as well as these 2 cells I have simply given the data given in the text. The text here shows the X value; this is the Y value and X corresponds to t in seconds; I have put seconds in brackets, Y is data T minus T ambient actually designed to use Celsius. I have not put it here because I have to increase the width.

For example I make it wider and put the temperature indicated there if we want. That means the cell can be resized by going up here and then dragging it and making it bigger or smaller as you want. Now let us look at the data, the data is given as 6 data points: 1, 2, 3, 4, 5, 6 and there are corresponding 6 values of temperature. So 0.35 60, 0.6 50, 0.937 40, 1.438 30, 2.175 20 and 3.25 10: this is the given data. The first thing which one will do is to make a plot of this, and for making a plot, you just block these cells. The first one corresponds to X values, the second one corresponds to Y values. These are pairs of values 0.35 60, 0.6 50 and so on. So I have blocked these two sets of cells here and if I go to the top, you will see something here, which is a chart wizard. So, if I press that, it will give me a way of making a plot; that is the great advantage of the EXCEL environment.

Because you can give the data in the form of table, put it in those cells and then just by using this chart wizard, I will be able to make a plot directly. For example, I will make a scatter plot. In fact you will see these are the types of plots, and what I will do is I will take a scatter plot. That means X and Y are going to be plotted normally like what you will do. And I can either have only points or I can have points as well as joining lines, only joining lines and so on and so forth. So, different options are available. I can choose whichever option I want, then go to next button, and then you see that the plot has come here. And if you want, you can make it bigger and you can either go back and make some changes, or you can go to finish and complete the plotting process. What I will do is I will cancel this because I have already plotted it. I will show the plot and the plot is the usual plot, which we normally do.

(Refer Slide Time 12:44)


Comparison of Data with Fit
Temperature Excess (T - T$_{amb}$), °C
80
60
40
20
0
0
1
2
3
4
Time t, s
$(T - T_{amb}) = 72.506e^{-0.6065t}$

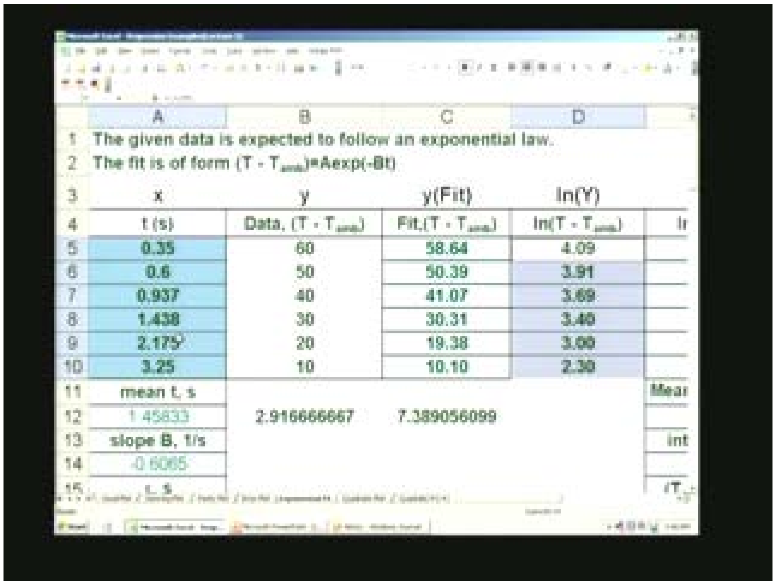
The plot has many options and these options are, I can use the text for the title, I can also indicate the axis titles and also the axis numbers corresponding to the divisions. I will come to this formula a little later. So what I have done here is to compare the data with fit but right now we don’t know what the fit is. We are going to come to that later; right now just look at the data. The data is the points shown here. The 6 points which have come here and the data is plotted just to indicate that it is not a straight line. It is very clear immediately that the plot is not a straight line. Therefore I want to find out whether there is a different way of plotting, which will probably give a straight line behavior.

That means if I take a look at the data which I took from the first worksheet, this data is not giving a straight line relationship between X and Y. So in order to find out whether we can linearize this, if you remember in the previous lecture we have said that we can either have a log–log graph or a semi-log plot to find out whether the plot or the relationship can be linear in any one of them. So what I will do is I will go and make a semi-log plot. For that I can actually calculate the logarithm as it is shown here (refer slide). This will be logarithm of Y or logarithm of T minus T ambient. This is the logarithm value—I have taken of the logarithm with respect to base e—therefore ln. Incidentally it is done by simply taking the logarithm with respect to base e ln; in bracket I have given B5. If you go back here, you will see that, B5 is this cell. So the formula here is logarithm of B5, which is in the cell number B5. Whatever number is here it will take the logarithm and indicate here.

There is another advantage. Suppose I remove this, I don’t have this. For example, I deliberately deleted those things. Now what I will do is I will drag this down and I will copy it; if you go to the edit menu you will say fill down and the numbers come back. That means I don’t have to write the formula every time; if the same formula is valid for more than one cell, all I have to do is to copy the formula down. By filling in the formula what happens is this will be logarithm of B5, next cell logarithm of B6. It means it has gone already to the next cell and therefore copying the formula will also change the reference to the appropriate cell. In the previous case it was B5 it became B6 and so on. It will automatically run down the formula will calculate as we go down by taking the suitable reference to the suitable cell.

So with this background let us see what I have done in the next plot. I have plotted this one and this one. X is as such without any change because I am making a semi-logarithmic plot.

(Refer Slide Time 16:17)

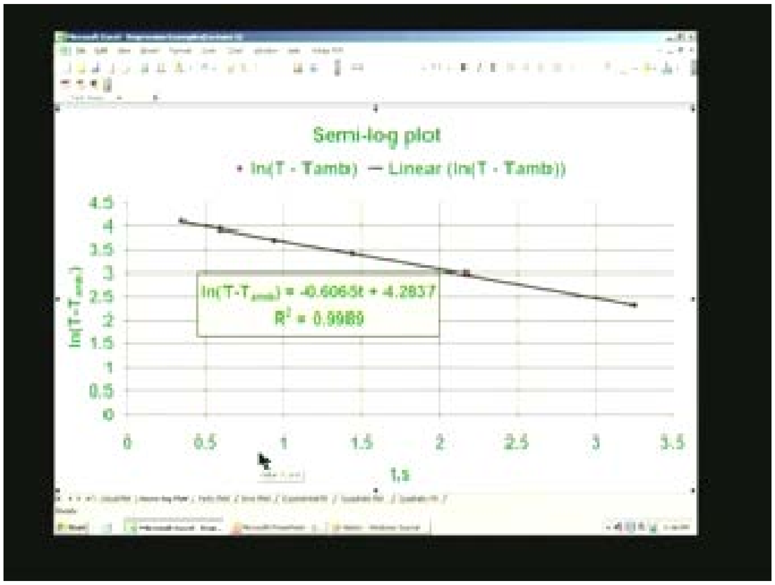


The given data is expected to follow an exponential law.
The fit is of form $(T - T_{amb}) = A\exp(-Bt)$

