Mass, Momentum and Energy Balances in Engineering Analysis
Prof. Pavitra Sandilya
Cryogenics Engineering Center
Indian Institute of Technology, Kharagpur

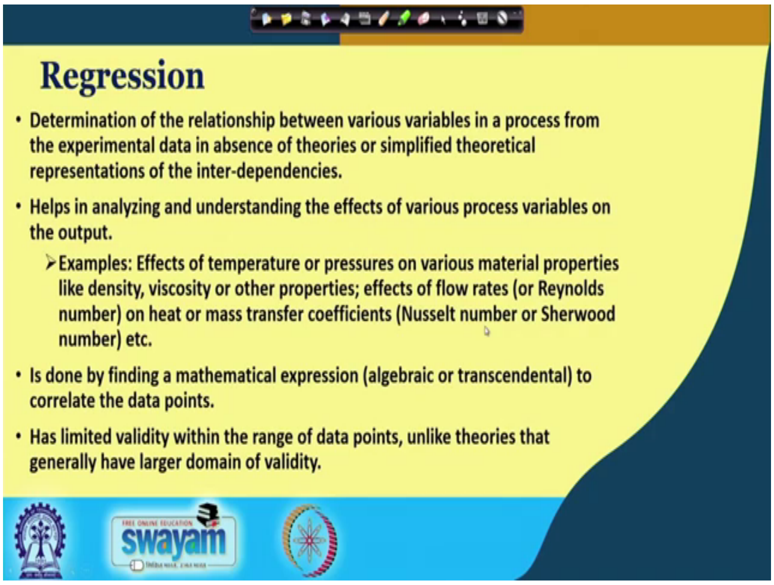
Lecture – 38
Regression

Welcome. After learning the various types of balanced equations now we shall be looking into the various ways of the data analysis and we have learned something about the methods of solution. Now many a times you find that when you have the model equations with you; you need to know some of the values of the properties or some coefficients like heat transfer coefficient mass transfer coefficient etcetera from some given data. The data may be given in terms of some correlations and sometimes in terms of some tabular form. Now in many cases you find that the exact data are not given to you and in those cases you have to use some kind of mathematical ways to find out those missing values.

So, in that respect we need to learns what are those mathematical techniques and many a times you also want to know the effects of the various variables on the process. For example, if you are doing some heat transfer experiment or heat transfer analysis, you find that you either you can find the effects of the say flow rate on the heat transfer performance. And sometimes you need also to figure out that what is the say effect of your given some kind of say data say flow rate versus the some performance of the process. So, you need to know that, what kind of effect is there in terms of that whether it is linear effect whether it is non-linear effect if there is non-linearity then what kind of non-linearity is there in the effects.

So, all these things can be understood if we do some analysis mathematically. So, in this context first we shall be looking into the regression analysis. So, in this particular thing we shall be looking into linear regression, then polynomial regression and exponential regression.

(Refer Slide Time: 02:27)



So, first let us understand what we mean by regression. Regression means that we are trying to determine the relationship between the various variables involved in a process and these variables will be given in terms of some experimental data. Now when do we need to know these relationships from the experimental data is this that, many a times we may not have a proper theory to correlate the variables or many a times we find that. Even if there are some theories, but the theoretical expressions becomes so, complicated and complex, that we are not able to adopt them in our analysis easily and sometimes we also do not need to have the exact theoretical expressions.

So, in those cases we try to just figure out that what are the primary variables which are of our concerned in a given analysis and from those values of the variables, we try to adopt some simplified way of correlating these variables and that is what we do with regression. So, in this way what we find that this regression helps us in analyzing and understanding the effects of the process variables easily ok. Now examples are like for example, we want to know the effect of the temperature or pressures on the various material properties like density, viscosity and other properties.

Like in this case we do not need to always carry outs the; carry out the experiments or we do not need to go into the theoretical analysis just by taking some kind of experimental data at some specific points, we can regress those data and by that we can also find out the very values of the particular properties at some other values of the temperature pressure which are not given to us.

And in another example that, for example, we need to know the effects of flow rate generally flow rate comes in the through the Reynolds number on the heat and mass transfer coefficients, which are again generally come true Nusselt number and Sherwood number.

