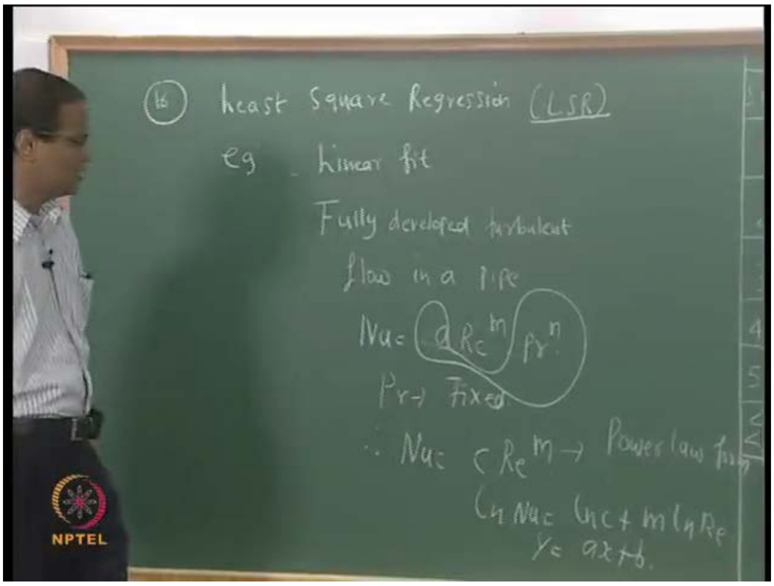
Design and Optimization of Energy Systems
Prof. C. Balaji
Department of Mechanical Engineering
Indian Institute of Technology, Madras

Module No. # 01
Lecture No. # 17
Least Square Regression

We will continue with the discussion on LSR - Least Square Regression. Towards the end of yesterday's class, we looked at the classical problem of fully developed turbulent flow in a pipe, where the heat transfer, the dimensionless heat transfer which is given by the Nusselt number can be written as a Reynolds number to the power of m and Prandtl number to the power of n.

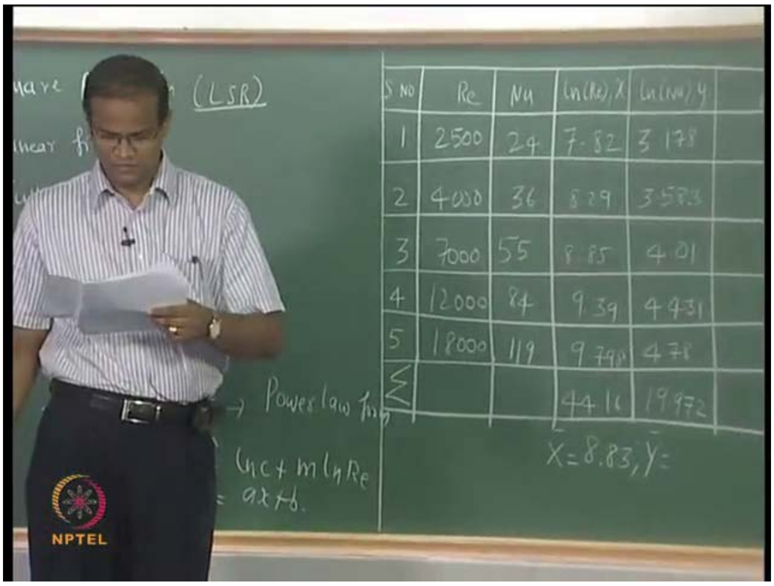
(Refer Slide Time: 00:41)



Suppose, you are considering experiments over water or any other fluid, if the fluid is fixed, the Prandtl number is fixed. So, a into Prandtl number to the power of n gets fixed for the problem under question. So, you can put it as a c Reynolds to the power of m, but so far we have learnt only linear regression that is y equal to a x plus b. So, this is basically what is called as a power law form; something goes as Reynolds number to the power of something; so, it is possible to take natural log on both sides and reduce it to the form y is equal a x plus b. And I gave some typical data, I gave results of a hypothetical experiments which we conducted by varying the velocity and then determining the heat transfer rate, and then from the heat transfer rate, we get the heat transfer coefficient; then we convert it in to Nusselt number using thermal conductivity, and all those stuff. And then, I gave five representative values. Reynolds numbers varies from 2500 to 18000; the Nusselt number Nusselt number is appropriately varying.

I started with Reynolds number 2500 because Reynolds number 2300 is the limit for transition to transition turbulence for a pipe flow, right. Now let us do, let us work out this; I will give you ten minutes; I will start filling it out; do not look at the board; you do it yourself and independently check whether you are also getting the same thing, right.

(Refer Slide Time: 01:55)



So, the log will be… is it ok? Just check this; is it alright? What is the sigma of, what is sigma x? Good 44.1. What is x bar? I do not know whether the x bar is required or not; we will just calculate and keep.

Student: 8.83.

8.83. Keep it. Of course, this can be done using Excel. You can use it, you do it on various software, write a program and all that, but this is the first time you are doing regression. If it is the first time you are doing regression, it is important to do a hand calculation. Hand calculation is the best way. Even when we teach finite element to students, we ask them to take two elements and do hand calculation first, and you should know what is going on in the computer.