	A	B	C	D	
3	x	y	y(Fit)	ln(Y)	
4	t (s)	Data, $(T - T_{amb})$	Fit,$(T - T_{amb})$	$\ln(T - T_{amb})$	ln
5	0.35	60	58.64	4.09	
6	0.6	50	50.39	3.91	
7	0.937	40	41.07	3.69	
8	1.438	30	30.31	3.40	
9	2.175	20	19.38	3.00	
10	3.25	10	10.10	2.30	
11	mean t, s				Mear
12	1.45833	2.916666667	7.389056099		
13	slope B, 1/s				int
14	-0.6065				
15	t, s				$(T_{-}$

Time axis is linear; the temperature axis is nonlinear and logarithmic. So I am going to choose these two cells, these two columns and make the plot. That is what I have done here. It is a semi-log plot, which shows the logarithm of T minus T ambient as a function of t in seconds, and again we will come back to this later. So we have 6 data points here and you see that it looks like there is a nice straight line relationship between the data points and in fact I have shown the data, the straight line which looks like a reasonable representation of the data.

(Refer Slide Time 16:45)

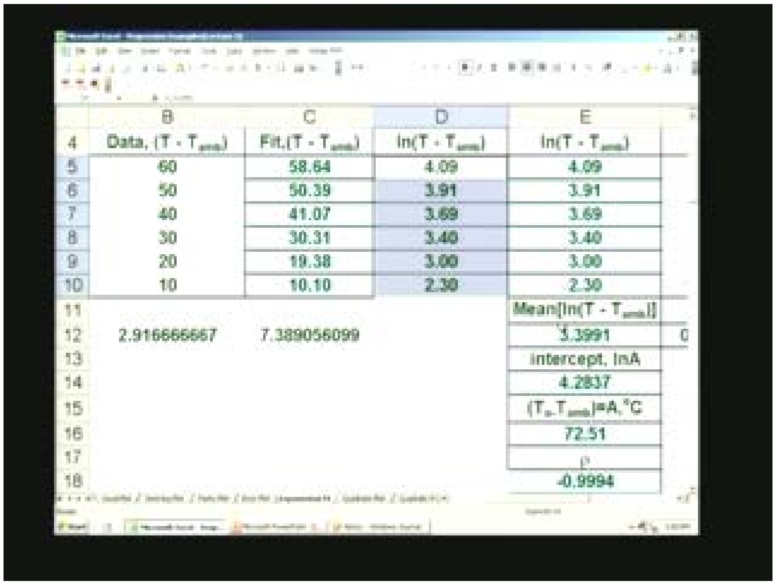


So in order to find out how to get the line, all I have to do is to go to the chart here and there are several options here. There is an option called add trend line. If I press this trend line it invokes several possibilities. The trend can be: this is linear, this is logarithmic, this is polynomial, this is power log type, exponential and moving average. So let us not worry about all of them. Right now I know that I am expecting a straight line relationship so I will click this. That means the trend line I am going to get is going to be a straight line and there are options. I can, for example, ask it to display the value of the correlation coefficient. If I want, I can also ask it to display the equation on the chart. That means the regression equation will immediately appear on the screen and I can do that. So these are all the choices I have. If I just say ok it is already going to give the equation. This is the equation. I will make it slightly bigger so that we can see it better; so that is the equation. In fact that is the one I brought here. What I will do is I will go back and undo what I did, so that the trend line option is not there. So I will just clear this so that I don't have to worry about it.

What I see here is the formula logarithm of T minus T ambient is equal to minus 0.6065t; so this is the slope parameter, plus 4.2837 is the intercept parameter and also it gives me the R square value 0.999 or 0.9989. That means the trend line was obtained simply by invoking the trend line from the chart option. If I add trend line, all I have to do is just do that and the calculation which we did in the last class of making the tables: adding all the x values and the y values, the product of xy, the product x square and then finding out the covariance, variance and then determining the parameters, the slope parameter and the intercept parameter, all that is done automatically by the program. And that is the great advantage in that. In fact what I did was after doing semi-log plot and finding what the appropriate relationship was, I have gone back and I have plotted that in the form of an exponential plot because, if you remember, the semi-log plot shows a straight line relationship between these 2 quantities. The relationship between the 2 variables is exponential in nature. And in fact the exponential relationship T minus T_{amb} equal to

72.506 comes from the constant. e to the power of minus 6065t is the slope parameter, that is the slope parameter and constant and the 2 of them are going to come out alike. In fact I am going to show how I do that. So let us look at the next slide here.

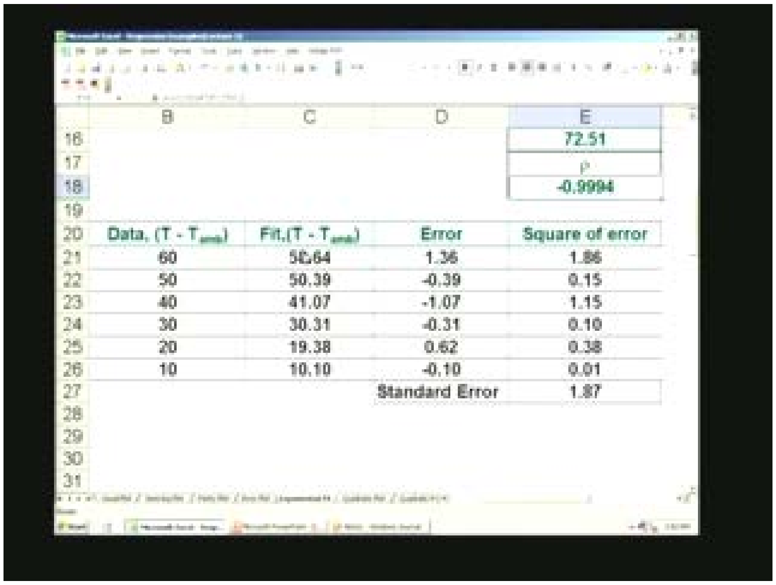
(Refer Slide Time 21:12)



	B	C	D	E	
4	Data, (T - T_{amb})	Fit,(T - T_{amb})	ln(T - T_{amb})	ln(T - T_{amb})	
5	60	58.64	4.09	4.09	
6	50	50.39	3.91	3.91	
7	40	41.07	3.69	3.69	
8	30	30.31	3.40	3.40	
9	20	19.38	3.00	3.00	
10	10	10.10	2.30	2.30	
11				Mean[ln(T - T_{amb})]	
12	2.916666667	7.389056099		3.3991	0
13				intercept, lnA	
14				4.2837	
15				(T_o-T_{amb})=A,°C	
16				72.51	
17				r	
18				-0.9994	

So I get the intercept, 4.2837, the logarithm of the intercept. Therefore all I have to do is take the exponential of this. That is nothing but the inverse of the logarithm and I have that here, e to the power of E14, E14 is the number I have here. E14 is the number corresponding to E and 14 is the number I have taken. The exponential of that and that is going to come in this, actually if I go and just click here and put a cursor there in the text bar or the equation bar, it immediately indicates there is E14. It immediately shows, with a blue border in this case, the active cell from which I have taken the information. So this cell takes the value in this cell and immediately calculates with whatever formula I have given and the formula is the exponential of E14, and that is 72.51. And the correlation coefficient can be calculated. You saw there that the correlation coefficient was given here already. It can also be calculated, if one wants, by doing the calculations like what we have indicated earlier, to calculate the mean values, slope and so on and so forth. You will be able to do that, so one can do that. In fact one can also do the following: I want to calculate what is the expected error in this fit, or the so-called standard error in the fit. So for that, I compare the data with the fit.