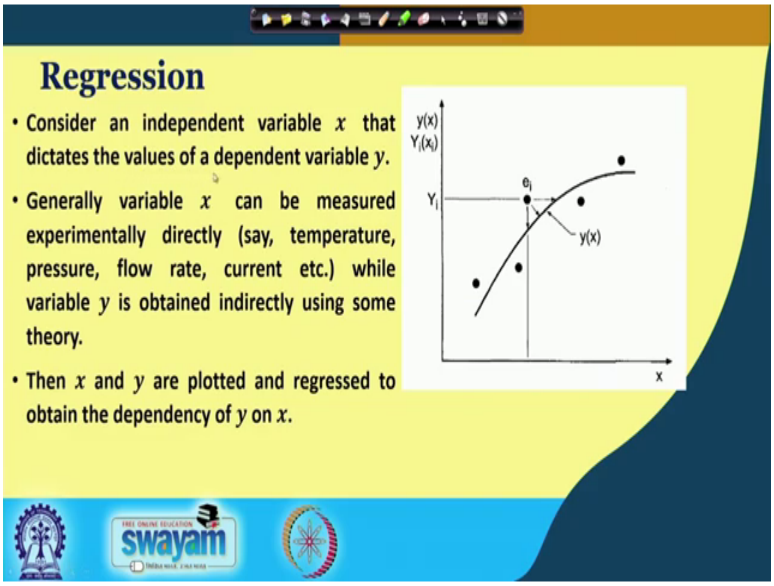
So, here you will find that a somebody might have carried out some experiments in those experiments somebody might have reported the flow rates and with each flow rate they might be reporting the values of these coefficients values and then what you do that you find out the Reynolds number from those flow rates and then you convert these coefficient values into either Nusselt number or Sherwood number and then try to regress these values. So, that you can obtain a generalized correlation which will be having a you more universal validity.

Now, I told you in my earlier lecture what is the importance of putting the variables, process variables in terms of the non-dimensional numbers. Now this regression is generally done by finding some mathematical expression to correlate the data points. Now these mathematical expressions could be algebraic expressions for example, we can have linear expression, we can have quadratic expression or we can have some polynomic polynomial or they could be transcendental. Transcendental means it could be say exponential kind of logarithmic or say hyperbolic. So, all these kind of expressions are called transcendental.

So, we can have either algebraic expressions or transcendental expressions to correlate this data. So, despite all these simplifications being provided by the regression, we find that they have some limitation. And limitation comes because we are taking a fixed number of data points to understand the behavior. Now that the whatever regression after regression, we are finding the correlation we find that those correlations will be valid within the range for which we have taken the data.

And anything outside this range if you want to predict, will be a bit could be you know newest if you find that there is high non-linearity. Now if there are linear expressions then perhaps we can extrapolate to some extent, but other than that if we find that there is some highly non-linear behavior of the 2 types of variables, then we should be careful not to extend or use this regression expression for outside the domain of our determination.

(Refer Slide Time: 07:13)



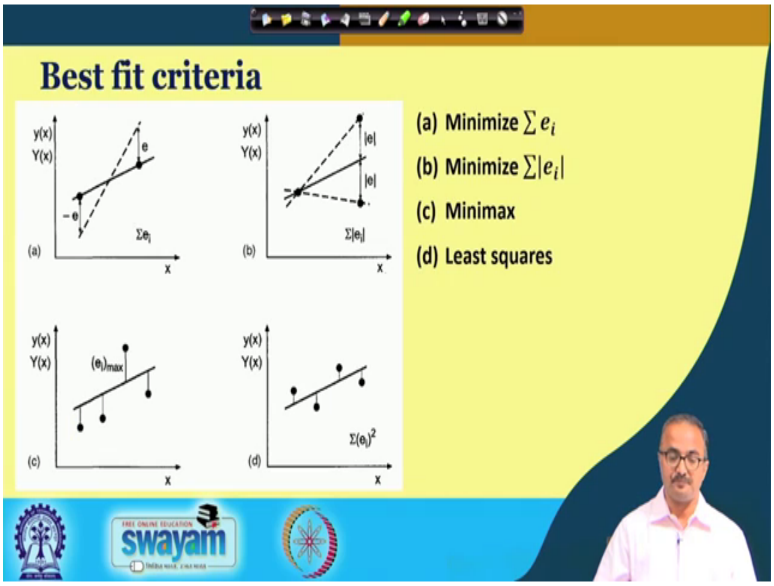
Now, here we show that how mathematically how we represent it or graphically we represent it. So, if let us see that we have 2 variables one is x another is y now x is what? x is that that dictates the values often dependent variable y. So, x is taken to be the independent variable, and y is taken to be the dependent variable ok. And generally x can be the measured variable for example, we can measure temperature, we can measure flow rate etcetera can directly and y is the variable which is obtained indirectly; that means, for example, if you are talking about specific heat.

Then specific heat cannot be obtained directly by measure through by any kind of sensor, but it has to be obtained by indirectly by combining the various types of properties or the measured variables like temperature pressure etcetera ok. Now, what we do that, now we plot this y versus x here and suppose these dot dotted points are the values of this at each x point or the value of y.