So, if you look at this, 55 is 4.01 is it? 0 0 okay 01. You are able to see Vikram or you want to come here? Sum is 19.97, alright? Please check; please do not copy from the board. I am not solving a very tough partial differential equation which you cannot follow; I am just doing simple math. Senthil what is the mean, y bar? 3.9944 right?

3.9944. Now, the more tedious operation has come. You have to start; get x square; 7.82 square; this is 68.72, 88.17, 96.00; so, the x square is 392.36. From this side, Gimin, Vikram, tell me, are you getting these values? You are able to see the board? Akshay, what are you doing? SMS?

Student: Minus 3.1

You can discuss at 9:50; what exam was that? This was the post ponded exam.

What was that?

Student: Linear algebra

Linear algebra? What did you learn in that? What are you learning?

Matrix inversion

Student: Actually for mechanical course fundamentals of engineering drawing we are studying in a, but Stiffness matrix and all that

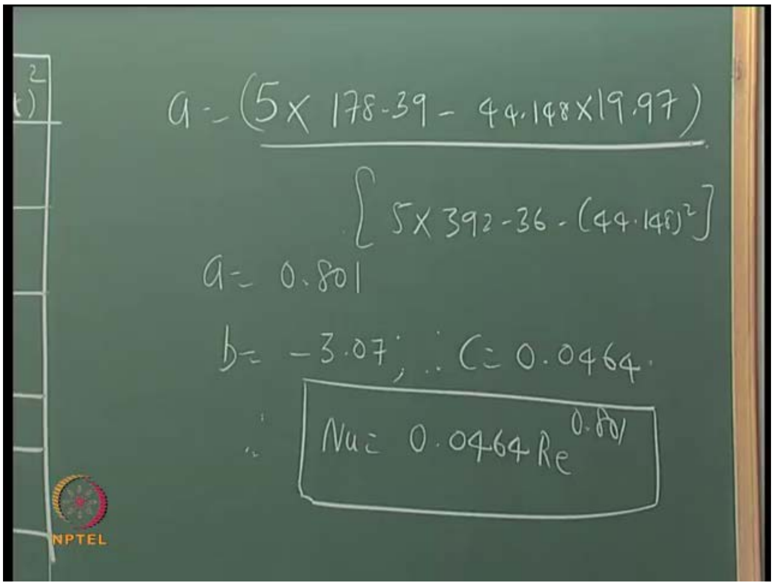
Divergence of stress tensor; vector itself is difficult, then second order tense are there; many second order tensors. There is a book on continuum mechanics by Prager, but of course, as engineers, if you want to do advance studies in mechanics, you need tensors; basically, as a notation it helps us; the notation scares people; that is the thing.

So, xy, patience is required if you want to do linear, if you want to do regression all that, you cannot make mistake; do it slowly; it is not like a puzzle or whatever; it is most mostly a test of a rigour, whether without mistake you can do; it is not very hi funda and all that; if you are systematic, it very simple and straight forward, right.

So, what is the… we require x y 178.39. I think we now have enough. Now, we have all the quantities required to calculate a and b. Please go ahead and calculate a and b, and declare that that is the best fit that represents the Nusselt number as a function of the

Reynolds number. So, y bar was…, I am simply writing that y bar in all the columns; so, a is equal to…

(Refer Slide Time: 08:49)



What is 6 to the power of 0.4?

Student: 2.04

6 to the power of 0.4 is 2.04. So, what I have essentially done is I have I have conducted some hypothetical experiments; I started out with the relation which I am already aware of namely, the details bolder equation. So, it is 0.023 Reynolds number to the power of 0.8 into Prandtl number to the power of 0.4; I assume that the fluid is water. So, 6 to the power of 0.4 is 2; 0.023 get multiplied by 2; I am getting 0.04604.

So, I start with 0.04604 Reynolds number to the power of 0.8 and I will just write this value, and then I will smear these values with some error. This is how we Prof generate problem. We do not take from text books. So, it is a pretty straight forward; it is also it is also physical no, you know that if you conduct a heat transfer experiment, if it is carefully done, you will get the result like this; you will get the result that is worth knowing.

So, I put two more columns; why did you do that? To find the goodness of the fit or the absence of fit, whatever, right. So, now I want to do y fit; I want to calculate y fit; in order to calculate y fit, what I will do is for each data point, I will substitute the value of

Reynolds number, raise it to the power of 0.801, multiply by 0.0464 and write out this column; can you do that? So, this Reynolds number to the power of 0.8 is very typical of turbulent flow; it is wrong. I do not get you.

Student: In that equation, we have a and b, directly we can put y equal to a x plus b.

