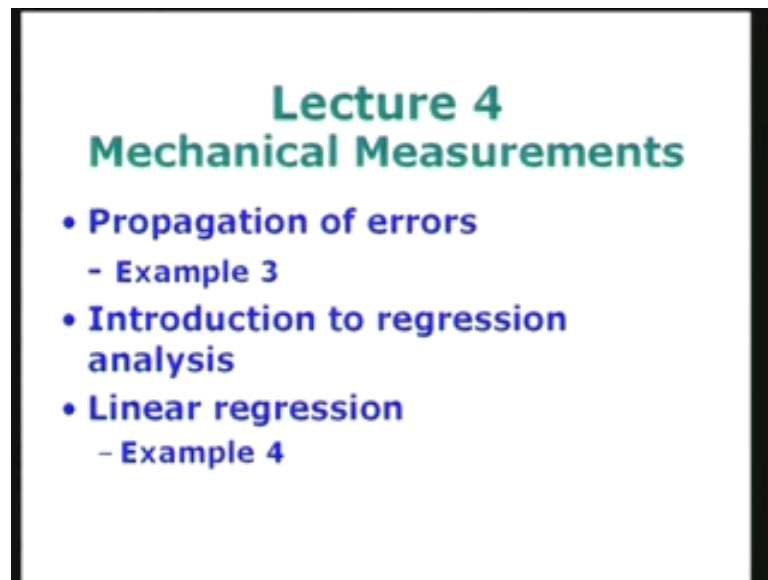


Mechanical Measurements and Metrology
Prof. S. P. Venkateshan
Department of Mechanical Engineering
Indian Institute of Technology, Madras
Module - 1
Lecture - 4
Propagation of Errors

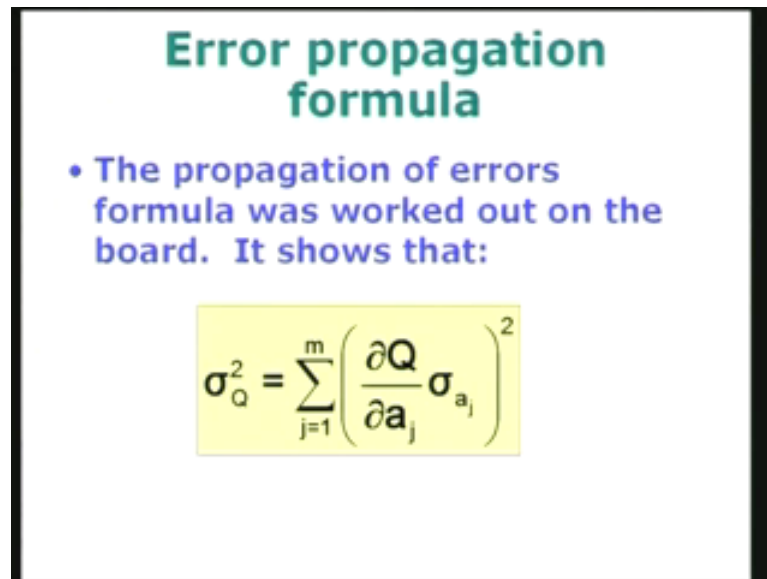
This will be lecture number 4 in our series on mechanical measurements. In the previous lecture, that is, lecture number 3, we were looking at the question of propagation of errors; that is the propagation of errors from the measured primary quantities to the estimated value of the derived quantity.

(Refer Slide Time 1.31)



We have already derived the formula and we will resume from there. We will also give an example, which will consider two different ways of solving a particular problem. Then subsequently, we will look at introduction to regression analysis. Regression analysis is the study of relationship between different measured quantities. They may be linear, nonlinear and so on. So we will be looking at the basic ideas of regression analysis and then subsequently we will take up linear regression analysis in a more detailed fashion. If time permits, we will be actually solving a specific problem. Towards the end of the last lecture, we actually derived a relationship, which is rewritten here in this slide.

(Refer Slide Time 2.11)



Error propagation formula

- The propagation of errors formula was worked out on the board. It shows that:

$$\sigma_Q^2 = \sum_{j=1}^m \left(\frac{\partial Q}{\partial a_j} \sigma_{a_j} \right)^2$$

So we have Q as a derived quantity and a_j s are the quantities, which are primarily measured, and Q is estimated. Q is a function of a_j for j equal to 1 to m . If the precision of each on the measurement of a_j is known, for example, σ_{a_j} gives you the standard deviation for each one of these measurements, then I found a product of the partial derivative (*doh*) Q with respect to a_j multiplied with σ_{a_j} . This is the first part. The partial derivative is nothing but the influence coefficient and σ_{a_j} is the precision of the particular measurement. I add such terms and square them from j equal to 1 to m . This gives you the variance in the error of the measured value or estimated value of the quantity Q . What is happening here is that the error in the primary quantity that is measured, which is given by σ_{a_j} , is contributing a certain amount to the error in the quantity Q and that is given by the influence coefficient multiplied by precision of that particular measurement, the sum of the squares. This is just like taking the magnitude of the vector or by the use of Pythagoras theorem, which gives you the sum of the squares is equal to the squares of the hypotenuse. In this case we are extending the idea of Pythagoras theorem to several dimensions instead of just 2 dimensions, which we normally use in geometry.

(Refer Slide Time 2.11)

Example 3

- The quantities A, p and T are measured and the quantity M is estimated. The formula relating the quantity M to A, p and T is given by:

$$M = A \frac{p}{\sqrt{T}}$$

So with this background let us take a look at a simple example, example 3. In this case we measure A, p and T. These are the 3 measured quantities and the quantity M is estimated and the formula, which links the value of M to A, p and T, is given by the formula here: M equal to A times p by root T. Let me just digress and tell you what we are trying to do. What is this problem? In measuring fluid flow sometimes we use sonic orifice and in that case the mass flow rate of the gas is given by some constant, which is obtained by some kind of a calibration process multiplied by the stagnation pressure and the stagnation temperature upstream of the orifice. So this is an example taken from fluid flow and let us see what we are going to do.

Further specification in this problem is as follows. The nominal values are A equal to 10, p equal to 20 and T equal to 4.84. What I have done is that I have chosen the values such that it is dimensionally homogenous and these values are just numerical values. The standard errors in these quantities A, p and T are also specified: sigma A is 0.1%, sigma p is 0.2% and sigma T is 1%. We would like to specify an error bar on the derived quantity M.

(Refer Slide Time 5.14)

Example 3 (Continued)

- The nominal values are:
 - ◊ $A=10$, $p=20$ and $T=4.84$.
- The standard errors in these are:
$$\sigma_A = 0.1\%, \sigma_p = 0.2\%, \sigma_T = 1\%$$

We would like to specify an error bar on the derived quantity M .

(Refer Slide Time: 6:05)

On the board we work out one way of solving this problem. A second method will then be presented in the form of slides.

I am going to work out on the board what I call as method number 1. So I will go to the board and in fact this is the formula which we have written down, towards the end of the last lecture. So I am going to use this as the basis for solving the problem and let us write down the formula which is given to us: M equal to Ap by square root of T .

(Refer Slide Time 9.34)

$$M = \frac{Ap}{\sqrt{T}} = \frac{10 \times 20}{\sqrt{4.84}} = 90.91 \checkmark$$

$$\checkmark I_A = \frac{\partial M}{\partial A} = \frac{p}{\sqrt{T}} = \frac{20}{\sqrt{4.84}} = 9.09 \checkmark$$

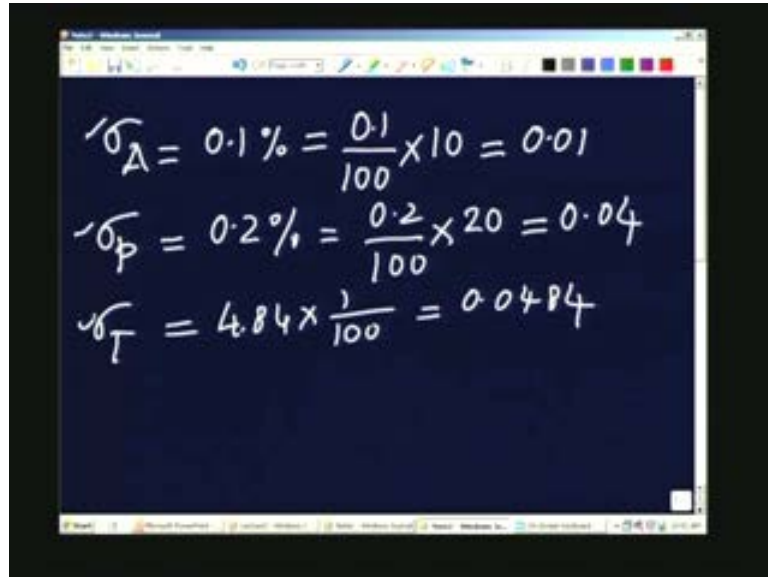
$$\checkmark I_p = \frac{\partial M}{\partial p} = \frac{A}{\sqrt{T}} = \frac{10}{\sqrt{4.84}} = 4.545 \checkmark$$

$$\checkmark I_T = \frac{\partial M}{\partial T} = -\frac{1}{2} \frac{Ap}{T^{3/2}} = -9.39$$

If you remember the value of M or the best value of M is, what is determined by using the values of A, p and T, which are specified in the problem. Therefore I can calculate this value A is 10, p is 20 divided by square root of 4.84. This was specified in the problem and this gives you a value of 90.91. I have just rounded off to 2 significant figures of the decimal point. This is the value of M, which is expected for the values of A, p and T, given in the problem. Now I have to find the error. To determine the error I have to find out what are influence coefficients.