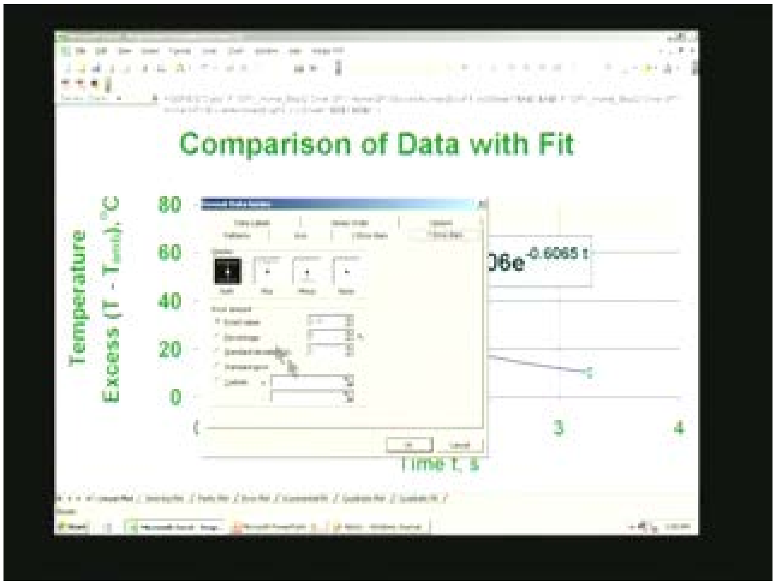
(Refer Slide Time 23:08)



	B	C	D	E
16				72.51
17				ρ
18				-0.9994
19				
20	Data, (T - T$_{amb}$)	Fit,(T - T$_{amb}$)	Error	Square of error
21	60	58.64	1.36	1.86
22	50	50.39	-0.39	0.15
23	40	41.07	-1.07	1.15
24	30	30.31	-0.31	0.10
25	20	19.38	0.62	0.38
26	10	10.10	-0.10	0.01
27			Standard Error	1.87
28				
29				
30				
31				

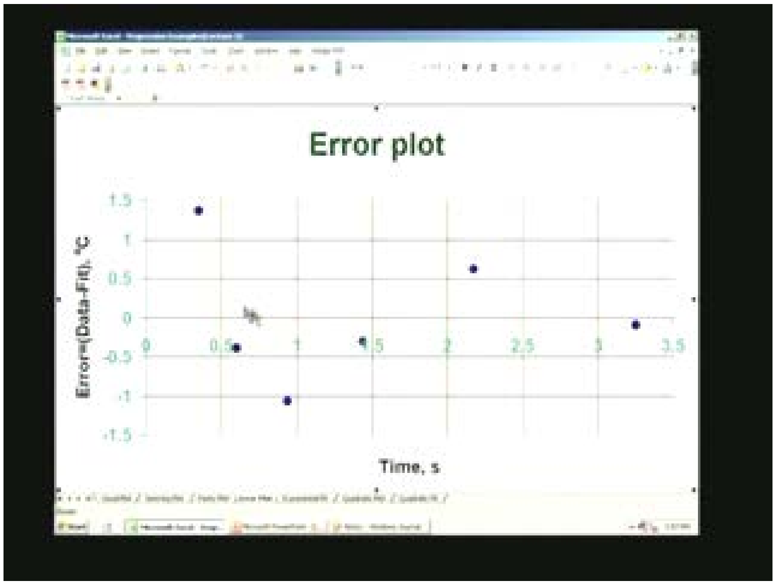
I have written here the numbers corresponding to the fit. How do I do that? It is very simple. The fit is given by the formula, 97.minus 25. This is your 72.51 e to the power of minus A5 by 1.6487. This A5 is the t. 1.6487 is nothing but 1 over the value, which was shown here, I have taken that slope parameter, the 1 over slope instead of this form of slope. So this is the formula which I have given here. And I have calculated this formula down and therefore these are the fit values and you can also see that the value of the given data is 60. This is 58.64; this 50 50.39 40 41.07: you see that they are close to each other. That means the fit is certainly a good representation of the data and in fact in order to further find out whether it is good or bad, I am going to compare or obtain the standard error and that is obtained by taking square of the difference between T minus T$_{amb}$, the data and the fit, and then adding all these squares. These are the errors and then the square of the error is given here. You find out the standard error. I am taking 1.96 for 95% confidence level, square root of sum of these. Actually the sum is a function available in the calculation and I will just explain how it is done, E21 to E26 are the values I am adding and then dividing by 4. I am going to divide by 4 because the number of degrees of freedom that I have is n, the number of points, 1, 2, 3, 4, 5, 6 minus the number of parameters 2; that is going to become 4. So this is the standard error expected in the calculation. I calculate it here. Now I can go to the plot and in fact show the standard error as an error bar. So if I go to the data point Forma data series, I have an option for giving the error bar. What I have done is that I have written 1.87 here in this option, so that it will give a bar like what is shown here and that's what you see here.

(Refer Slide Time 25:38)



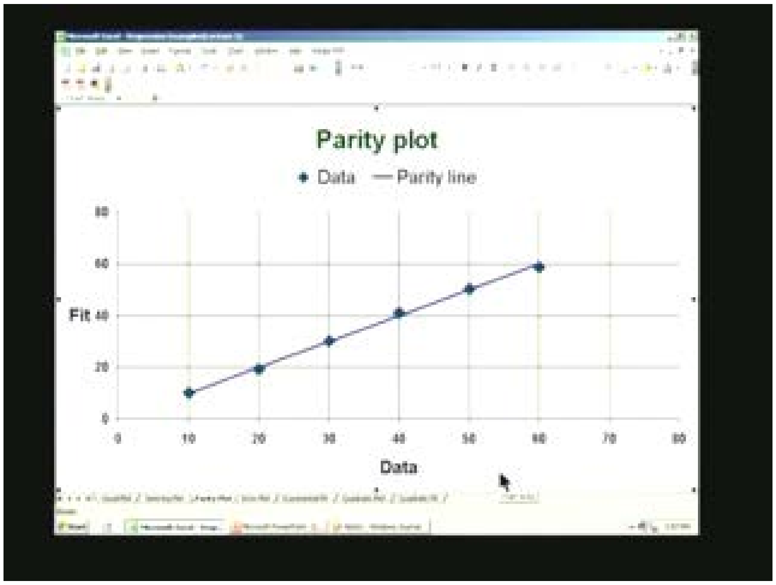
So if you want, we can expand it slightly so that we can look at that. In fact I can do a little better; I think the option is not here. Anyway you can see that by simply keying in the data on to the worksheet and by manipulating the data very easily without much sweat, you are able to get the plot either in the form of a semi-log plot or in the form of usual plot, XY plot, or in fact I can do some other types of plots. What are the other types of plots? These plots are helpful in further finding out or further demonstrating how good the regression analysis has been. Whether the data we collected is well represented by the formula we have decided to have, the regression line—in this case exponential or straight line or any other formulas which we might have obtained or chosen. For that I can make 2 more plots. I can make an error plot and it shows the error at each data point. This plot is very useful because in this case you see that the error points are lying on both sides of the zero line. This is the zero error line; this line would have been the line on which all these data points would have been lying if there were no errors, because there are errors in each data point, these 2 points are on the positive side, these 4 points are on the negative side.