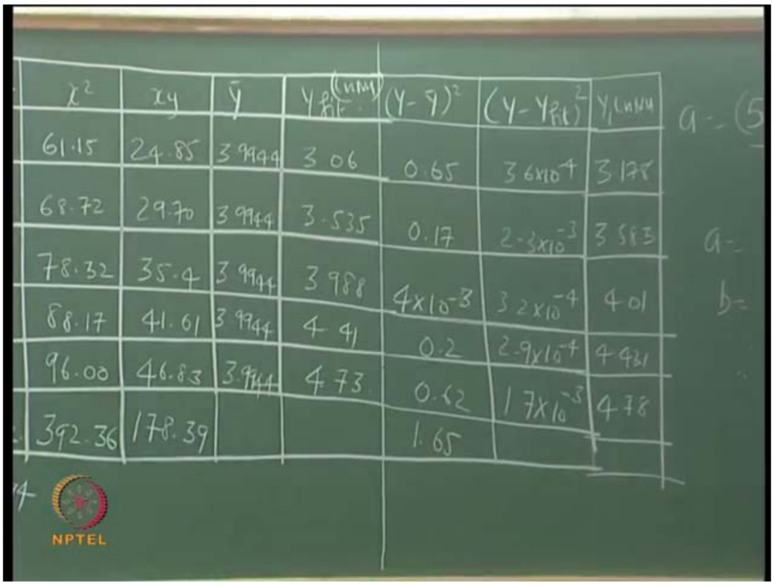
No, there are two ways now; whether you want to check on directly on this equation or you want to directly check we want to check on this; there is a dilemma there; where do you want to check? If you want to check out this , then I will erase this; are you getting the point? So, there is confusion as to whether we have to check on the lon or the actual performance on this. So, if the class feels that we should not check on this oln and then because… why are you giving a blank face? The debate starts; whether we should be checking on the logarithmic values or ultimately on the Nusselt numbers itself.

So, in my notes, I have done it based on the Reynolds, based on this. In the class, I wanted to do this, but whatever you want to do, we will do now; you have to decide. Whatever on the regression, then we will do this. Now, let us complete the other two parts. I hope others got the point. Since we, since it was a Reynolds, it was c Reynolds number to the power of m, we did some ln and reduced it; whether we should test it on the ln of something or on the actual this thing; that is the doubt, somebody can check it out whether there is some difference between the two. So, y fit please get it, the y fit, Deepak can you tell me?

No, now we say that what will be the regressed values of y according to that equation. Now, this y is this ln of n u; so, ln n u. take 0.06446 the Reynolds number to the power of 0.10, then take ln of that; what are you getting? You will get values 2 3 4 5 6 something like that, what is the first value?

3.06. Next?

(Refer Slide Time: 14:26)



Student: 3.535

Student: 3.989

Student: 4.41

Student: 4.73

So, that is fine. Now, there is a pain; so, this also we need a log of this to calculate this. y minus y bar whole square is

Student: 0.65.

Student: 0.17

4 into 10 power 6

Student: 0.62

Sigma?

Why are we finding y minus y bar square? Good question. I am going to the answer all these questions.

Student: 1.65

1.65 yeah, do the y minus y fit; I think it will help if you have one more column; what is that?

Y. Deepak just give me the values ln 24, you already have no?

Yeah, this is just to help you.

Student: 3.6 10 power of minus 4

Student: 2.3 10 power minus 3

Student: 3.2 10 power minus 4

Student: 2.9 10 power minus 4

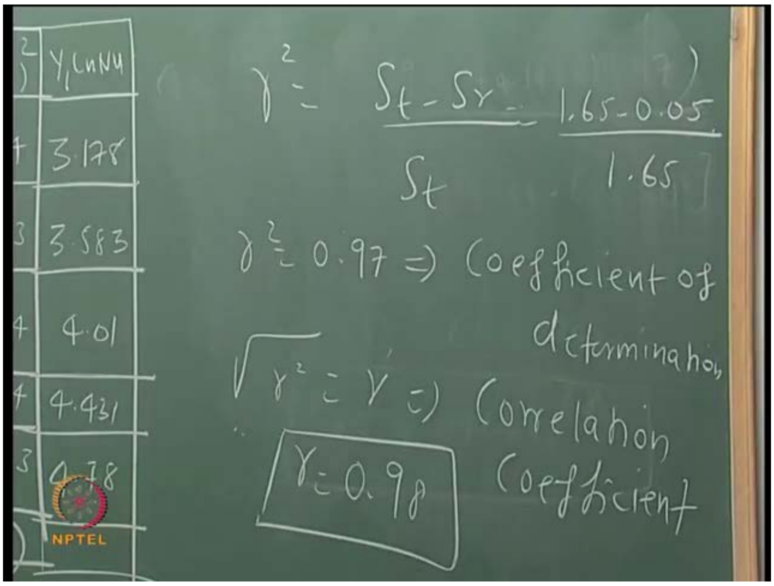
Student: 1.7 10 power minus 3

Sigma, yeah in the exam also you have to draw so many rows, but you guys will finish in 15 minutes. It will take about, to calculate everything it will take about 20 to 30 minutes if you know the procedure; is it okay? Vikram, sigma is how much?

Student: 0.05

0.05, alright. Is everybody through with this? Sowmiya you got it? So, I call this as S t total I mean s is the sum of the residue, the sum of the residue with respect to the mean and sum of the residue with respect to the fit is S r.

(Refer Slide Time: 18:48)



Now, I can define, I can introduce a term called r squared equal to S t minus S r divided by S t; calculate that; this r square is called the coefficient of determination; the root of r square which is r is called the correlation coefficient. You get hope seemly high value; what is it? 98 that is all, am I endlessly introducing more and more quantities and what is all this, what is going on? There may be genuine concern. Now, if you are done with this then I will I will put the [FL] now. Is everybody through with this?