I will call the influence coefficient as I_A , the influence coefficient for I_A is equal to *doh* M by *doh* A, which is simply given by root of T and that will be 20 divided by square root of 4.84 and this works out to be 90.91. And similarly if I represent the influence coefficient due to p as I_p , it will be *doh* M by *doh* p when I am taking partial derivative all other things are kept constant. So this will be A by root T, that is, 10 divided by square root of 4.84, which is equal to 4.545. So there are 2 influence coefficients. The third one is I_T equal to *doh* M by *doh* T, which will be minus half Ap by T to the power of 3 by 2 and I can substitute and obtain the value as minus 9.39. So the 3 influence coefficients are available here and the values of the errors or the standard deviations of the errors A, p and T are specified as % ages.

(Refer Slide Time 11.02)

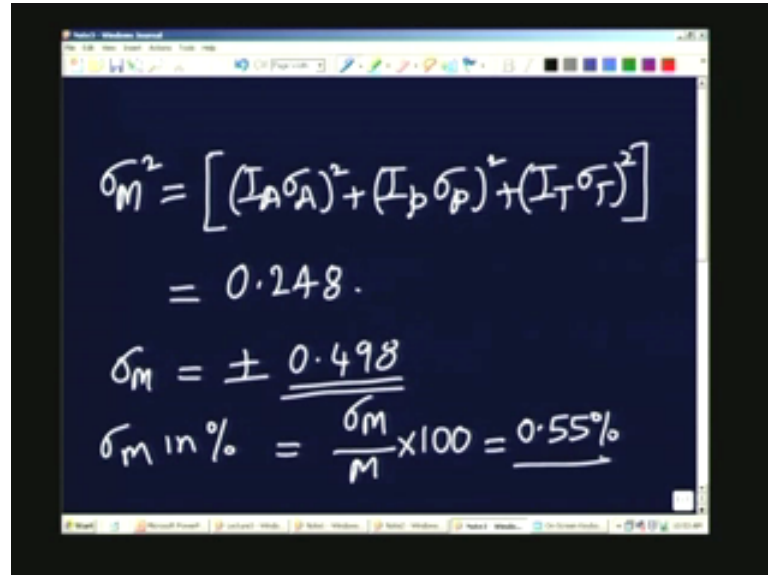


The image shows a digital blackboard with handwritten calculations for three sigma values. The first line is $\sigma_A = 0.1\% = \frac{0.1}{100} \times 10 = 0.01$. The second line is $\sigma_p = 0.2\% = \frac{0.2}{100} \times 20 = 0.04$. The third line is $\sigma_T = 4.84 \times \frac{1}{100} = 0.0484$.

Therefore I will first obtain the actual values. So if you take sigma A is 0.1%, this will be 0.1 divided by 100, multiplied by 10. This is the actual value, so this will give you 0.01. And sigma p is 0.2% that will give you 0.2 divided by 100 into the value of the parameter p, which is 20. That will give you 0.04. And sigma T will be 1%. That means 4.84 into 1 over 100, which is 0.0484.

So once we have obtained all the sigmas and if we go to the previous slide we have obtained the influence coefficients, we can obtain using the error propagation formula. Sigma M squared will be the sum of 3 quantities I_A and sigma A whole squared plus I_p sigma p whole squared plus I_T sigma T whole squared. If you want sigma M, all you have to do is to take square root and put a plus or minus sign. In fact you can substitute the values, which were obtained in the previous slide. And then I can just substitute here, I will leave it as an exercise for you to complete. It will give you 0.248. That is the value of sigma M squared. If you calculate sigma M it will be square root of this, which will be plus or minus 0.498. I can also calculate the %age value; for example, the value of M is given and I have got the value of sigma M. So sigma M in %age will be sigma M given above divided by M, multiplied by 100 and this comes to about 0.55%. So you see that the problem of propagation can be calculated in a methodical fashion as we have done in this particular case.

(Refer Slide Time 12.56)



The image shows a digital blackboard with handwritten mathematical calculations. The first line is the variance formula: $\sigma_M^2 = [(I_A \sigma_A)^2 + (I_p \sigma_p)^2 + (I_T \sigma_T)^2]$. The second line shows the result: $= 0.248$. The third line shows the standard deviation: $\sigma_M = \pm \underline{\underline{0.498}}$. The fourth line shows the percentage error: $\sigma_M \text{ in } \% = \frac{\sigma_M}{M} \times 100 = \underline{\underline{0.55\%}}$. The blackboard has a toolbar at the top with various drawing tools and a taskbar at the bottom.

$$\sigma_M^2 = [(I_A \sigma_A)^2 + (I_p \sigma_p)^2 + (I_T \sigma_T)^2]$$
$$= 0.248.$$
$$\sigma_M = \pm \underline{\underline{0.498}}$$
$$\sigma_M \text{ in } \% = \frac{\sigma_M}{M} \times 100 = \underline{\underline{0.55\%}}$$

Now I will go back to the presentation where we left it. So I will go to the next slide and discuss a different way of doing the same problem. So let me just explain to you what we are trying to do. On the board we have indicated how to work out in detail the several steps involved in the process. I have obtained the influence coefficients; I first obtained the value of the quantity M in this case, which corresponds to Q in the general formula we have worked out earlier. Then we calculated individual sigmas from the %ages, in terms of actual values, we were able to use the propagation formula to find out what is the variance in the value of Q. And of course by taking the square root you can get the value in terms of the standard error or the standard deviation of the error also we refer to as the standard error in Q. If you again divide this by the value of Q and multiply by 100 you get the %age of error which was 0.55. And if you look at this next slide the relationship between the various quantities is given in the form of product of quantities. That means M equal to pA into p divided by square root of T. It involves product of this and therefore I can do what is called logarithmic differentiation and that is shown here.

(Refer Slide Time 14.30)

Example 3 (Continued)

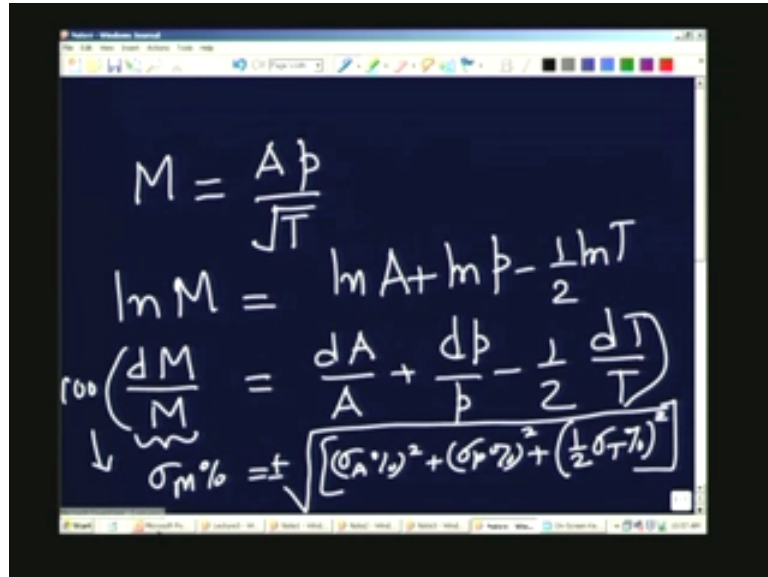
Since the relationship given involves product of quantities on the right hand side logarithmic differentiation will automatically yield the fractional influence coefficients. Thus, we have

$$\frac{dM}{M} = \frac{dA}{A} + \frac{dp}{p} - \frac{1}{2} \frac{dT}{T}.$$

What is logarithmic differentiation?