(Refer Slide Time 27:12)



And you will also see that the closer these points are to the zero line, the better is the fit. Indeed we can see that they are lying very close to the lines, the maximum being about 1.5 and the minimum being about minus 1.1 or something like that. So the error is between plus or minus 1.5.

(Refer Slide Time 28:12)

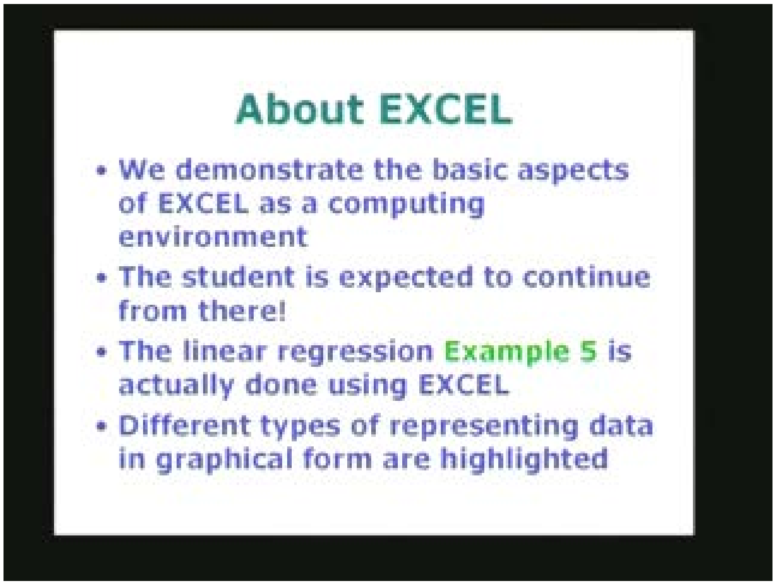


And in fact I can make one more plot which also gives me an indication of whether the fit is good or bad. What is this line? We have the plot, the fit on the y axis. That is the value of the y obtained by the regression line on the y axis. On the x axis I am going to use the data which is given. And if these two data and the fits of these two are going to agree completely or perfectly with each other, they will lie on a line. If I plot it using the same scale on this axis as well as this axis, it would lie on a 45° line. And in fact I have called that 45° line, the parity line. Parity line means the line of parity between the data and the fit. So you can see from here that the points are all lying very close to the parity line and therefore the fit is a good representation of the data, which has been collected. So it is very important.

To just summarize; I can make a plot like this to compare the data and the fit. I can also put the error bar on it. I can make a semi-log plot and I can do the same thing. I can make a parity plot and indicate how good the regression line is corresponding to the data. I can also make an error plot to find out what is the error between the line of the fit line and the data. So there are several ways of plotting: each one supporting the other or each one supporting the fact as to whether the fit is good or bad, and that is one possibility. Of course, the other possibility is to find out whether the correlation coefficient is good or bad and that is another thing which we can do.

And in fact I have calculated the correlation coefficient also. It is minus 0.9994; the negative sign because there is a decreasing trend as I have already indicated earlier, and this is another indication of how good the fit is. I have shown how the EXCEL program can be used to do a linear regression. Of course the example I took was of exponential fit, but as we recall from the previous lectures, an exponential fit can be converted into a linear regression model by a simple semi-log plot or by taking the logarithm of y values.

(Refer Slide Time 31:00)

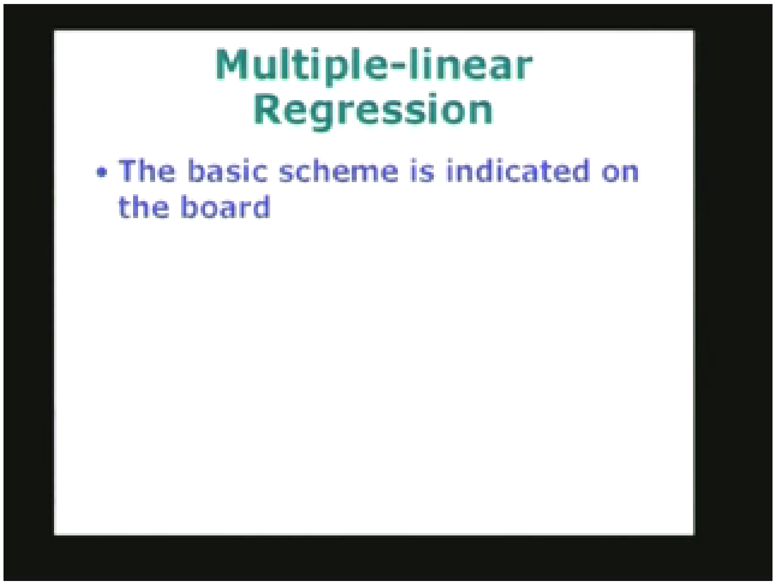


About EXCEL

- We demonstrate the basic aspects of EXCEL as a computing environment
- The student is expected to continue from there!
- The linear regression Example 5 is actually done using EXCEL
- Different types of representing data in graphical form are highlighted

So to just recapitulate we demonstrated the basic aspects of EXCEL as a computing environment and hopefully the student will have learnt enough so that he can continue from there. And also we took the linear regression as an example of the exponential fit and we demonstrated how it is actually done using EXCEL spreadsheet. And we also have indicated how different types of plots can be made to represent the data in a graphical form. And all these plots can be done by simply invoking the chart wizard, which was actually demonstrated to just show how the plotting is done. With this basic background, let us see what the other kinds of regression are that one may be interested in doing. I am going to talk about two things: a multiple linear regression, which is also a useful regression model. For many problems or many situations multiple regression may be useful.