Then what we do by regression? We find some kind of curve which is passing through somewhere between these points, it is not collecting the joining the points, but it is passing somewhere between these points and then this particular curve which is called y in terms of x, this particular curve is obtained after with the do the regression analysis.

Now we shall be looking into this concept and try to understand it by some examples now.

(Refer Slide Time: 08:50)



So, when we say that we have to regress, what we mean that we have to find the best fit. Now best fit means that various ways we can define the best fit. Here in this example for example, you see that this particular these two points are the actual data points. And suppose through this we are sort of taking this regression regressed curve and what we say that let us see that defined some error. The error is that between the regressed value that is coming from here and the actual value ok.

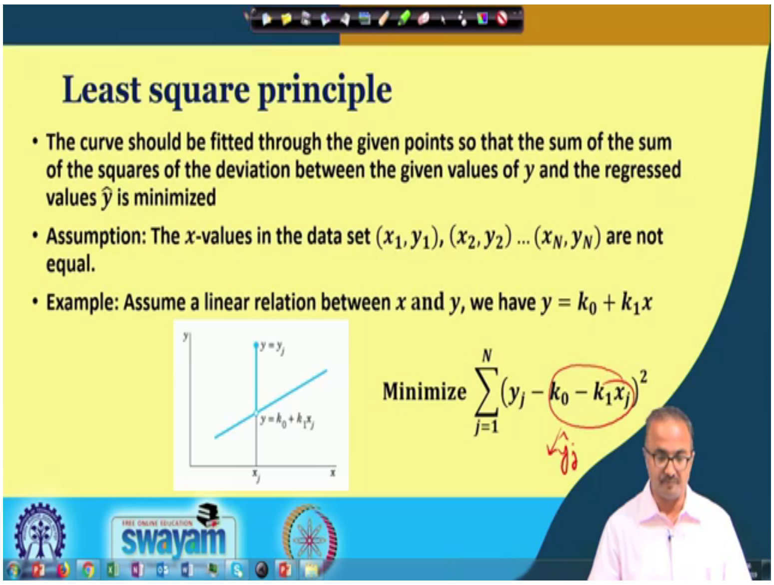
So, if I say that the actual this regressed value minus actual value in this case we find this is coming to be negative, because this is regressed value is less than the actual value and similarly in this case we find that regressed value minus the actual value is coming out to be positive. So, we are what we are trying to do that, we are trying to minimize the summation of the errors; that means, whatever for say the there are n number of data points for each of the points, we are finding the error between the regressed value and the actual value and we are summing up the errors and we are trying to minimize the total error.

Now in this case what we find that, the errors could be negative and positive and many a times this negative and positive values may get cancelled. And if they get cancelled it would look as if the total will be 0 in that case it would look as if there is the it is this regressed value and the actual value are matching perfectly, but that is not the case. So, in that case we have to devise some other way, one way is that we minimize the total of the absolute value.

So, here we have that we have this absolute value like we take whether its positive negative we take the absolute values and then sum them up ok. This is another way for doing this fitting. Third way is this that we again may find the errors at different regressed points and here we show that different regressed points, but in the errors and then what we do? We try to minimize the maximum error among all these errors ok.

So, this is called the minimax problem and lastly the least square error in the least square error what we are doing? We are trying to least square means we are squaring the errors for each regressed value we are squaring the error. So, you see that we are showing the error here and then summing them up ok. So, this least square method is a very popular method and in this particular lecture we shall be looking into this least square method.

(Refer Slide Time: 11:33)



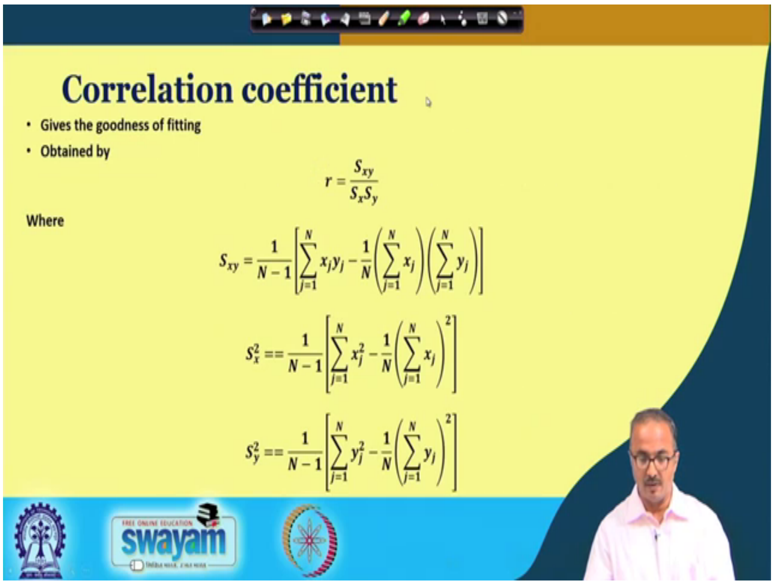
So, let us first understand the principle of least square. I am not going to detail of these things because these are dealt separately in the dedicated courses on numerical techniques and statistical methods. So, here I shall be just giving you the overview of these methods. So, in the least square principle, what we have that the curve should be fitted through the given points so, that the sum of the squares of deviation between the given values of y and the regressed value say y (Refer Time: 12:07) is minimized ok. So, that is the meaning that in this we the repetition of this.