So, we are able to get the values of a and b; we are able to find out y minus y bar y minus y fit; we are able to determine some statistical quantities and something is close to one; so, you believe that is good and all that. Now, I am going to tell you what is what is going on behind the whole thing.

How was this? Please put down your pens and if you have completed, just listen to this. Now, we did some heat transformer experiments; we varied the Reynolds number and got the Nusselt number, and we believed that the Nusselt number and the Reynolds number are related by a power law form like this and then we went ahead and did the and did a regression or least square regression, and finally, we got some values. Has it really helped or not? We require some statistical measures. Suppose I did not have any heat transfer knowledge, what I had was only the Reynolds number versus the Nusselt number; so, what I will do at this? What I will do? The first step is basically look at the

Nusselt number y and then get the mean; then, I will tell the outside world, I do not know any functional relationship, but the mean value is likely to be like that.

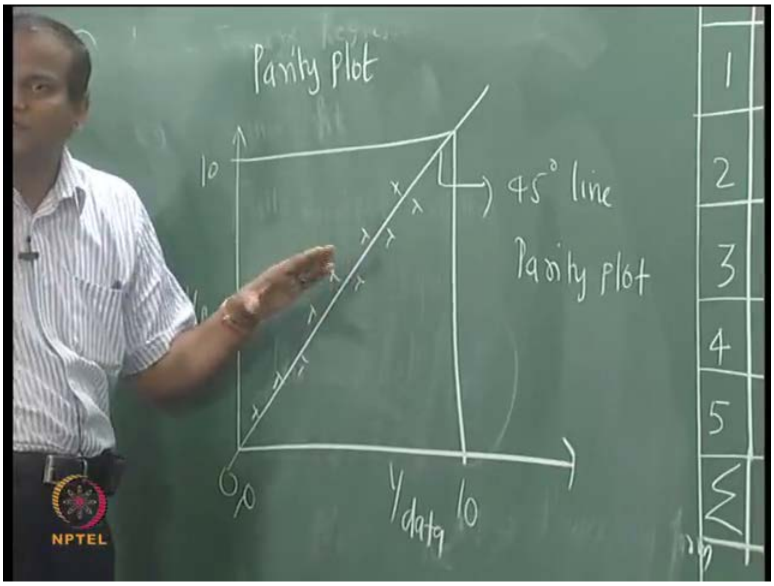
You can get a mean of, in this case, if you take ln of this, what is the mean? You are likely to get a ln of the Nusselt number as 3.9944, but do you stop with the mean of the mean of the ln of the Nusselt number? You do not stop with that. You say that if you have; if I say that the mean is like this, the mean is expected to expected to have a variance which is like this; every data point will have y minus y bar which is like this. Therefore, the total residue is, the total residue is which is given by the difference between the log of the Nusselt number minus log of the Nusselt number average; if you sum all that it will come to 1.65.

However, if I get smart and I say that instead of saying that I will qualify only by the mean, but I put y is equal to a x plus b, as it help, now I am saying that with respect to the fit, I am able to reduce the S t from 1.65 to 0.05. So, out of 1.65, 1.60 of the 1.65 is already explained by the fact that, Nusselt number goes as a Reynolds number to the power of m. Therefore, 97 percent of the variance in the data is explained by this correlation. Therefore, it is a good correlation. So, that is a story.

In the absence of a correlation, you will report only the mean and the standard deviation, but apart from the mean and standard deviation, you go deeper and then find out some physics behind this, and then propose a line and you are able to explain much of the variants in the data with this, then and your r squared is very high; that means, you have a good correlation to work with. Are you getting this point? So, that is the whole thing about the coefficient of determination and the correlation coefficient. Yeah, so that plus or minus .98 could be some books will call minus 1; it is also alright; that is, it is when x is increasing, y is decreasing; it is basically like that.

No, that is not the… here we say that 97 percent of the variants, 97 percent of the variants in the data is explained by this, explained by this correlation. Suppose, you have a correlation which gives only 60 percent, either your experiments are erroneous or there are additional variables which you have not taken into, that you have not taken into account. So, when you do a regression, it will it will tell you so many things; it is just not your experiment or the simulation; it will help you understand the subject. When we do a regression, it well tell you so many things; which are the point which required to be repeated and so on. So, that minus is basically sometimes it when x is increases, y is decreases, and so on. Apart from this we also have, we also can draw what is called a parity plot.

(Refer Slide Time: 24:38)



So, if you have something like this, so, let us say this is 10 and 10 or something; you should have points likes this; this is called the 45 degree line. When you do, when you write your project reports or when you write, when you do, when you do when you write a research article, there is also graphically there is also another way of graphically depicting the relationship between the y data and the y fit using this way.

So, the points must be equally speared on both side to the 45 degree line; this 45 degree line is called the, this is called the parity line. This whole thing is called the parity plot; this called the parity line. If all the points lie on this parity line, it is incredible; that means you doctored data. If you do any experiment, you will have a natural variation because we will have measure voltage, you will measure flow rate, you will measure temperature and so on, there will be natural, there will be natural variability.