(Refer Slide Time 32:00)

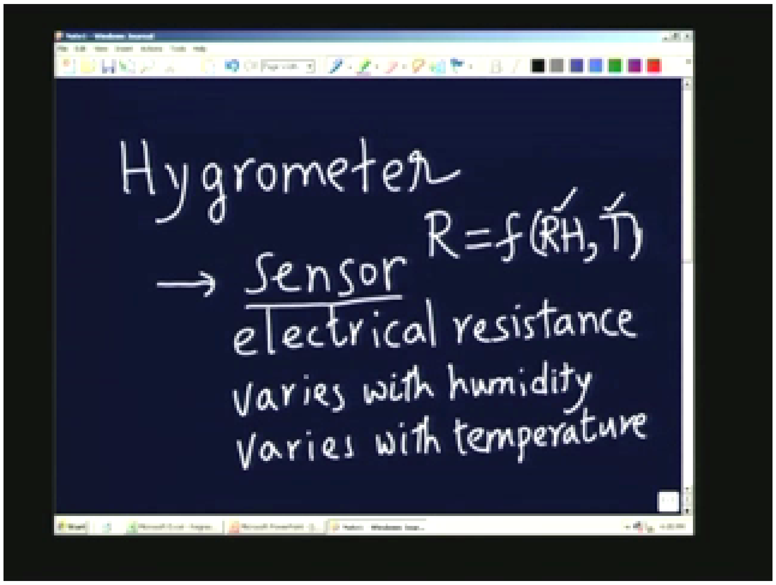


Multiple-linear Regression

- The basic scheme is indicated on the board

I will indicate the basic scheme on the board. Let me just briefly say where you expect this kind of a model to come. So I will just take a simple example. Let us take a sensor, which measures the presence of water vapor, moisture. It is referred to as hygrometer and hygrometer uses a sensor. We will not go into great detail. We will just assume there is a sensor and its property, some property of the materials used in the sensor. For example, the electrical resistance of the sensor element varies with humidity. Hygrometer measures the humidity in the atmosphere and the sensor, which is used in the hygrometer, has electrical property called electrical resistance, which varies with humidity. Unfortunately, it also varies with temperature. That is, if it were a function only of humidity, I wouldn’t have any problem, but it also varies with temperature. That means,if I say R is the resistance of the element, it is a function of humidity, which we can indicate as relative humidity, RH, and temperature.

(Refer Slide Time 34:18)



Hygrometer

$R = f(\overset{\checkmark}{RH}, \overset{\checkmark}{T})$

→ Sensor

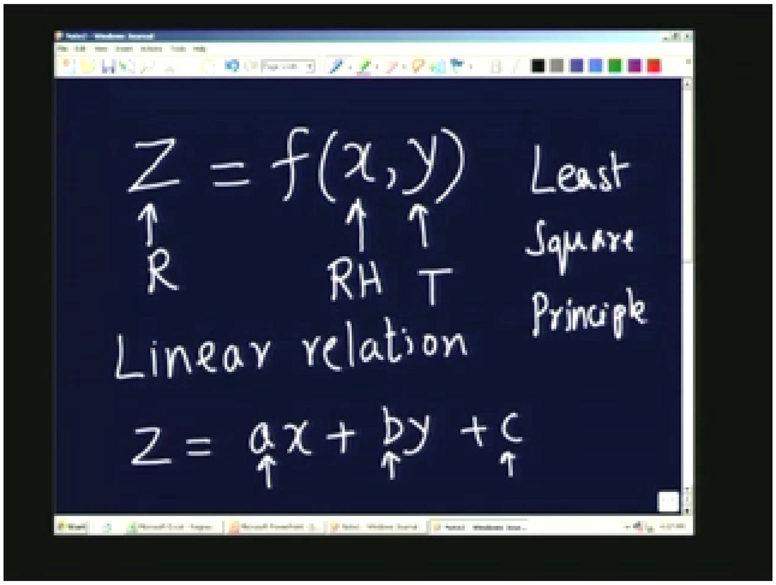
electrical resistance

varies with humidity

Varies with temperature

So the sensor responds to two quantities at the same time; I am just trying to take a very typical example. So let us look at how we are going to do this. Of course we do not know, to start with, what will be the type of function relationship. Suppose we now generalize, instead of taking the hygrometer example, I can take a general example. Suppose I have a quantity z which is a function of x and y. In the example of the hygrometer, this will be the RH. This will be the temperature and this will be the resistance. Of course we need not have only two variables here; we have more than two variables. However, if the relationship z equal to f(x,y) is linear that means we have z equal to ax plus by plus c. There are three parameters a, b and c. The regression model will have to obtain three parameters by using the same principle, which we have used earlier, the least square principle.

(Refer Slide Time 36:22)



$$\underset{\overset{\uparrow}{R}}{Z} = f(\underset{\overset{\uparrow}{RH}}{x}, \underset{\overset{\uparrow}{T}}{y})$$

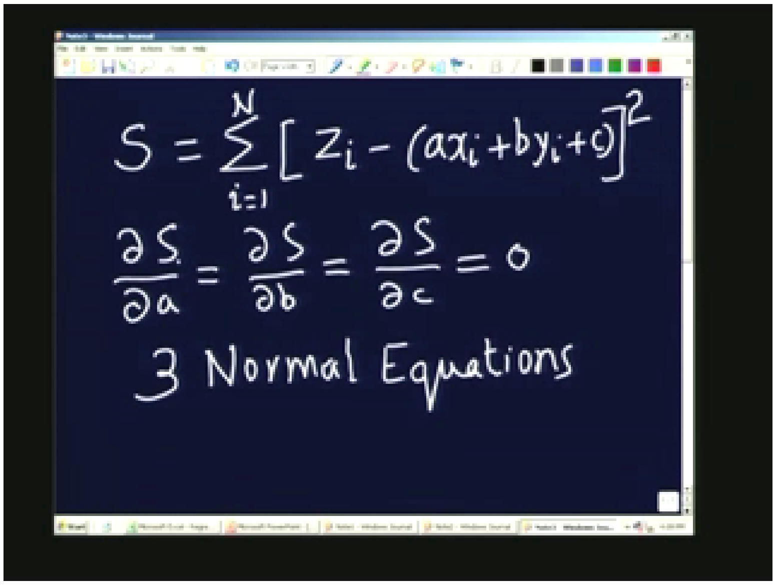
Least Square Principle

Linear relation

$$Z = \underset{\uparrow}{a}x + \underset{\uparrow}{b}y + \underset{\uparrow}{c}$$

So let me just indicate the steps involved in this without taking any specific example. I will just indicate the steps involved.