So, here we see that we have the regressed value say this is the y (Refer Time: 12:22) and this y is the actual value. So, what we do? We try to minimize the sum of the squares of these errors and we assume that the x values in the data set this x 1 x 2 etcetera whatever the data set we are taking in this x values are not the same they are different. Now suppose we want a linear regression; that means, we say that the y depends linearly on x. So, this is a linear expression you can see that y equal to k naught plus k 1 into x.

Now, what is k naught? K naught is the what we got the intercept and k 1 is the slope ok. So, now, you see that if you plot y versus x here and here we see that this is the line which is the y equal to k naught plus k 1 into x ok. So, this is y j; so, this is this particular one is y j. So, y j equal to this and then we this y is the sorry this y is not the y j, this is the regressed value that is y tilde this y tilde and this is the actual value y j. So, what we are trying to do? We see that there is an error between this y j and y ok. So, this is the regressed expression this is a regressed curve ok.

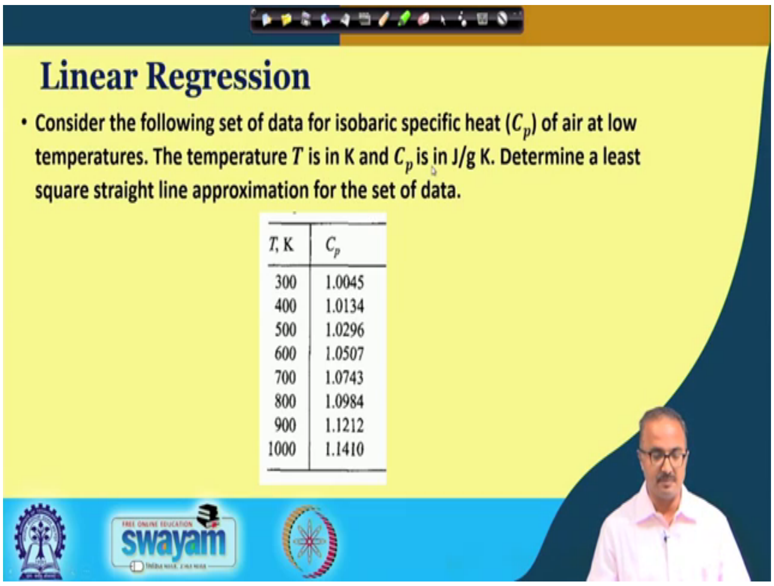
So, this particular error a square of the error we try to minimize. So, this is the expression and minimize the sum of the squares of the error this y j is the we can put it y j prime this tilde this y j (Refer Time: 14:01) is minus this is the regressed value the actual value. So, we find the this is the actual value this is the regressed value this whole thing; this whole thing is the regressed value ok. So, we try to minimize the square of the errors.

(Refer Slide Time: 14:28)



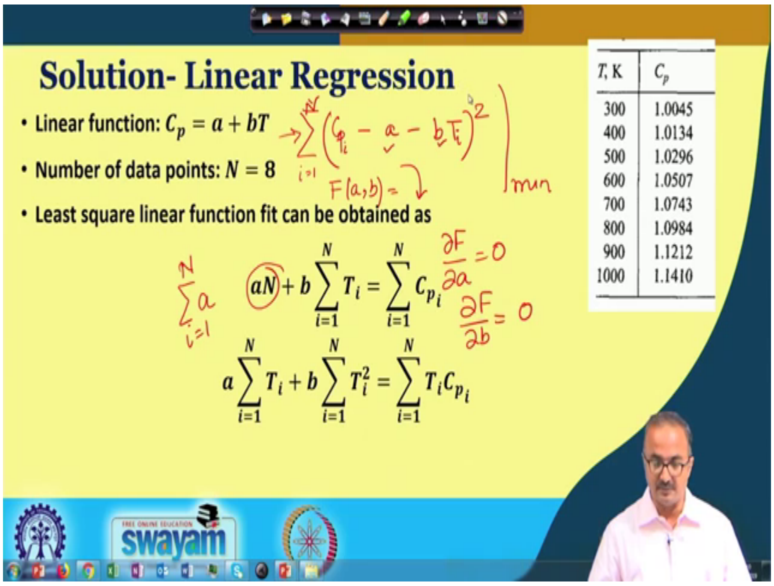
Now whenever we are regressing with some kind of equation; whether it is linear or polynomial. So, first we have to understand that what is the goodness of this regression. So, in that to understand the goodness we define something what we call the correlation coefficient and this is defined like this ok. And in this expression we find this is the S x y this is the S x square and this is the S y square. So, you can see here that these coefficients are obtained from the given values of the data points now what we do?