You take logarithms on both the sides and may be I can indicate it on the board. So in this case we have M equal to $A p$ by root T so I take logarithm of M ; that is, $\ln M$ equal to $\ln A$ plus $\ln p$ minus half $\ln T$: very simple and straight forward. And therefore if I differentiate now, it gives dM by M is equal to dA by A plus dp by p and minus half dT by T . In fact we can look at dM as σM , dA as σA , dp as σp and dT as σT or if you multiply dM by M throughout by 100, this will be the $\sigma M\%$, this will be $\sigma A\%$. While writing this step, I am going to use the error propagation formula. This will become square root of plus or minus $\sigma A\%$ whole squared plus $\sigma p\%$ whole squared plus half of this. Half is coming because the influence of this term is limited to that $\sigma T\%$ whole squared. That is what we are trying to work out.

(Refer Slide Time 17.17)


$$M = \frac{Ap}{\sqrt{T}}$$
$$\ln M = \ln A + \ln p - \frac{1}{2} \ln T$$
$$\text{root} \left(\frac{dM}{M} = \frac{dA}{A} + \frac{dp}{p} - \frac{1}{2} \frac{dT}{T} \right)$$
$$\sigma_M \% = \pm \sqrt{(\sigma_A \%)^2 + (\sigma_p \%)^2 + \left(\frac{1}{2} \sigma_T \%\right)^2}$$

Let me go back to the slide and indicate what has been done and I have simply written down that formula here, after logarithmic differentiation I get this and what I can do is I can go next to the error propagation formula and this is what I indicated on the board: sigma M% is equal to square root of sigma A% whole squared plus sigma p% whole squared plus half of sigma T% whole squared. That minus sign here is not going to make any difference because I am taking the square of the quantity. You need not worry whether the influence coefficient is negative or positive; it is not going to matter. So it will be square root of 0.1 square plus 0.2 square plus half of 1 whole squared, which is 0.55%. In this case because the relationship between the various quantities was in the form of a product, we are able to directly solve the problem using the percentage errors which are given. Therefore it can be used only in such cases. In more general cases, the first method which I used must be used.

(Refer Slide Time 18.33)

Example 3 (Concluded)

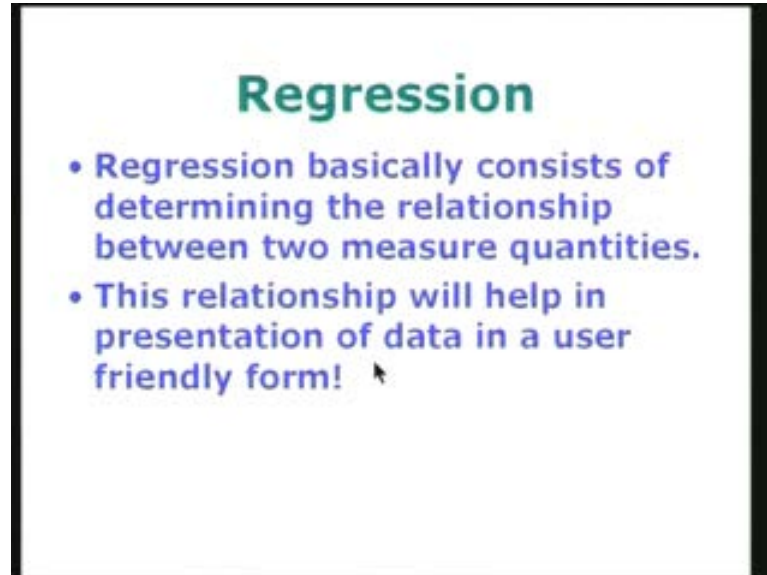
The propagation of error formula will then take the form

$$\begin{aligned}\sigma_M \% &= \sqrt{(\sigma_A \%)^2 + (\sigma_p \%)^2 + \left(\frac{1}{2} \sigma_T \%\right)^2} \\ &= \sqrt{0.1^2 + 0.2^2 + \left(\frac{1}{2} \times 1\right)^2} = 0.55\%\end{aligned}$$

Now with this background, I am going to look at regression. What is regression? How are we going to perform a regression analysis? What is the method to be employed? These are the things which we are going to discuss in this part of the lecture. Regression basically involves or consists of determining the relationship that may exist between two measured quantities, which should have been measured for two measured quantities or regression may be determining the relationship between any number of measured quantities. It may be one quantity, like in the previous case, Q equal to $A p$ by root T is also a regression formula. A , p and T are measured and the value of Q is estimated and the relationship is Q equal to $A p$ by root T . That is a formula and that is also a regression formula for that particular problem. This relationship helps us in presentation of data in a user-friendly form.

Either you can present the data in the form of a table and many graphs and the person who is going to read through your report is going to look at all these and he is going to worry about how to use it or if you are able to make it in the form of a simple relationship like M equal to $A p$ by root T for any other relationship, it is going to be much better for the person. So what I have to do when I am reporting a regression formula is to tell the user what the formula is; that is, the relationship between a derived quantity and a certain measured quantity. Then I must also tell him the range of the values for which this particular formula is valid. Or in the case of an experiment which is performed in the laboratory, you will know the values of the variables which appear in or the measurements you have done, the range of these measurements.

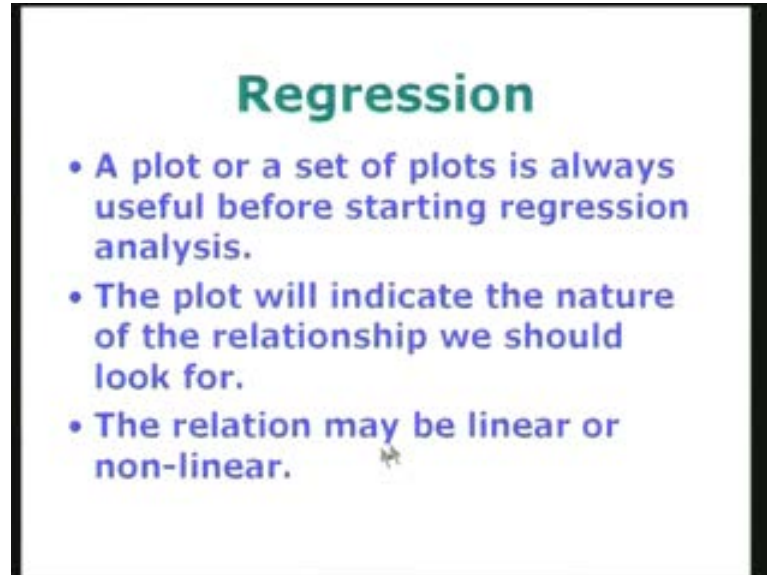
(Refer Slide Time: 20:45)



For example in the case of M equal to A_p by root T , the pressure was given as 20. It may be valid for between 5 and 25. That is the range in which that formula is valid. Then T may have a value of 4.84. Maybe it is valid from 3 to 10. So what I have to do is to give the relationship in the form of a regression equation and also I have to give the values or the ranges of the various things, going into the formula. I also should, ideally speaking, tell the user what is the error I am going to make by using the formula, what is the plus or minus within the band in which it is going to give the value of the function. So Q will be Q plus or minus some error margin and I must be able to give the user the error margin also. So regression analysis consists of all these things.

So the very first step however in doing regression analysis is to make a plot. You either make one plot if it is only 2 variables, Y and X . If there are more than one variable of course one plot will not give all the information so one may have to go for number of plots. So a plot or set of plots is always useful before starting regression analysis. The plot will indicate the type of relationship we should expect between the variables. If it is a single variable Y and X it may be linear or a nonlinear relationship. By making a plot, I will immediately come to the conclusion as to what is the best possible relationship that may exist between the two quantities. The plot will also indicate the nature of relationship we should look for; that is, whether it should be a simple relationship or it may be a complex relationship.

(Refer Slide Time 21.47)



Regression

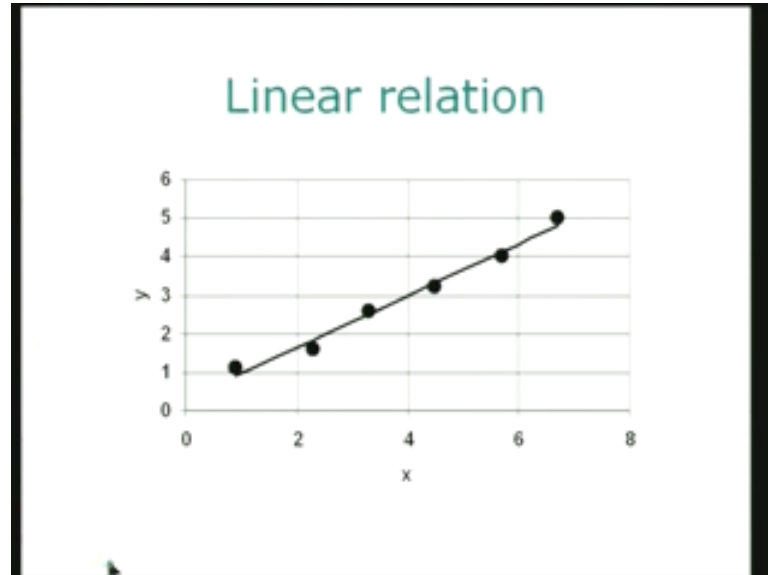
- A plot or a set of plots is always useful before starting regression analysis.
- The plot will indicate the nature of the relationship we should look for.
- The relation may be linear or non-linear.