(Refer Slide Time 37:55)



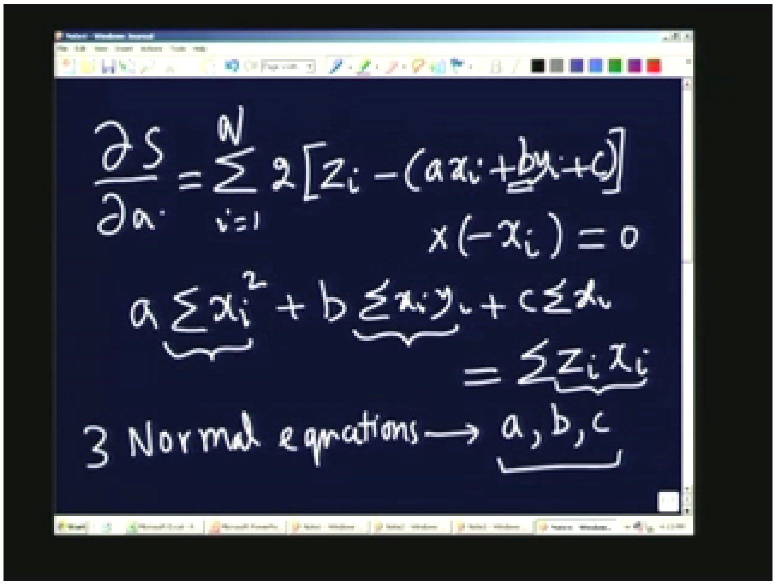
$$S = \sum_{i=1}^{N} [z_i - (ax_i + by_i + c)]^2$$

$$\frac{\partial S}{\partial a} = \frac{\partial S}{\partial b} = \frac{\partial S}{\partial c} = 0$$

3 Normal Equations

So least squares principle means, I am going to minimize some quantities S, sigma over i equal to 1 to N, which are the number of data I have: z_i minus ax_i plus by_i plus c whole square. So I have to minimize the quantity, which is shown here. So again you will see that the procedure is very simple. I have to set *doh* s by *doh* a, *doh* s by *doh* b, *doh* s by *doh* c, all equal to 0 and these will give 3 normal equations, which have to be solved simultaneously. Actually I will just give one normal equation, the students can work out how the others are going to look. So suppose I take *doh* s by *doh* a, let us see what is going to happen.

(Refer Slide Time 40:50)



$$\frac{\partial S}{\partial a} = \sum_{i=1}^{N} 2\left[z_i - (a x_i + b y_i + c)\right] \times (-x_i) = 0$$

$$a \underbrace{\sum x_i^2} + b \underbrace{\sum x_i y_i} + c \sum x_i = \underbrace{\sum z_i x_i}$$

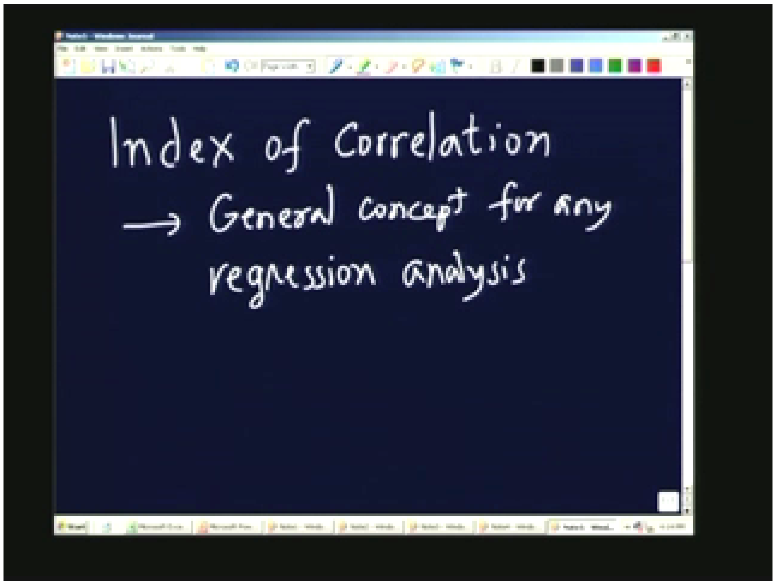
3 Normal equations $\longrightarrow \underline{a, b, c}$

So we have *doh* s by *doh* a equal to sigma over i equal to 1 to n, 2 into z_i minus ax_i plus by_i plus c multiplied by minus x_i; this is equal to 0. This is one normal equation. Similarly I have to obtain the normal equation for *doh* s by *doh* b equal to 0, *doh* s by *doh* c equal to 0, so 3 equations will come. Let us just look at the forms of these equations. So I have z_i multiplied by x_i; I will send it to this; minus into minus this becomes plus. This will become minus; I will send it to the right-hand side. This 2 is a factor, which multiplies every term. Therefore I need not include it.

So you will see that I have a sigma x_i square plus b sigma x_iy_i plus c sigma x_i is equal to sigma z_ix_i. So you see that the first normal equation is written down by taking the derivative with respect to a, and setting it equal to zero, and you will see that of course I have to put i equal to 1 to n for all the summation which I am just not doing it because we assume that we can supply that information, so it involves the square of the x values. Then you see that it involves the product of x and y; it also involves the product of z and x. In fact you can immediately see that when I take the derivative with respect to b, instead of x_i square, I will get x_iy_i here I will get y_i square, here I will get y_i and here I will get z_iy_i. So that is the second normal equation. When you take it with respect to c I will not get any product like this. So it will be just, a sigma x_i plus b sigma y_i plus c times n the number of data. This is equal to sigma z. So, three normal equations are written down and solved to get a, b and c. So it is a very straightforward and simple thing and in fact, the rest of the thing is to be done by the following method.

So in the case of single variable, if you remember, we talked about the correlation coefficient. Here we have a problem because there are too many variables which are there: two independent variables and 1 dependent variable. So how do we define the correlation between these two? So for that, what I will do is, I will introduce a general concept of index of correlation.

(Refer Slide Time 42:22)

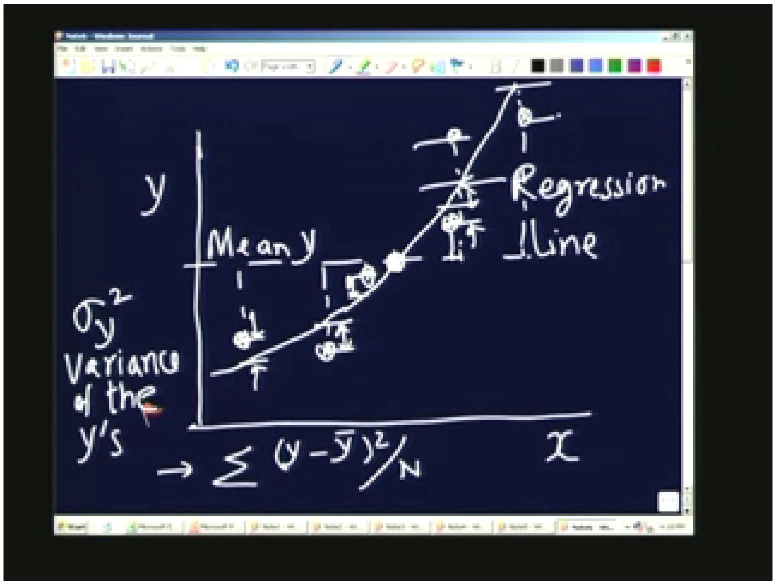


Index of Correlation

→ General concept for any regression analysis

In fact this is a general concept valid for any regression model; it need not be linear, it need not be multi-linear like what we are talking about. It can even be a nonlinear expression involving any complicated explanation for any regression analysis. General concept valid for any analysis. So let me just find out how to do this. So let us just try to see how one can go about doing this. Suppose I make a plot which of course can be done using EXCEL if you have actual data. So this is the y value; I am just doing it for a single variable but there is no need to restrict to single variable. This is x; the data was like this. So let us take data points like this and the relationship, the regression line or the regression curve, is like this. This is your regression line; I will generally say line. Line includes a straight line; it also includes a curved line. So if I look at this, suppose I identify the mean value of the y. So let us say mean value of y is here, with the mean y.