(Refer Slide Time: 15:11)



We take this example here we have we have asked that we are given some data set of T versus C p ok. So, T is the temperature in Kelvin and C p is the isobaric specific heat and that is given in terms of joule per gram per Kelvin. So, we have been asked to determine a least squares straight line approximation for the data set. So, we have to find a linear fit to this data set.

(Refer Slide Time: 15:39)



So, here we have that we have been. So, we put this specific heat in terms of linearly in terms of temperature. So, let us put C p equal to a plus b T and the number of data points is taken to be 8. So, here we have the expressions.

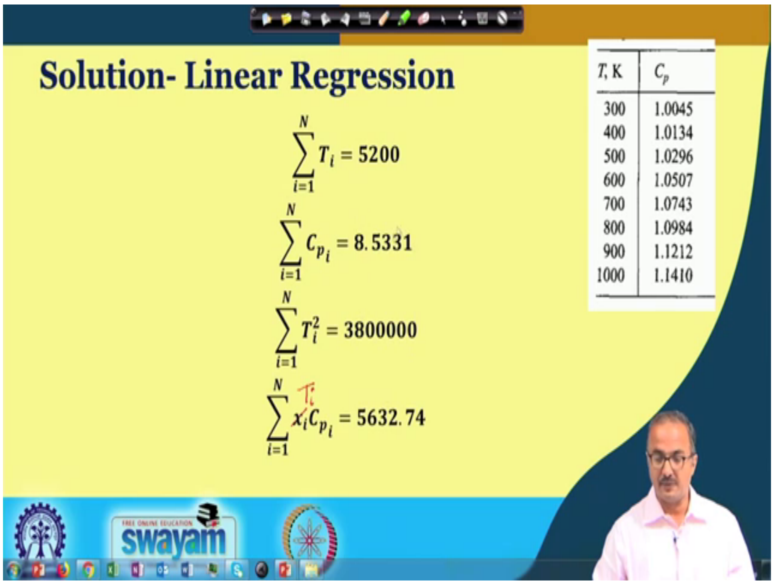
Now I will given a hint to you that how we get these expressions its very simple that you first see that from this expression what you find that C p minus a minus b T. So, this is the error and this you square and take a sum and this sum is basically say i to N ok. So, this is how it is. Now if you to minimize then this you have to minimize this thing you have to minimize now minimize with respect to what? Minimize with respect to A and B ok; that means, you have to adjust this a and b value so, that you can minimize this error and what to do that what we do that as we have learnt in calculus, to minimize we have take a derive first derivative first to take it to be 0.

So, suppose I define this expression as some kind of a function in terms of a and b now please understand we are not defining these functions in terms of T because at each point of to this T i actually this for each 1 of T i we have been given the value of C p i. So, it is functional basically a and b and this is the function. So, this is the function here and to minimize this error function what we do? We first take this dou F by dou a equal to 0 and then what we do? Dou F by dou b equal to 0. So, you do this to you can do this exercise very easily, you find the derivative of this error function with respect to first a and then with respect to b and equate them to 0.

Once you equate them to 0 and the rearrange them, you will find you are getting this particular expression and this expression. Now this expression you will find this will come because we will find that this thing will come that you will get something like this i equal to 1 to N. So, from that you will get this a is coming out and you will get a into N that is how we get the expression and rest of the things you can find out easily. So, you see that we are taking this a and b out of the summation because they do not depend on the ith the value of the particular coefficient. And now you see that we have 2 expressions and this in this 2 expressions, we know the temperatures we know the specific heats what we do not know are the a and b values.