It may involve several parameters. We will make several things clear as we go along. The relationship may be linear or nonlinear.

I have just taken a few examples. The first example is linear relationship. There is a measured quantity Y and there is another quantity X. These two are either measured or we expect these two to have relationship between themselves. Y and X are supposed to vary systematically with each other. In fact you look for regression or you look for expression only when there is a systematic relationship between the 2 variables which we have measured or 2 quantities which we have measured. If there is no relationship between Y and X and if I make a plot, the plot will not show any definite relationship. The points will be scattered all over the diagram and therefore you can easily come to a conclusion that there is no relationship between these two.

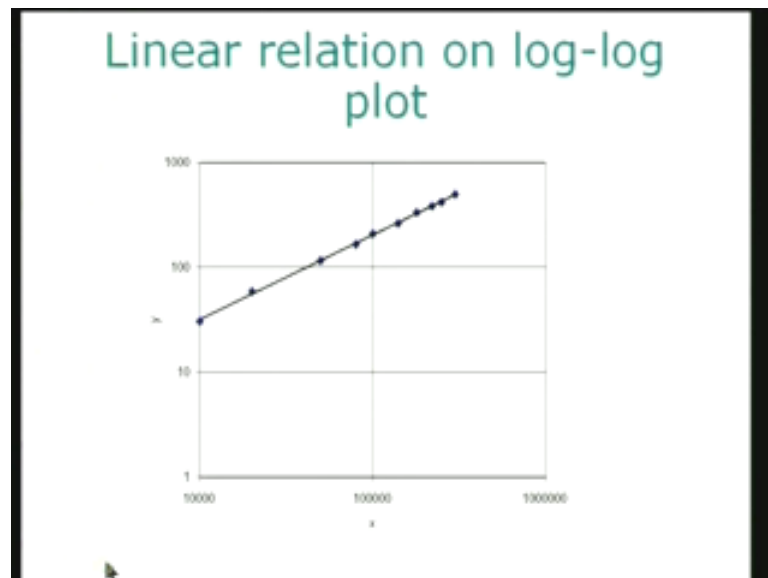
Suppose Y varies and X varies independently, the two are not going to have any relationship and therefore you will not get a systematic relationship as shown here. In this case Y and X are showing some kind of a linear relationship. The data collected are shown by the black circles here and the straight line is what I am expecting. And I will find out how to draw this straight line. That is the basic idea about regression. The regression analysis first of all, determines what is the regression equation or the relationship it should have. Secondly, it tells us what is that relationship and then thirdly, it will also tell us the error incurred in using this relationship instead of the original data.

(Refer Slide Time 22.51)



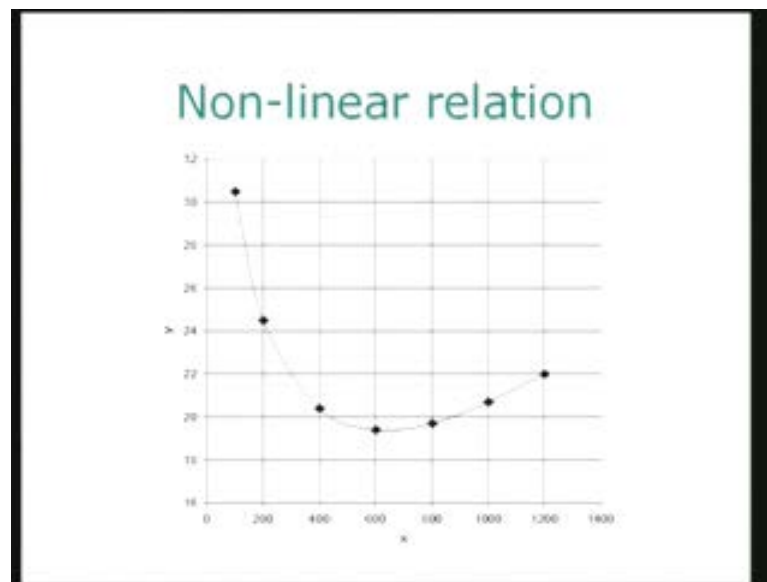
Let us look at the second type of relationship, which you may have. Suppose I make a plot. In this case I have taken log-log plot: log of X here and log of Y here. The logarithm may be either with respect to natural logarithm, with respect to base e or you can take it with respect to 10; it doesn't matter. Suppose you make a plot here on the log-log plot and if it appears as a straight line the original data can be converted into logarithm of the respective quantities and now I can look for a straight line relationship. So what is the relationship that gives a straight line in a log-log sheet? We will see it a little later.

(Refer Slide Time 24.37)



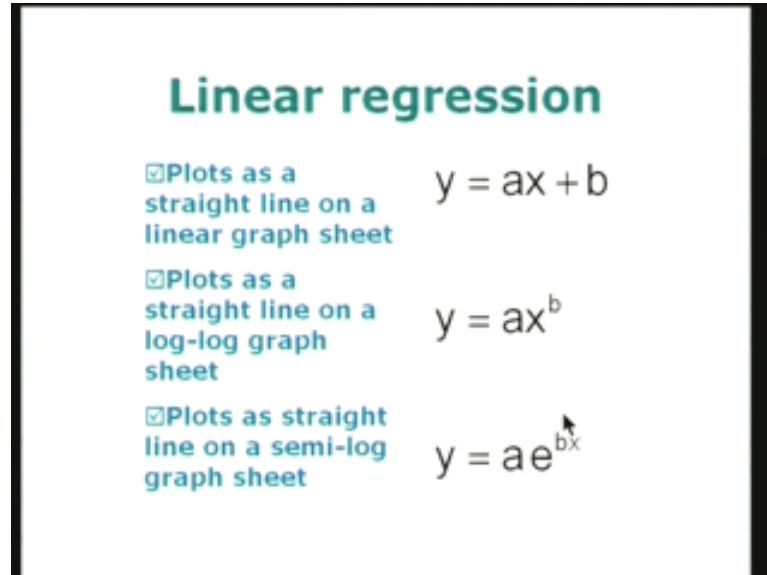
A third kind of thing which can happen is that the Y and X actually related in a nonlinear fashion. In this case you can see X and Y. I cannot linearize this relationship: this is possibly a polynomial may be quadratic or cubic. This cannot be approximated by a straight line. So what I am going to do is to find out what are all the situations where we can have a linear relationship. Regression could be a linear regression model. I have shown 2 examples: the first relationship was between Y and X. Second one was when there was linear relationship between logarithm of Y and logarithm of X. And in fact there is a third possibility when you plot it on a semi-log paper. That means you plot X versus log Y or Y versus log X; then also you may get a straight line.

(Refer Slide Time 25.23)



So these are the three possibilities that exist for linear regression and they are shown in the next slide. So, in the first case, y equal to ax plus b plots as a straight line on a linear graph sheet. Then y equal to ax to the power of b plots as a straight line on a log-log graph sheet because you can take $\log Y$ as equal to $\log a$ plus b . $\log x$ and $\log y$ are linear in this particular thing. So you are going to replace x by $\log x$, y by $\log y$, and I get a linear relationship. And the third one is when you have semi-log plot graph sheet and the plot looks like a straight line. I will have y equal to e^{bx} or it may be 10^{bx} ; I can use both. You can see that $\log y$ equal to $\log a$ plus bx , logarithm of e to the power of bx is simply bx . So $\log y$ is equal to a plus bx and what is common between all these three cases is that there are two parameters a and b . So a straight line or a linear regression gives rise to a two parameter model.