So, point should be on approximately 50 percent of the point should be above the line, 50 percent of the point should be below the line, and it should not be smeared; that is there is lot of scattered here means, there is some higher order physics. That is when the Nusselt number is increasing, when the Nusselt number is increasing when I get more error, there are some second order effects; there are addition variable like effect of variation of viscosity or some other parameter which you have not taken into account; that is very very important. So, you can draw ten percent plot like this.

You will draw like this; you say this is plus 10 percent minus 10 percent whatever. It looks to this thing, but we can draw what is called bands. This is another way of presenting the results. So, this is called a scatterogram; scatter plot or it is also called as scattergram. I can do additional thing. So, I can have, I can on the same parity plot, I can have red color, green color, blue color to give to denote fluids have different Prandtl number. So, it is all, I mean beyond a certain point, just creativity; how you present the results, how you plot, and so on, alright.

Now, what is a [FL] behind this whole thing? Can I wipe it off?

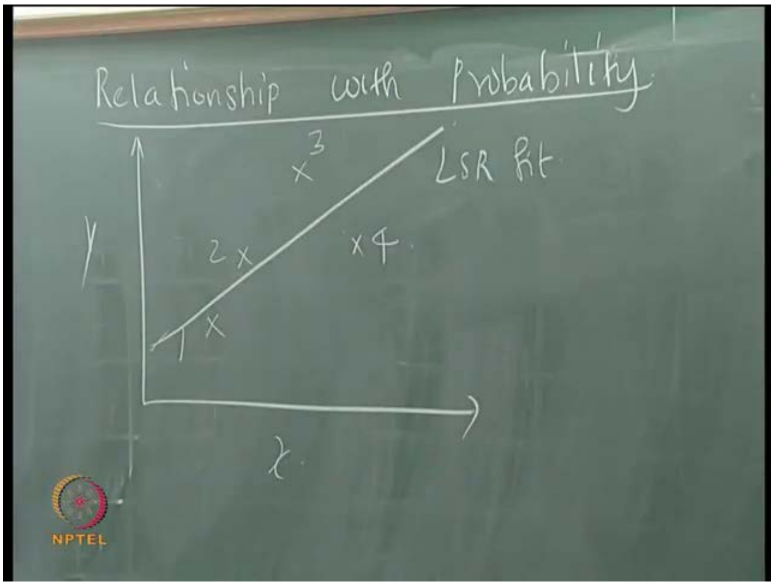
Yeah

No, in the absence of anything only mean and standard mean. For example, in this class room, I want to find suppose I want do get a, want do get the height of particular students in this class or whatever, I want the measure of it. I measure the heights of all of you and then I will take average; then I say that it is the average height is likely to be this, but then that S r will come; that is the variance. Of course, if Surekha has to be left out; Surekha and Mridula have to be left out; they are not part of this model; they have to be in separate model because boys and girls there may be some variable in it, ok.

So, what I am saying is in this, the first step would be I can take the average of I can take all the data points, take the average, and then, I can find out the difference of the individual height from this and report that plus or minus 3 sigma 8 and if the class it is large, it is likely to follow Gaussian or a Normal distribution.

Now, I say that I want to fit a correlation. Now, I say, we can go one step further; suppose I know the date of birth of each of these fellows and now I say that the height is directly proportional to the date of birth of these fellows, that is y is equal to a x plus b, with the date of birth I am getting a much better fit than taking all the this thing in Excel and taking the average rate. Now, if you tell me the date of birth, I am in a much better position to give you the height; is that clear. So, what is the reference? Whether with the date of birth I am getting something good or not, that that bench marking is without the date of birth, with just raw analysis of the data with respect to a crude mean, how it performs.

(Refer Slide Time: 29:37)

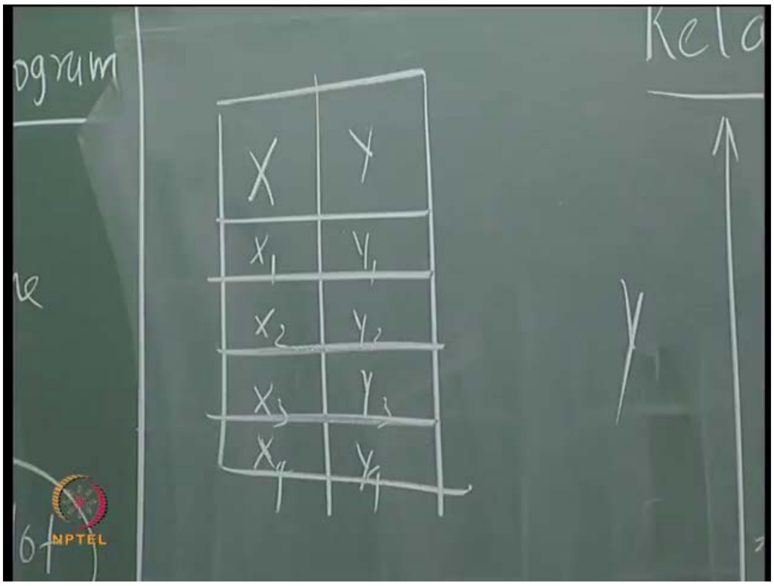


Relationship with Probability: Why do we have to draw horizontal lining? We could have drawn, we know that as r increases y increases, and then should not couldn't we have chosen better? No, if you if you already start thinking, then you are regressing.