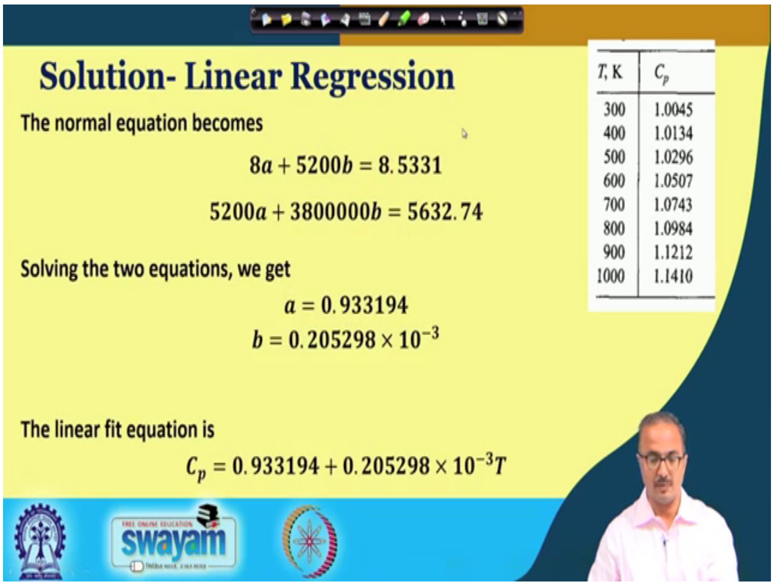
So, we have 2 unknowns and 2 equations and now from these 2 unknowns and 2 equations you can easily find out the values of the a and b. So, this is how we will going to we are going to solve these expressions.

(Refer Slide Time: 18:52)



Now after understanding this now what we do we get all these values of these summations of temperature, the specific heat and the square of the temperature and heat will be this will be T i into C p i. So, this summation of the products we get ok.

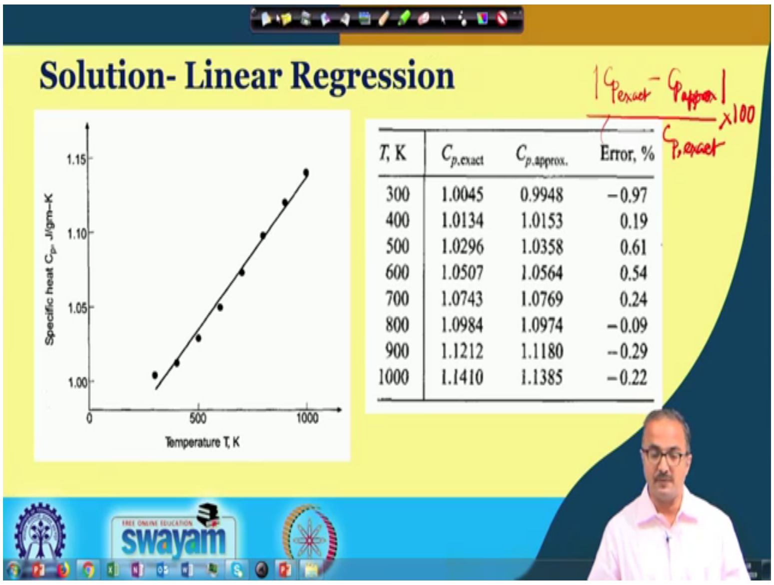
(Refer Slide Time: 19:22)



And after obtaining all these we put them in these expressions which found earlier and from these 2 expression equations we solve for a and b and these are the values of the a and b. So, now, we find that the linear fit the linear expression which is fitting this specific heat be the temperature comes out to be like this ok.

Now, whenever you are using any of these fitted equations one thing you must be very careful about is this that whatever units you have used to regress, they must be maintained whenever you are going to use this expression in future for any kind of determination of the C p or the dependent variable. If the temperature is suppose in centigrade or Fahrenheit you must first convert them into Kelvin to get the value of C p. If you make any mistake in using the proper units you are going to land up with wrong answers. So, that you should be very careful about whenever you are going to use any kind of regressed equation.

(Refer Slide Time: 20:34)



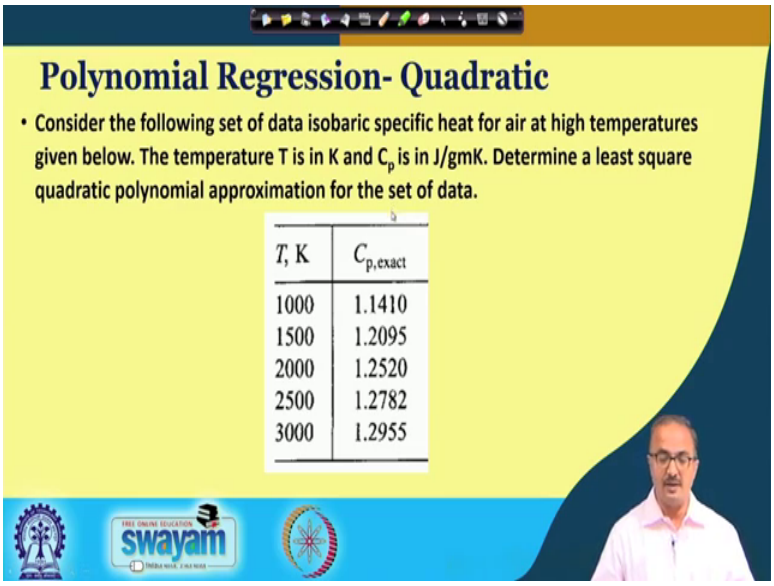
So, here we see that after regression whatever values of the C p we are obtaining. So, we are putting those values here. So, this is the solid line shows our regressed expression and these points are the actual data points ok. Now we put in this table the temperature values and here we put the exact or the given values of the specific heat and in this we are now getting the values of the C p, which we are calling approximate because how we are getting these values?