(Refer Slide Time 26.29)

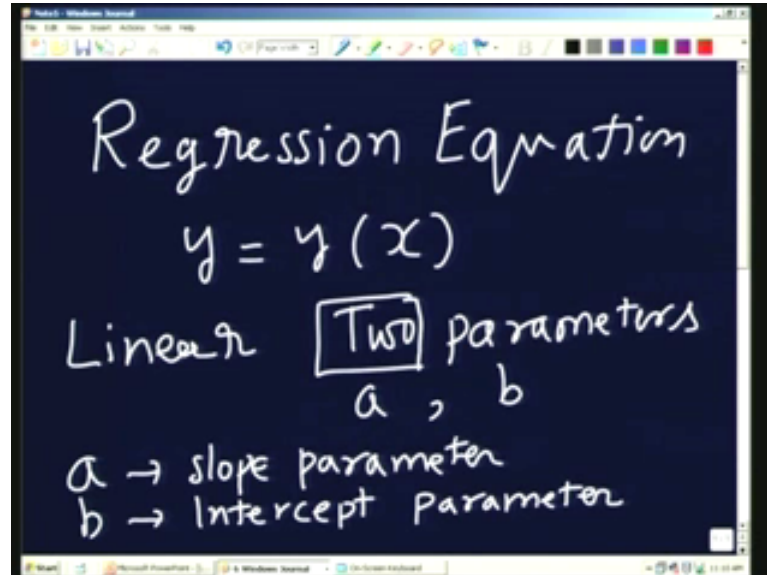


Linear regression

- ☑Plots as a straight line on a linear graph sheet $y = ax + b$
- ☑Plots as a straight line on a log-log graph sheet $y = ax^b$
- ☑Plots as straight line on a semi-log graph sheet $y = ae^{bx}$

Let me just go to the board and explain to you in slightly more detail. So we have a regression equation, which is a relationship; I am talking about only two variables. y is some function $y(x)$. Linear gives rise to 2 parameters a and b : a is the slope parameter. I am calling it slope parameter because a itself may be slope or if I am using log-log, a may become something else, b is an intercept parameter. I am using these terms slope parameter and intercept parameter in a general sense applicable to the three situations shown on the slide: (1) x and y plot linearly on a normal graph sheet; (2) x and y plot as a straight line on a log-log sheet and (3) x and y plot as a straight line on a semi-log sheet. So a and b have different meaning depending on the particular case I am thinking of. The number of parameters is always 2; in all these 3 cases we have 2 parameter models.

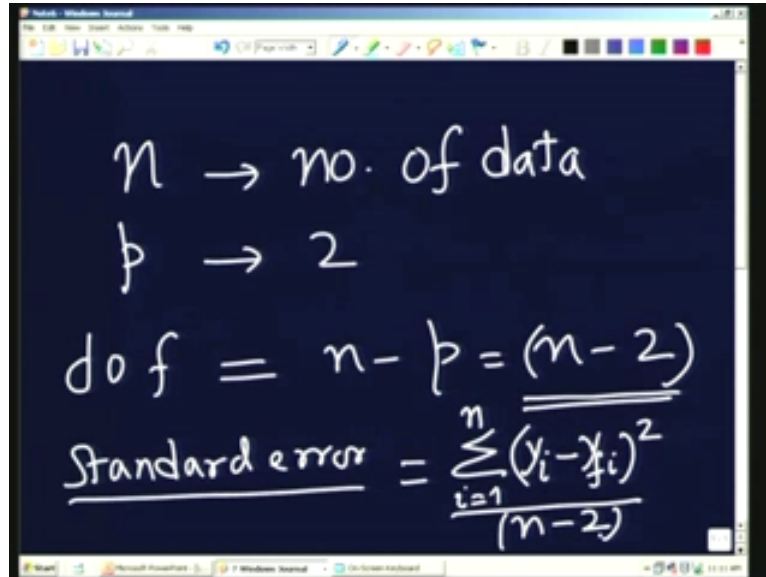
(Refer Slide Time 29.56)



And what does it mean in terms of regression? We are finding two. Suppose the number of data is n , parameters are two so you see the degree of freedom is equal to n minus p , which is equal to n minus 2. So when I want to find out the likely error, I have to use n minus 2 in the denominator. For example the standard error will be given by σ_i equal to 1 to n ; the value of y_i , which is the data given minus y_{fi} . y_f stands for the fit or the regression line, its whole square divided by n minus 2. Normally we would have used n here; its number of degrees of freedom is n minus p .

Now I am going to replace it by n minus 2. Therefore I am calculating both a and b using the data. So as I indicated earlier, 2 bits of information or 2 parameters have been obtained from the same set of data and therefore I have less information available now to calculate the third parameter, that's the significance. With this background, let us go back to the slide and what I am trying to do is I am trying to see these three types of relationship. All of them are going to be linearized. That means I am going to take logarithm on both sides in log-log graph; and in semi-log graph case, I am going to logarithm on only 1 side, semi-log means I am going to plot in a different way.

(Refer Slide Time: 31.17)

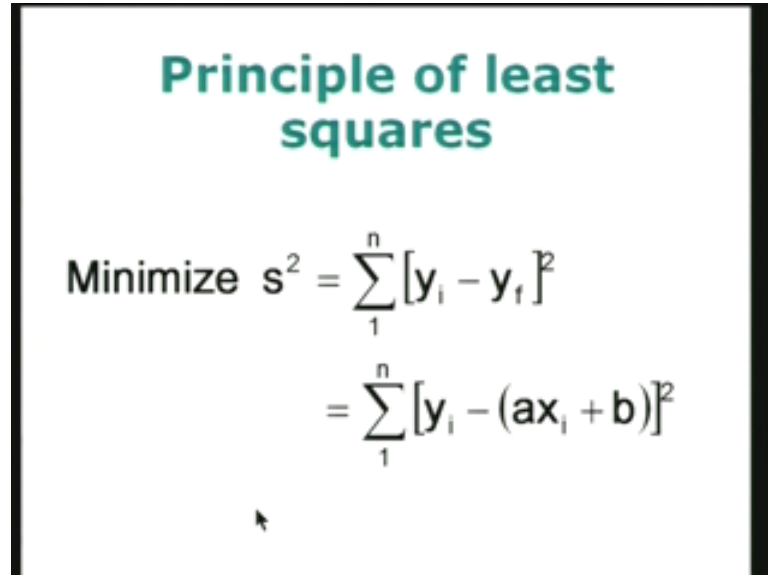


The image shows a digital blackboard with handwritten mathematical formulas. The first line is $n \rightarrow \text{no. of data}$. The second line is $p \rightarrow 2$. The third line is $\text{dof} = n - p = \underline{\underline{(n - 2)}}$. The fourth line is $\underline{\text{Standard error}} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{(n - 2)}$.

So if I do that, the principle of least squares is going to be used for obtaining the relationship which I am looking for. So what is the relationship I am looking for? The general relationship now is a linear model. The relationship between y and x is a linear relationship or a straight line and what I am going to do is I am going to look for a minimum value for the sum of the squares of the errors—the disparity between the values given by the linear relationship and the actual data which I have got as the measured set of data. So let us look at the formula here. I want to minimize s square. s square is the sum of the squares given from sigma 1 to n number of data $y_i - y_f$ the whole square.

Of course y_f is calculated for the corresponding value of x . Therefore it is equal to sigma 1 to n $y_i - y_f$; at the value of x_i will be ax_i plus b whole square. So I have to minimize this value of s square because we have already seen when we are talking about normal distribution of errors and so on that the sum of the squares must have the smallest value. That is when we found out that the mean value represents the best value of the quantity. There we were talking about single quantity we want to measure, but here we are talking about two related values y and x , which are related. And now I am looking for the difference between the data y for a particular value of x and the value calculated using the straight line relationship at the same value of x and the difference between the two. I am taking the square of this and summing over all the data points available from 1 to n and I am trying to minimize this value.

(Refer Slide Time 32.21)



Principle of least squares

Minimize $s^2 = \sum_{i=1}^n [y_i - y_f]^2$

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

So minimization requires that we evaluate 2 derivatives *doh* s square by *doh* a, this is one of the parameters. Second parameter is *doh* s square by *doh* b. Please notice that we are taking the derivative with respect to the parameters x and y. The derivative is taken with respect to the parameters a and b as I have done here: minus of sigma 1 to n, 2 into y_i minus ax_i plus b multiplied by x_i equal to 0, because I am taking derivative with respect to a y_i minus y_f whole square. It will give 2 into y_i minus y_f multiplied by the differentiation of this with respect to a will give x_i with a negative term so I have put a negative here. In fact that negative is not going to play a role. Similarly if I take it with respect to b, I get the second equation: sigma 1 to n, 2 into y_i minus ax_i plus b. With respect to b it gives 1 with a negative sign that is equal to 0. So I have got two conditions for the minimization of the sum of squares.

(Refer Slide Time 34.22)

Conditions for minimum s^2

$$\frac{\partial s^2}{\partial a} = -\sum_1^n 2[y_i - (ax_i + b)]x_i = 0$$

$$\frac{\partial s^2}{\partial b} = -\sum_1^n 2[y_i - (ax_i + b)] = 0$$

And this leads to two equations: what are the two equations? What I am doing is I am rewriting this in the form of an equation. So it will be 2 into y_i into x_i . This 2 is common to all the terms so I can remove the 2. Similarly, here I don't have to keep the 2, negative, I can also remove it. I am taking it to the other side sigma 1 to n a times x_i , a is a constant so I take it outside. So you see that I write it as ca into sigma x_i square plus b into sigma x_i equal to sigma $x_i y_i$. All this summation is over the equation i equal to 1 over n and I have not explicitly shown because it is understood from the symbol here; so this is the first equation. The second equation is obtained from this equation (refer previous slide). The first equation was from here, second equation was from here. You can see that sigma y_i is equal to sigma x_i into a plus n times b because I am adding b again and again n times it will be n times b. If you add the same concept again and again then you are going to get n times b. So what I have is 2 normal equations, which are shown here.