The basic point in data analysis is if nothing else is known, you will start with mean and standard deviation. You will try to plot, you will try to draw histogram, and if it is a large sample, it will follow the Gaussian model. If you have already start started, if already started thinking about something, then you already started regression.

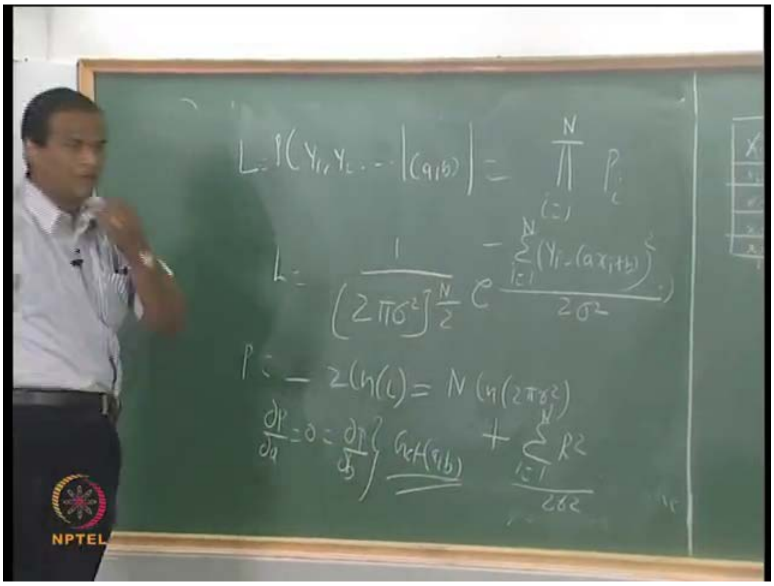
So, the relationship with probability: Now, let us say let us talk about 4 points. So, we want draw some straight line through this. So, this is our LSR fit. For example, now assuming that all the measurements are made with the same instruments, I can assume that the error, the errors associated with the measurement follow Gaussian with the same standard deviations sigma.

(Refer Slide Time: 31:30)



Now, for x 1; so, for the purpose of illustration, I have taken only 4 points.

(Refer Slide Time: 33:27)



Now, assuming the distribution of errors is Gaussian and I am saying that y is equal to a x plus b, this is basically y fit; y fit is equal to a x plus b; the probability of getting y 1 for this particular set a b should be true; a b are the two parameters which are yet to be retrieved is given by… is it correct? y 1 right this comes from Gaussian distribution, normal distribution (Refer Slide Time: 32:47).

Similarly, the probability of getting y 2 for the same given a b is given by…. So, therefore, the probability of getting y 1, y 2, y 3 and so on for a given a b is just given by the product of all the probabilities. The probability of getting, the probability of getting y 1 multiplied by the probability of, we cannot add the probabilities have to be multiplied because these are what are called IIDS right, you learnt about all this. These are all independent samples; I mean there is no connection; there is no connection between first and the second; these are truly independent even if this were to be so. Then, this total probability can be given by l, l is given by…(Refer Slide Time: 34:25)

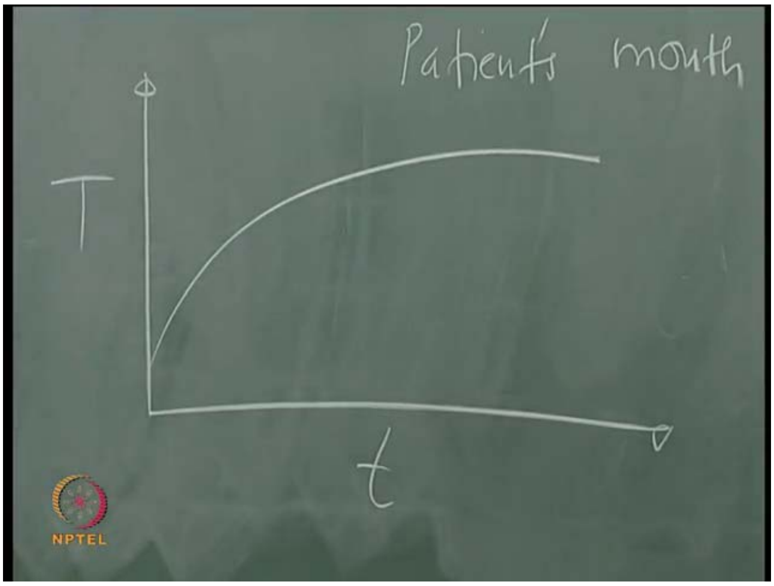
Now, you turn around and find the best values of a and b which will maximize p. Correct? You want to find out, you want to find out those values of a and b which will give the maximum probability. This strategy if you solve this equation to get the, if you want to solve these equation for the maximum values of p, you are essentially doing what is called the maximum likelihood estimation.

So, you can actually do minus to 2 ln l or minus ln l, if you do minus ln l, what you get? into 2 n plus this. This is also 3 sigma square, sigma is not a variable in the problem right; sigma is known it is an instrument characteristic. Now, you can go ahead and call this as p; you can do dou p by dou a is equal to 0 (Refer Slide Time: 35:26).