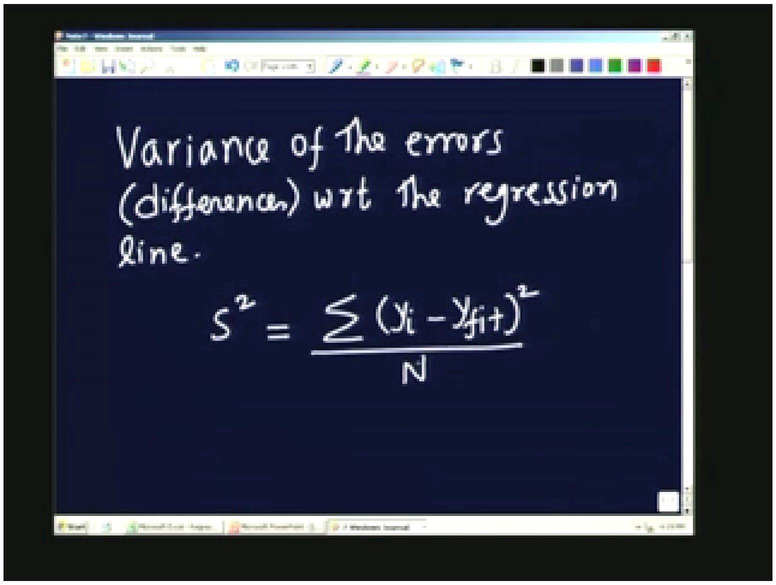
(Refer Slide Time 45:52)



So if you look at this mean y, of course, we know that this line will pass through the mean value. You see that with respect to the mean value of y, these are the differences. If I draw a line from there to there, these are the differences. I have just taken a few points. If we have more points, we will see a much better description of what is happening. So now I can look at the variation of y with respect to the mean and define sigma y square as the variance of the y's. This concept is already familiar to us: the variance of y's. Actually how do we define? This will be given by sigma y minus y bar whole square divided by N. Now if you look at this graph, I have got a line called regression line here, and with respect to the regression line there are errors. So this is the error or the difference between the regression line and the data.

Again we have a difference here and we have a difference here, so I am just indicating the differences like this. This is the difference I am talking about, and similarly here and here and so on. So now I can also define a variance for the errors. So I will just say I will go to the next screen so that it is not cluttered. So I can also define variance of the errors or if you want to call them as differences with respect to the regression line; it is a very interesting thing to do. So let us call that as s square or s square is the variance of those things. This will be nothing but sigma of y_i minus y_{fit} whole square divided by N.

(Refer Slide Time 47:05)



Variance of the errors (differences) wrt the regression line.

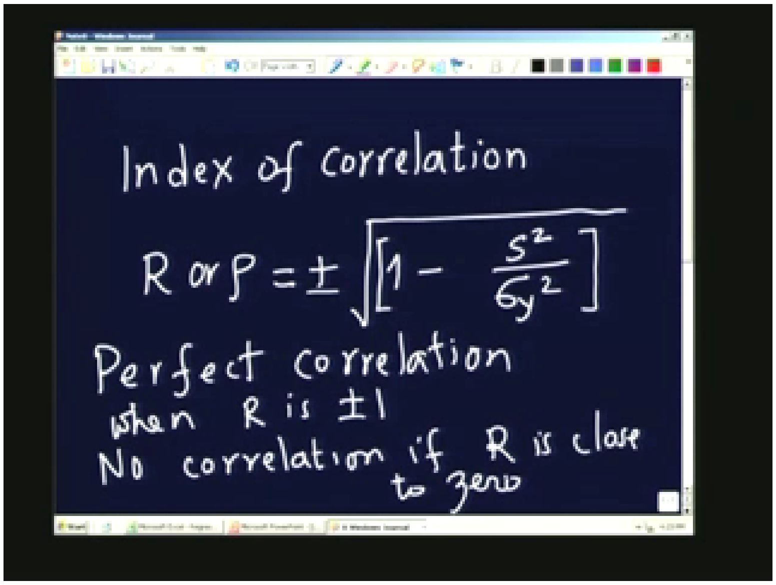
$$s^2 = \frac{\sum (y_i - y_{fit})^2}{N}$$

So we see that we have defined a variance of the y's given by sigma y square, which is given by this formula here: sigma i equal to 1 to n, y minus y bar the whole square divided by N or y_i minus y bar whole square divided by N. y bar is of course sigma y by N or sigma y_i divided by N. So just look at this quantity; this quantity is something which tells you how the data points are distributed with respect to the mean value of y.

Now we will look at the other one. The variance of the errors is given by this quantity. This is nothing but y_i minus y_{fit} whole square divided by N, y_{fiti}. Let us put it here, because at each point you must find out what is the value corresponding to that value of x. Find out the y_{fit} value, take the square of all these errors and divide by N, you get the variance of the errors.

Now I am going to compare the variance of the errors with respect to the variance of the ys. If the variance of the errors is small compared to the variance of y's, that means if I go back to the previous page, I am comparing this quantity, that means I am going to compare the variance with respect to the mean value of y bar and the variance with respect to the local mean value because the fit represents the local mean. So I am going to compare these two variances to know whether the variance with respect to the local value is small compared to the variance with respect to the mean values of y. Then I can say that whatever relationship I have obtained is good. So, how to put that mathematically? Let us define what I call the index of correlation,which is a mathematical expression. I will say this is the index of correlation; instead of calling it coefficient of correlation, I am going to call it an index of correlation. I can use the same symbol as I used earlier: either R or rho.