(Refer Slide Time: 35.40)

Normal equations

- The following two (Normal) equations have to be solved for obtaining the fit parameters a and b

$$(\sum x_i^2) a + (\sum x_i) b = \sum x_i y_i$$
$$(\sum x_i) a + n b = \sum y_i$$

The solution to these two equations is indicated in the next slide. We use a Cramer's rule, which is familiar to all of you. The constant a or the parameter a is obtained as the ratio of two determinants: the top one we have $\sum x_i y_i$, $\sum x_i$, $\sum y_i$ and n as the elements. The denominator is $\sum x_i^2$, $\sum x_i$ and n. The denominator is the same for the two parameters: only the numerator is different because we are going to replace the columns and then obtain the two quantities, a and b. What I am going to do is, I am going to use some definitions which are statistical in nature and using these definitions I am going to rewrite the values of a and b by using the statistical parameters.

(Refer Slide Time 37.13)

Solution for a and b

- Use Kramer's rule to get

$$a = \frac{\begin{vmatrix} \sum x_i y_i & \sum x_i \\ \sum y_i & n \end{vmatrix}}{\begin{vmatrix} \sum x_i^2 & \sum x_i \\ \sum x_i & n \end{vmatrix}}, \quad b = \frac{\begin{vmatrix} \sum x_i^2 & \sum x_i y_i \\ \sum x_i & \sum y_i \end{vmatrix}}{\begin{vmatrix} \sum x_i^2 & \sum x_i \\ \sum x_i & n \end{vmatrix}}$$

So this is shown in the next slide. The statistical parameters I am going to use to describe the data are given here. There are two means: \bar{x} is the mean of the values of x_i . So $\sum x_i$ by n where n is the number of data points is the mean x . Similarly, I have the mean y value, which is given by \bar{y} equal to $\sum y_i$ by n . And I have got two variances, one corresponds to the variance with respect to x and the other one variance corresponds with respect to y . So the definition is very simple. We have already given this definition earlier: σ_x^2 or σ_x^2 is equal to $\sum x_i^2$ by n minus \bar{x}^2 . This \bar{x} is already calculated and is given here. Similarly σ_y^2 is the variance of the y with respect to its own mean and $\sum y_i^2$ divided by n minus \bar{y}^2 .

The third quantity which makes its appearance if you go back to the normal equation is the product like $\sum x_i y_i$. This product or the summation of the product of two quantities is called the covariance. The covariance is nothing but an indication between the variability or the influence of y_i on x_i or a x_i on y_i . So the two are going to have some mutual understanding with each other and therefore the variations in the two are related. And therefore the covariance tells me what the relation between x_i and y_i is. In fact we will talk more about it a little later. So $\sum x_i y_i$ is the summation of the products of x and y divided by $n - \bar{x} \bar{y}$, which is from here. So the two quantities a and b , which were given in terms of the ratio of two determinants here and the ratio of the other two determinants here, can be recast in a form that involves the quantities defined here. This $\bar{x} \bar{y} \sigma_x^2 \sigma_y^2$ and σ_{xy} and the intermediate steps can be supplied by you. You can work it out and I will leave it as an exercise for you to do it.

(Refer Slide Time 38.07)

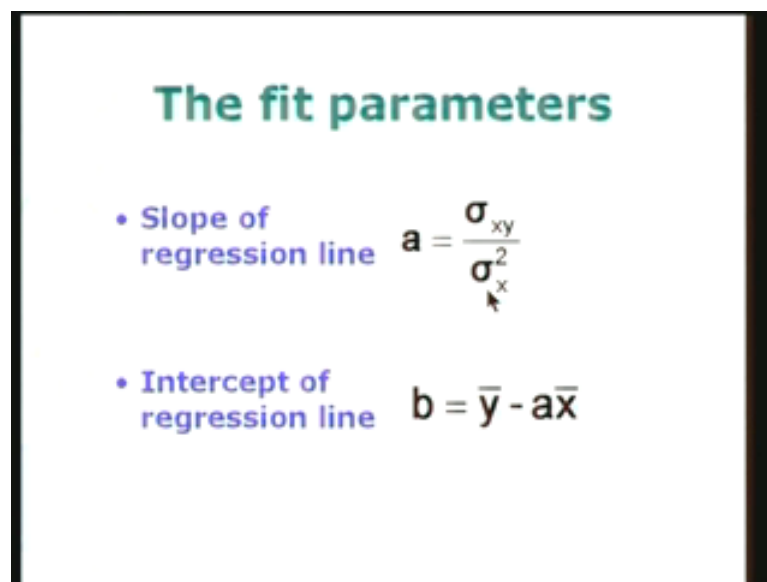
Statistical parameters

- **Means** $\bar{x} = \frac{\sum x_i}{n}, \quad \bar{y} = \frac{\sum y_i}{n},$
- **Variances** $\sigma_x^2 = \frac{\sum x_i^2}{n} - \bar{x}^2, \quad \sigma_y^2 = \frac{\sum y_i^2}{n} - \bar{y}^2$
- **Covariance** $\sigma_{xy} = \frac{\sum x_i y_i}{n} - \bar{x} \bar{y}$

The slope of the regression line a is equal to σ_{xy} , the covariance divided by the variance of the x or σ_x^2 . Similarly, if we go back to the normal equation we can see that $\sum x_i$ is nothing but $n \bar{x}$. $\sum y_i$ is nothing but $n \bar{y}$ and n is here. So if I cancel all the n 's or if I divide this equation by n , I will get $\bar{x} a + b = \bar{y}$. That means y equal to $ax + b$ is a straight line and \bar{y} is equal to $a \bar{x} + b$. That means the point $\bar{x} \bar{y}$ is the

point which the regression line passes through the point \bar{x} \bar{y} . So the intercept can be obtained by noticing, as we did just now, that \bar{y} and \bar{x} is a point on the straight line or the regression line. So we can say that b equal to \bar{y} minus $a \bar{x}$ and a is already obtained here at the slope and is given by σ_{xy} divided by σ_x^2 . So the 2 fit parameters a and b , have been obtained by a simple procedure, which involves the statistical parameter σ_{xy} , which is called the covariance. The statistical parameter variance of x , which is called σ_x^2 , and the mean values of x and y are here. With these 4: 1, 2, 3, 4 statistical parameters, we are able to get the regression coefficients. So we now have a way of calculating the parameters a and b , the slope parameter and intercept parameter. And now we would like to ask the question as to whether this particular straight line we have found represents the data well or not.

(Refer Slide Time 40.35)



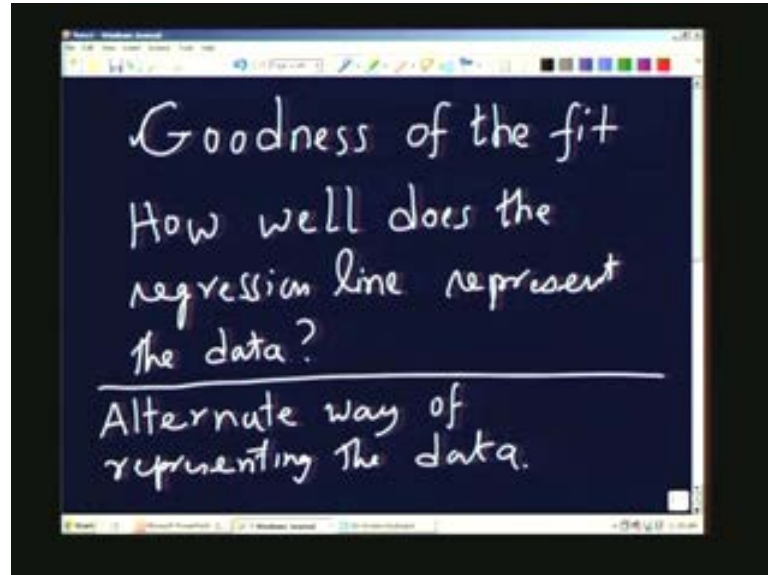
The slide is titled "The fit parameters" in green text. It contains two bullet points in blue text, each followed by a mathematical formula. The first bullet point is "Slope of regression line" followed by the formula $a = \frac{\sigma_{xy}}{\sigma_x^2}$. The second bullet point is "Intercept of regression line" followed by the formula $b = \bar{y} - a\bar{x}$.

The fit parameters

- Slope of regression line $a = \frac{\sigma_{xy}}{\sigma_x^2}$
- Intercept of regression line $b = \bar{y} - a\bar{x}$

So the question we are going to ask is about the quality of the fit. It is also referred to as the goodness of the fit. The question is how well does the regression line represents the data. So in order to do that, I must look for some internal consistency or some parameter which is internal to this procedure and which is going to indicate to me whether the fit is good or not. So for this, what I am going to do is I am going to look at an alternate way of representing the data. What is this alternate way? Let us look at that.