So, in the last class, we started with just this. So, we should have started from here. So, there is a link between what you have done and probability according to, this basically comes from what is equal that gauss mark of theorem and all that; so, there is a relationship between, so there is a relationship between the probability as defined by the Gaussian distribution. So, these are basically maximum likelihood estimates. Correct? If you solve for this, you will get a and b.

So, this is the connection between what we did yesterday; somebody had a doubt; I told them; I promised that, promised him that Deepak, somebody had a doubt; I promised that I will bring out the connection between this; is this clear? Let us not get into this too much. So, it is fine.

(Refer Slide Time: 37:22)



Now, let us solve one more problem, problem number 17. Let us do some unsteady… Mercury in glass thermometer in a patient's mouth; I am looking at an example which is analogous to this; how does the how does the temperature respond when the doctor treats?

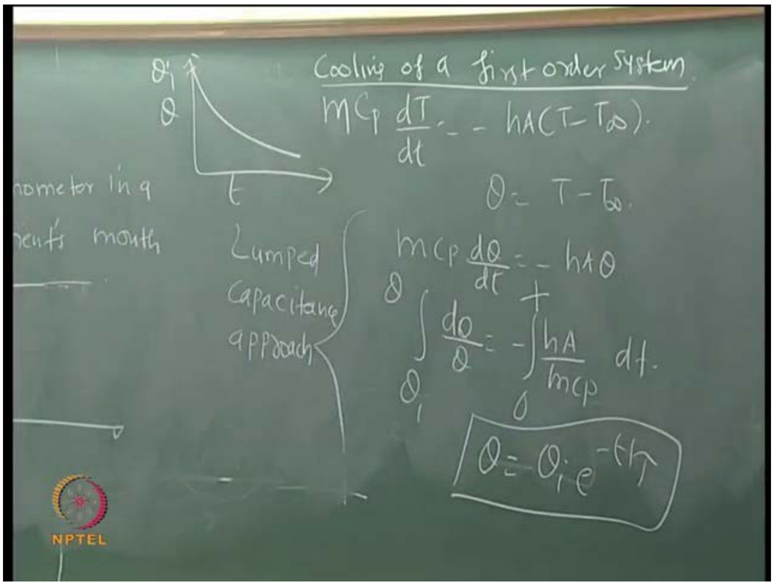
When will it reach the core temperature of the human body? Infinite times; what is infinity for the doctor? Two minutes; two minutes he will get 99 percent of the information because the emerging glass thermometer so designed that the time constant is very small. A thermometer kept in the arm pit will take five minutes; you will take five minutes, but however, if you drink Pepsi and go to the doctor, then it will sure be incorrect reading, or if you drink coffee, then you will be admitted.

So, half an hour before and after, you should not take all these. Therefore these are all, I mean, it has to be treated with caution. So, in Europe they use ear temperature; they have a thermometer where it just touches the ear drum or even in India, now we get something which measures the forehead temperature, or like in the swine flu detection, you will have a thermal infrared camera and then, but that is Ak 47 to kill a mosquito. So, that is normal clinical thermometer, doctor will even at the end of one minute, you have a good idea how it responds; two minutes, you have a fairly good because this 98.6 is not the you are not going to have 98.6 throughout 24 hours; you know that, right.

The human body has a lowest temperature at about 4 am and as the highest temperature at 3 pm in the afternoon or so. There is a minor variation of plus or minus one degrees one degree Fahrenheit 0.5 degree centigrade; after food it goes up; after exercise it goes up; when you sleep, it goes down. So, early morning if the temperature is normal, it may still be fever because of it is that is why fevers will become very clear in the evening. When already the body is hot, then over and above this normal temperature, it will be 99. Then any fever of 1 or 1.5 will show as 100.5 or 101 and the fevers tend to increase; you feel it increasing because already there is a component loaded to it.

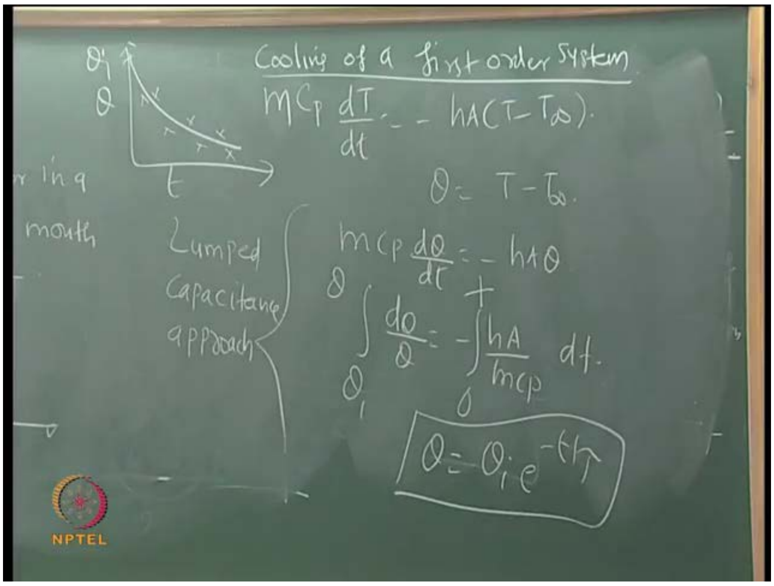
So, the 98.6 degrees is average body temperature of a British male is not having fever; no longer people no longer accept it 98.6 nobody cares; doctor will not care; anything less than 100, you will not worry. If it exceeds 100 in the morning time, morning if it exceeds 100, then you are in trouble. If it is 99 and all that, now a day's nobody gives medicine. So, it is part of the its part of the variations, part of the variations.