We are again putting these temperature values in the expression we have just obtained and from that we are getting the values of the specific heat. So, here we are putting those specific heat values and after that what we are doing? We are trying to find out the percentage error; that means this approximate value minus this exact value and take the absolute value ok.

So, this is for finding this percentage error what we are doing? This is nothing, but this mod of this exact minus approximate and divided by the exact into 100 ok. So, this is how we are able to get the percentage error. Understand one thing whenever you are finding the percentage error, it must always be found from the exact value with respect to the exact value what is the error. You never put here the regressed value not the approximate value ok.

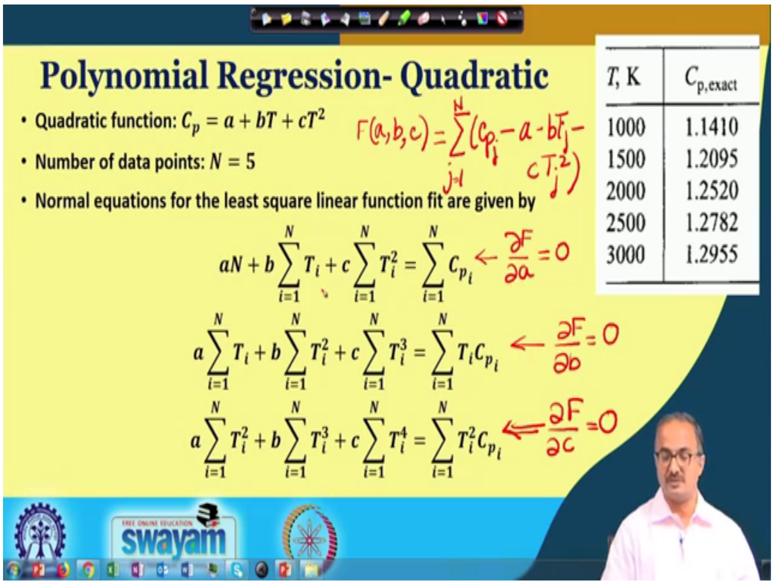
So, that is a one we see that after we make this error analysis what we find that, there is a good matching means the maximum error is coming to about 0.97 percent ok. So, even it is negative or positive does not matter, but what we are finding it is coming very small. So, that is how we can see that our regression or this assumption of a linear fit between the specific heat as and the temperature seems to be ok. Next we take another example.

(Refer Slide Time: 22:57)



Here again we are given a set of the specific heat verse as a function of temperature and this is given a high temperature and the same air ok. In the last problem was on was at low temperature this is for high temperature and again we find that here temperatures in terms of Kelvin and specific heat is in terms of joule per gram per Kelvin ok. Now we have been asked to find a least square quadratic polynomial.

(Refer Slide Time: 23:30)



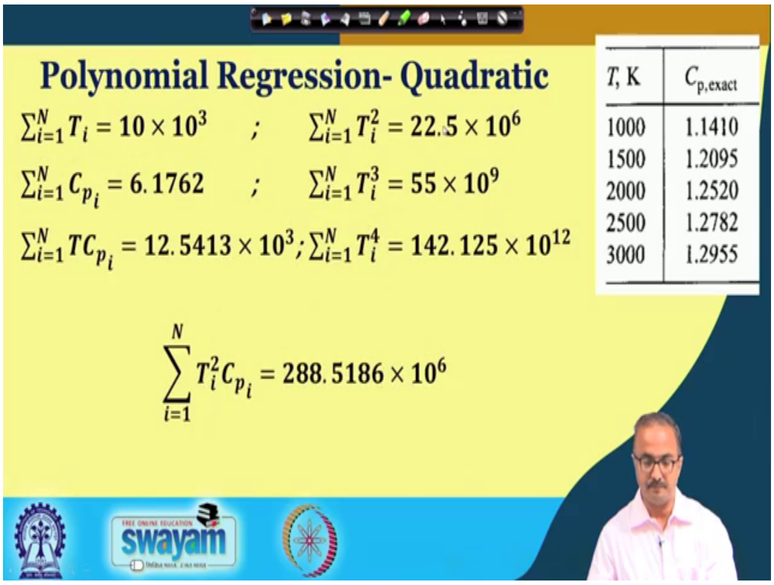
So what we shall do that now, we shall do this expression, get this expression we assume that C p equal to a plus b T plus c T square. Now in this case as I told you earlier what you can do is that, you can simply again put the error function and now it will be in function of a b and c and this you take to be like C p j which is given to you minus a minus b b T j minus c T j square and j is equal to 1 to N ok.