(Refer Slide Time 44.00)



Alternate way of representing the data is as follows: I took y equal to ax plus b ; I will call this the first line. I could have also taken x equal to some a dash y plus b dash; I will call this second line. So what I did earlier was to assume that y and x were linearly related in the form y equal to ax plus b , where a was the slope of that line and b was the intercept. Instead of that, why not take x as a function of y in the form x equal to a prime y plus b prime? Of course I can represent this slightly differently by solving for y , this y equal to, I will retain it here, send this b prime to the other side— 1 over a prime times x minus b prime divided by a prime. So I can say this is something like Ax plus B , where A is this quantity 1 over a dash, this is A and this minus b dash by a dash is your B . So what I am saying is that it is possible to relate the x and y in a different form by writing x equal to a prime y plus b prime, which is equivalent to writing in the form y equal to capital Ax plus B , where A is 1 over a prime.

We will just write it here. A equal to 1 over a prime and this B is minus b prime by a prime. That means if I perform the regression using this line, after performing this regression I will obtain a prime and b prime, exactly similar to the method which we used in the case of the first line and then I will replace it in this form by representing it as a straight line y equal to some Ax plus B and the slope of that line is 1 over a prime. The intercept of the line is minus b prime by a prime. So the method used for a prime and b prime is exactly similar to a and b , so we can easily obtain capital A and capital B as a slope of a second line and intercept of a second line.

(Refer Slide Time 47.00)

The image shows a digital chalkboard with handwritten mathematical equations. At the top, it says $y = ax + b$ - 1st line. Below that, it says $x = a'y + b'$ - 2nd line, which is enclosed in a rectangular box. A curved arrow points from the boxed equation down to the next line. The next line is $y = \left(\frac{1}{a'}\right)x - \left(\frac{b'}{a'}\right)$. Below this, it says $= \underline{A}x + B$. To the right of the boxed equation, there are two equations: $A = \frac{1}{a'}$ and $B = -\frac{b'}{a'}$. The $\frac{1}{a'}$ and $-\frac{b'}{a'}$ terms in the main equation are circled, and arrows point from these circles to the definitions of A and B respectively.

$$y = ax + b \text{ - 1st line}$$
$$x = a'y + b' \text{ - 2nd line}$$
$$\rightarrow y = \left(\frac{1}{a'}\right)x - \left(\frac{b'}{a'}\right)$$
$$= \underline{A}x + B$$
$$A = \frac{1}{a'}$$
$$B = -\frac{b'}{a'}$$

Suppose I compare the slopes of the two lines. If we compare the slopes of the two lines and if the slopes of the two lines are close to each other, we say that the correlation is good. If the two slopes calculated are different, obviously the fit is not good. We can only say whether the fit is good or bad by comparing the slopes of the two regression lines and in fact I can take the ratio of these two slopes and relate them to the goodness of fit. I will say that if I calculate, the ratios of these slopes come close to 1, it means that the regression is good. If it comes very small then the regression is not good. So ratio of slopes is a measure of the goodness of fit.

(Refer Slide Time 48.30)

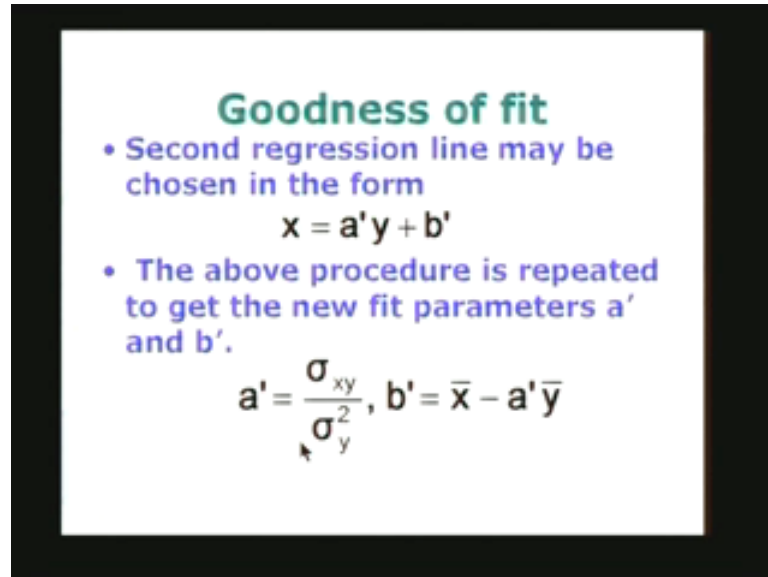
The image shows a digital chalkboard with handwritten text. The first line says "Compare the slopes of the two lines." The second line says "Ratio of the slopes is a measure of the goodness of the fit".

Compare the slopes of the two lines.

Ratio of the slopes is a measure of the goodness of the fit

So with this background, what I will do is I will show you how the second line is obtained and we can look at the two slopes and introduce a parameter, which is going to describe whether the correlation is good or bad. So I have indicated here the second regression line just as I showed it on the board x equal to a' y plus b' and now you remember what we did for obtaining a and b . A similar procedure gives you a' equal to σ_{xy} divided by σ_y^2 b' equal to \bar{x} minus a' \bar{y} . So b' and a' are obtained by statistical parameters.

(Refer Slide Time 48.46)



Goodness of fit

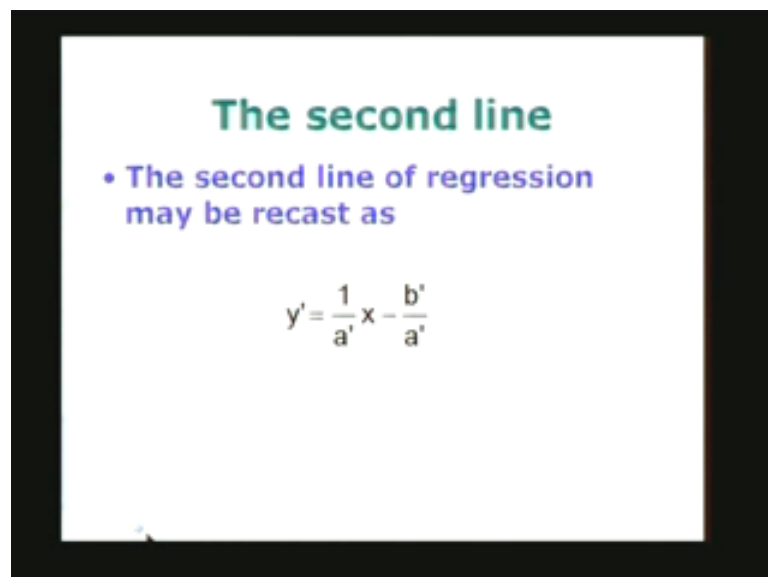
- Second regression line may be chosen in the form

$$x = a'y + b'$$
- The above procedure is repeated to get the new fit parameters a' and b' .

$$a' = \frac{\sigma_{xy}}{\sigma_y^2}, b' = \bar{x} - a'\bar{y}$$

And then the second line is given by y' equal to $1/a'$ x minus b'/a' .

(Refer Slide Time 49.12)

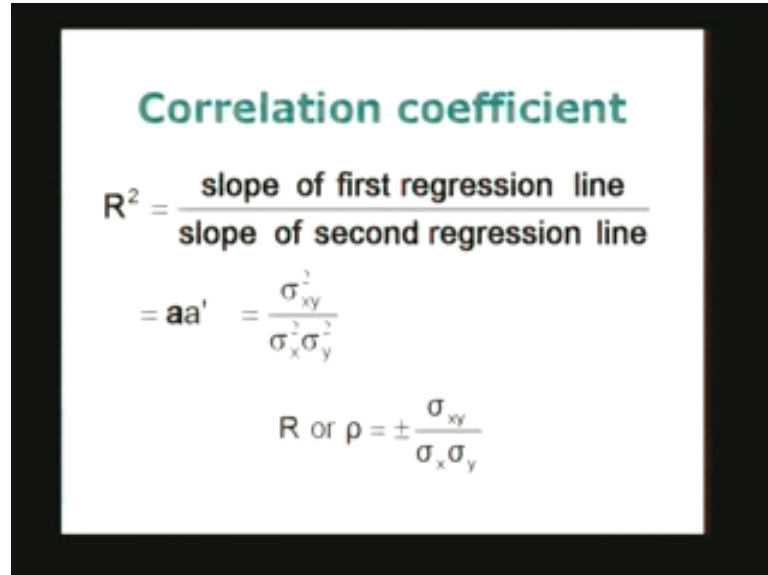


The second line

- The second line of regression may be recast as

$$y' = \frac{1}{a'}x - \frac{b'}{a'}$$

(Refer Slide Time 49.25)