(Refer Slide Time: 41:07)



Now, what is the story about this? It cannot go down. Now, if m is the mass of the system, the first order system c p; of course, in this case it will be plus hA theta. Now, the temperature is increasing. I am writing this equation for a system where theta decreases to temperature that is cooling. I can say theta temperature excess is given by t minus t infinity; I can now say m c p…. So, basically, I call this lumped capacitance approach method. So, at time t equal to 0, this will be theta i. So, this tau is equal to m c p by hA, what are the units of tau.

(Refer Slide Time: 42:56)



Student: Seconds

You are sure? Second inverse

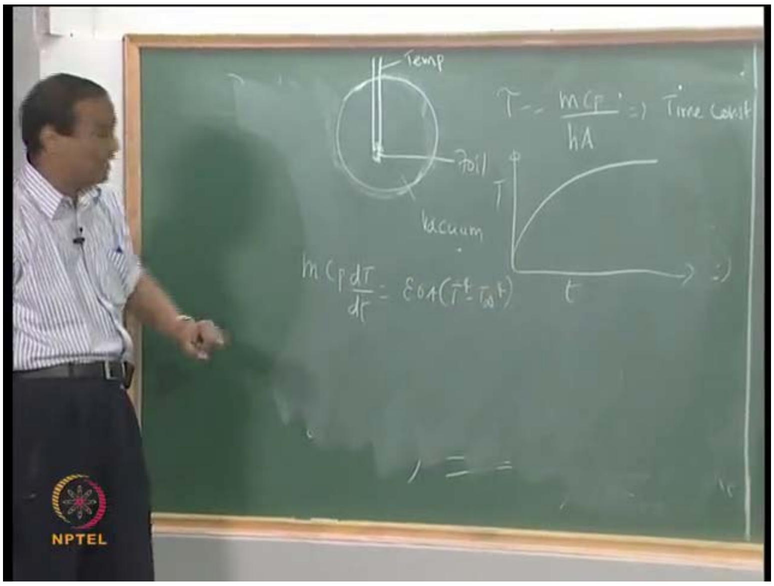
Student: Second

Time constant.

Now, is it a heat transfer class? No, it is an optimization class. So, we are doing regression. Why I am teaching at this? I am teaching the background so that we can work out a problem. So, basically, if you conduct unsteady heat transfer experiment, if you know theta i, then you will get theta as the function of t; you can get some experimental points like this. It is possible to treat this as a system, first order system with two unknowns, theta i and tau. And now you can solve for theta i and tau from tau if you know what the mass of the system, if you know the specific heat, if you know the surface area, and if you are pretty sure that the heat transfer coefficient is not changing during that narrow interval. So, this gives you one more method of evaluating the heat transfer coefficient using an experiment using unsteady heat transfer because if you have to do steady heat transfer, you have to repeatedly carry out several experiments.

Now, why do we want treat this theta a also as unknown because the moment it starts there is a rapid decrease in temperature; therefore, this theta i may be prone to error; it is good, it is good for you to treat theta I as an unknown, find out theta I, and compare with the experimental values which you have already obtained. So, this is the standard way of so this is the standard way of getting heat transfer coefficient using unsteady experiment.

(Refer Slide Time: 44:40)



Now, I can make it a little more interesting. Suppose, I can have I can have a doom like this; I can have a foil; I can have vacuum; I can heat it and if this is the m, c p, a everything of the foil is known; so, here I can measure the temperature; this is vacuum. Now, I completely avoid convection. Now, the temperature response will be like this; the temperature of this will go up in time, but now the governing equation is… Therefore, if you carefully do experiments, if I get temperature as the function of time, I can get the emissivity of the surface.

This is pretty interesting; d t by t to the power of minus t infinity to the power of 4. In partial fraction 1 by t square minus t infinity square minus 1 by t square plus t infinity square into 1 by 2 into 2 infinity t infinity square; that can be further broken on into 1 by 1 minus t t minus t infinity, 1 by 1 minus t infinity tan inverse of plus log of log of as analytical solution is possible. This should be asked in quiz. You can reduce it using partial fraction. So, this is standard way of getting emissivity, but you require a vacuum; are you getting this point?

So, in the exam, I do not know; depending on my mood, I will give some readings for emissivity, or the forward; this is called the forward model. If you solve this basically called the forward model, this becomes easier, but if I have a problem like this, it becomes more difficult and more involved for you to get the solution.

So, in the next class, I will again come up with some problem where something is cooling. I will give you some data, but now since it is e to the power of minus t by tau, if you have to take ln and then it just like today, you did ln of r e, ln of r u, we will do that and complete this. Then, we will go to multiple linear regressions where I will show it on the Microsoft Excel because it will take long time for doing in the class. Then, we will round off our discussion by taking a very realistic and very powerful case of non-linear regression, where this, for example, theta i and tau or epsilon, for example, it varies with iterations; then, it is going to be a lot more difficult.