(Refer Slide Time 51:09)



Index of correlation

$$R \text{ or } \rho = \pm \sqrt{\left[1 - \frac{s^2}{\sigma_y^2}\right]}$$

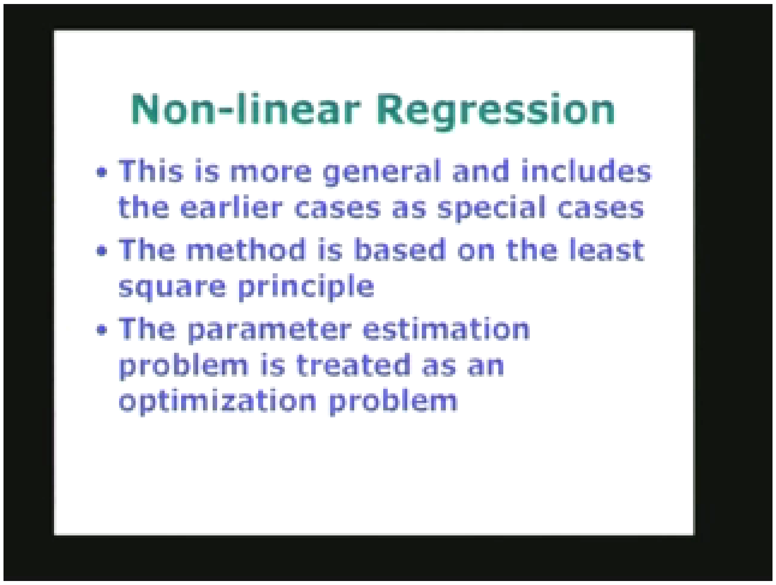
Perfect correlation when R is ±1

No correlation if R is close to zero

From the context, what we are talking about will be clear. It will be plus or minus square root of I am going to again define such that if this index of correlation is very close to plus 1 or minus 1, I get perfect correlation. If it is very small, I will have to make sure that it represents poor correlation. Therefore if I do the following, if I take 1 minus s square by sigma y square, put it under the bracket and then take the square root, this will have that property. That means perfect correlation when R is plus or minus 1 and close to plus or minus 1 and no correlation if R is close to 0. So the index of correlation is a more general quantity and in fact it can be shown that the index of correlation and the correlation coefficient are one and the same for a linear fit. So I am not going to do it here. I am just stating without proof. It can be shown that the index of correlation and the correlation coefficient are identical if we have a linear regression model.

So with this background let us look at the slide again and look at multiple linear regression. We have already shown how it is done. So I am going to look at general nonlinear regression. I will just be able to say a few words about the general nonlinear regression, and we will of course, continue in the next lecture. So this is more general and includes all the earlier cases we have considered and they will be simply special cases. The method is again based on the least square principle and the parameter estimation problem. If you remember, the least square principle gives me the sum of squares of differences between the data values which are given and the values obtained by the regression line or regression curve. That sum of squares must be made smallest. That means I am looking at a function whose value I must find or I must choose the parameter such that this sum becomes the smallest possible.

(Refer Slide Time 52:26)



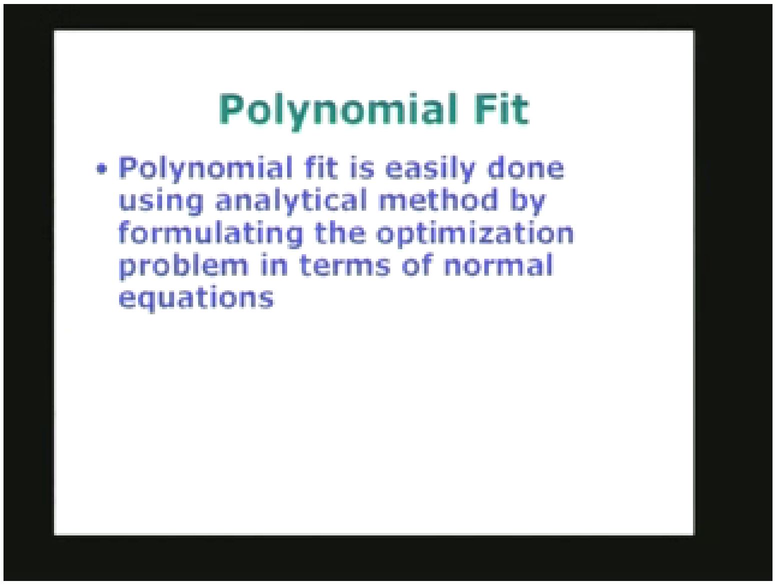
Non-linear Regression

- This is more general and includes the earlier cases as special cases
- The method is based on the least square principle
- The parameter estimation problem is treated as an optimization problem

That means it is like optimization formula. Optimization is finding either a maximum or a minimum for a function. Therefore in this case, we can treat it as an optimization problem and use optimization methods to obtain the solutions. Of course in few cases of nonlinear regression, like polynomial, it is possible to do regression again in terms of normal equation, solving them to get all the parameters. In the case of a general nonlinear regression model or problem, it is not possible to do that and therefore we have to do it using a mathematical technique, which is basically an optimization technique. There are several techniques available and what we will do is we will only indicate how it is done in a general way and not take it up for detailed discussion, because this will be outside the scope of the present set of lectures we are giving.

So I will talk about the polynomial fit and the polynomial fit is also a nonlinear fit. For example, I can just indicate a type of fit which is a polynomial fit.

(Refer Slide Time 54:12)

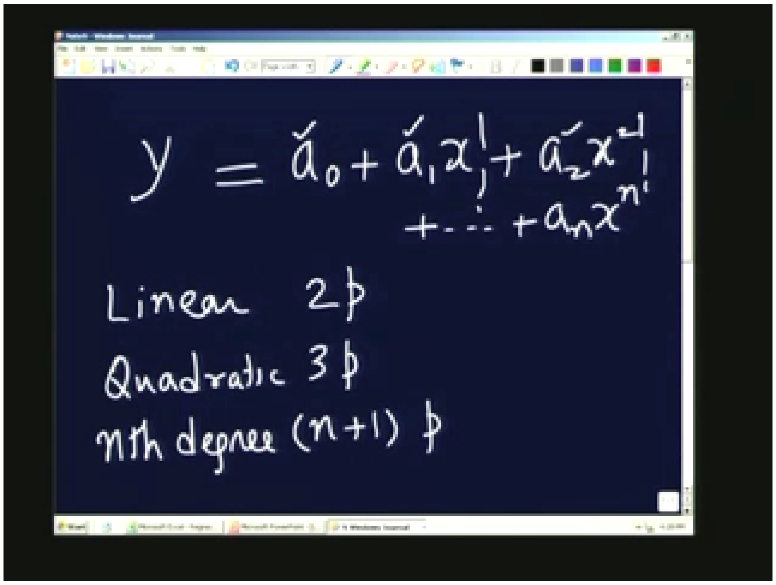


Polynomial Fit

- Polynomial fit is easily done using analytical method by formulating the optimization problem in terms of normal equations

Let me just indicate briefly.For example,if I have one variable y, which is related to x in the form of a polynomial, I can have the following. I can say a_0 plus a_1 x plus a_2 x square plus—it is going to stop somewhere. So you will see that for example linear stops here. It's a special case of a polynomial fit, it's actually a monomial. A quadratic will stop here, a cubic will stop with cubic term, and so on. And you will notice that linear has got 2 parameters. I will just say 2p, p stands for parameter, quadratic will be 3p, parameters a_0, a_1 and a_2. So this is 3p and in general, nth degree polynomial I have a_n ,a_0, so it is n plus 1 parameters.

(Refer Slide Time 55:50)



$$y = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n$$

Linear 2p

Quadratic 3p

nth degree (n+1) p

So what I will do in the next lecture is to take a simple example of a quadratic fit and show how it can be done using EXCEL and also try to understand how we calculate the goodness of the fit for that particular case. We will stop here and resume in the next lecture. Thank you.