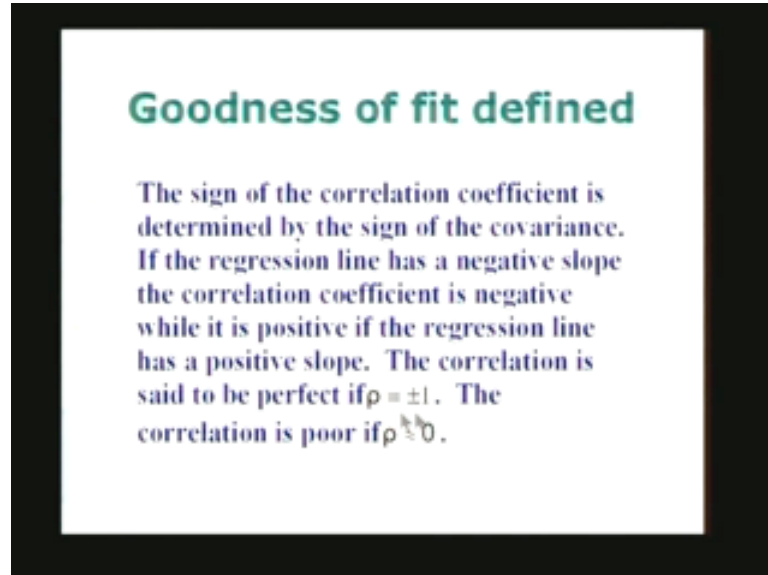
Correlation coefficient

$$R^2 = \frac{\text{slope of first regression line}}{\text{slope of second regression line}}$$
$$= aa' = \frac{\sigma_{xy}^2}{\sigma_x^2 \sigma_y^2}$$
$$R \text{ or } \rho = \pm \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

And therefore you can see that the ratio of these two slopes, slope of the first regression line divided by the slope of the second regression line. This will be nothing but a product of a and a prime because 1 over a prime was the slope of the second line, a was the slope of the first line. Therefore I have to take the product of these two. That will give you sigma xy square divided by sigma x square sigma y square. And therefore we call this the R square or the correlation coefficient or we also call it coefficient of correlation. R is the coefficient of correlation we refer to this as R square parameter and that is given by sigma xy whole square divided by sigma x square sigma y square or R.

Sometimes we use the symbol rho equal to plus or minus sigma xy by sigma x sigma y. This plus or minus will come into the picture because if you look at the numerator, numerator is the product of x and y; we have defined sigma xy earlier. The covariance is nothing but sigma x_i y_i divided by n minus x bar y bar. This may be either negative or positive. It will be negative if as x increases y decreases. If x and y increase at the same time, they behave alike, then the slope of the covariance will be positive. So the regression may give either a negative correlation or positive correlation negative 1 minus 1 will be equally good plus 1 is equally good. If the value is far away from either plus 1 or minus 1, we can say that the correlation is not good and the relationship is not a good relationship.

(Refer Slide Time 51.22)



So I have explained it here. The sign of the correlation coefficient is of course determined by the sign of the covariance and the correlation is said to be perfect if rho is equal to plus or minus 1. The correlation is very poor if rho is very close to 0 and in fact in practice, we should get the value of rho greater than 0.5 or less than minus 0.5. It should not lie between 0.5 and minus 0.5. If it is anything more than 0.5, there is some likelihood of correlation. Of course, the closer it is to plus 1 or closer it is to minus 1, the correlation is going to be the best we think of.

So with this, let us take an example that indicates how this procedure is to be done. So I am taking a data which is very simple, which is expected to show a behavior of the form y equal to ax plus b and I want to determine fit parameters by linear regression the procedure we have just given. x and y values are given in the tabular form. The first row gives the x values 0.9, 2.3, 3.3, 4.5, 5.7 and 6.7 and the y values are given in the second row 1.1, 1.6, 2.6, 3.2, 4 and 5 and we can immediately see that x is increasing and y is also increasing. Therefore we expect the covariance to be positive. That means the slope is going to be positive. y is equal to ax plus b , a is a positive quantity.

(Refer Slide Time 52.40)

Example 4

The following data is expected to follow a relation of the form $y = ax + b$. Determine the fit parameters by linear regression.

x	0.9	2.3	3.3	4.5	5.7	6.7
y	1.1	1.6	2.6	3.2	4	5

So in order to do this regression, it is necessary to put the whole thing in the form of a table so that all the statistical parameters can be obtained. So the data numbers 1 to 6, the x values, are given in the first column, y values in the second column. If you remember the statistical parameters required the calculation of sigma x square, sigma y square and sigma xy. Therefore, x square, y square and xy are calculated and put in the tabular form. Then I sum all the columns up to the number of data we have. So 23.4 is sigma x, sigma y is 17.5, and sigma x square y square and sigma xy. So you can see that the column sums are given and the column means can be calculated by simply taking this number and dividing it by the number of data points; that's also done here. So when we have got the column sum and column mean, I can calculate sigma x square sigma y square. This will be nothing but the column mean. Here it is 19.1033 minus the square of the column mean, here 3.9 square for this case and similarly sigma y square. And in fact I can calculate slope of the fit. It is covariance divided by sigma x square. Covariance is calculated by using this quantity, 13.9917 minus x bar y bar and then dividing it by sigma x square and you will see that it is 0.6721 and the intercept is 0.2955.

(Refer Slide Time 53.20)

Tabulation

Data No.	x	y	x^2	y^2	xy
1	0.9	1.1	0.8100	1.2100	0.9900
2	2.3	1.6	5.2900	2.5600	3.6800
3	3.3	2.6	10.8900	6.7600	8.5800
4	4.5	3.2	20.2500	10.2400	14.4000
5	5.7	4	32.4900	16.0000	22.8000
6	6.7	5	44.8900	25.0000	33.5000
Column Sum:	23.4	17.5	114.6200	61.7700	83.9500
Column Mean	3.9	2.9167	19.1033	10.2950	13.9917
σ_x^2	3.8933	Slope of the fit line is: $a =$			0.6721
σ_y^2	1.7881	The intercept is: $b =$			0.2955

So now I can compare the fit of the regression line with the data given and find out whether the fit is good or bad. So for the first one I can make the table like this: x and y are the given data and this is the fit line which is calculated using the regression analysis we have done earlier. You see that these values are close to the values of y . y_f is in the third column and close to the values, and the fit appears to be good. And in order to confirm further I can calculate the correlation coefficient by using the formula which we gave earlier: 2.6167 divided by square root of σ_x square σ_y square gives you 0.992. The correlation coefficient is 0.992—very close to 1 and therefore we expect the fit to be very good.

(Refer Slide Time 55.09)

Fit vs. Data

x	y	y_f
0.9	1.1	0.9
2.3	1.6	1.8
3.3	2.6	2.5
4.5	3.2	3.3
5.7	4	4.1
6.7	5	4.8

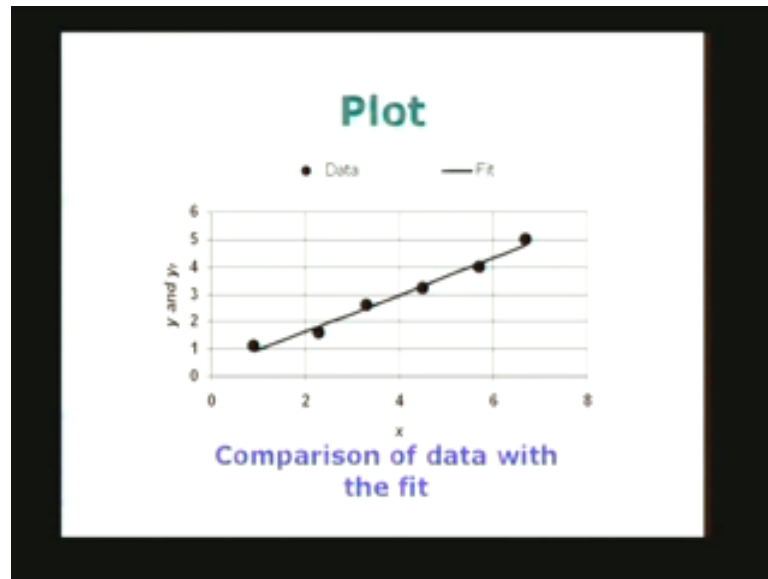
$$\rho = \frac{2.6167}{\sqrt{3.8933 \times 1.7811}}$$

$$= 0.992$$

Correlation coefficient

And in fact I can show the comparison also in the form of a plot. The plot also indicates that the dark circles here and the line are close to each other and therefore we expect the comparison to be good or the correlation to be good.

(Refer Slide Time 55.50)



We will stop this lecture here. In the next lecture, we will continue with some more discussion regarding both linear and nonlinear fit, take 1 or 2 examples and move on to the next topic. Thank you.