Now, what you do? You simply find now this expression has been obtained how this dou F by dou a is equal to 0, then this expression that have been obtained from this dou F by dou b equal to 0 and this expression has been obtained like dou F by dou c equal to 0 ok. So, by doing this you get these expressions and you find that these three equations and we have three unknowns that is a b and c ok. Now once you get these equations it becomes now easy for you to solve these equations and you can get the values of a b c. Now you see that as we keep on increasing the degree of the polynomial for regression, we find we will be having still more number of equations.

And when you have so, many equations what you can do is this, you can use any of the methods which we have learnt earlier matrix methods to solve. Up to say third degree second degree polynomial where we get three x equations there perhaps we can do it by manually.

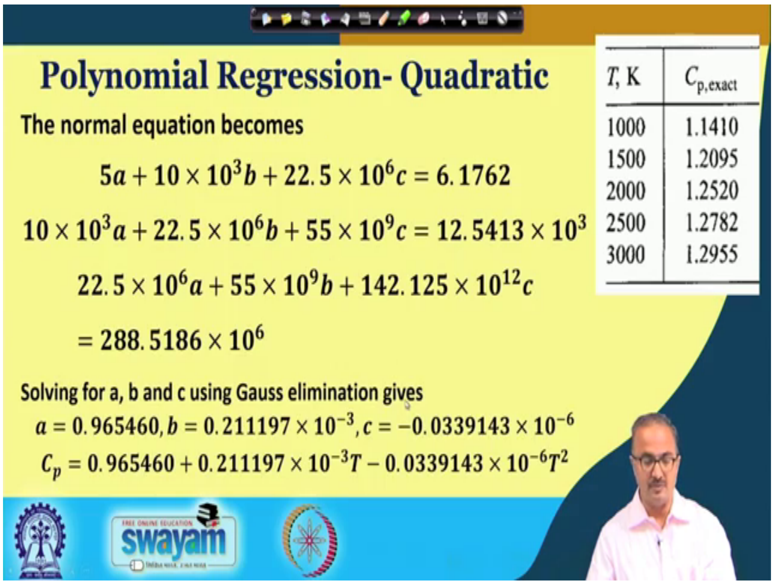
But as the number of equations increase, then you have to use an efficient matrix methods and sometimes you have to go for some coding to solve this expressions and these tools. For these codings it is always preferable that you can develop your own coding so, that you also develop some coding skill that always helps you. Otherwise you can also use the standard packages like excel or matlab to solve this set of equations. Now let us see how they will look like.

(Refer Slide Time: 26:12)



So, here you see that we are finding the values of the various summations involved in the given equation and then after plugging in these.

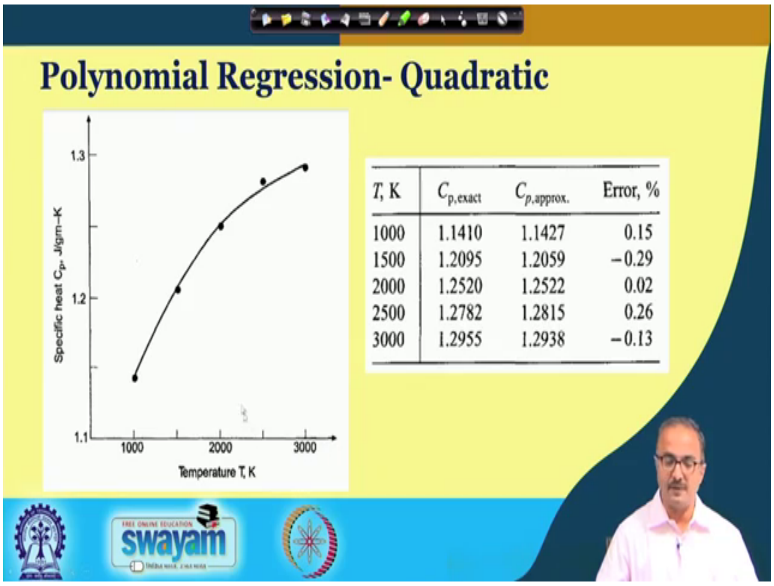
(Refer Slide Time: 26:22)



Values we find that we are getting these three equations and from these three equations by using the gauss elimination method we are able to solve for this a b and c and now we get the regressed expression for C p in terms of the temperature.

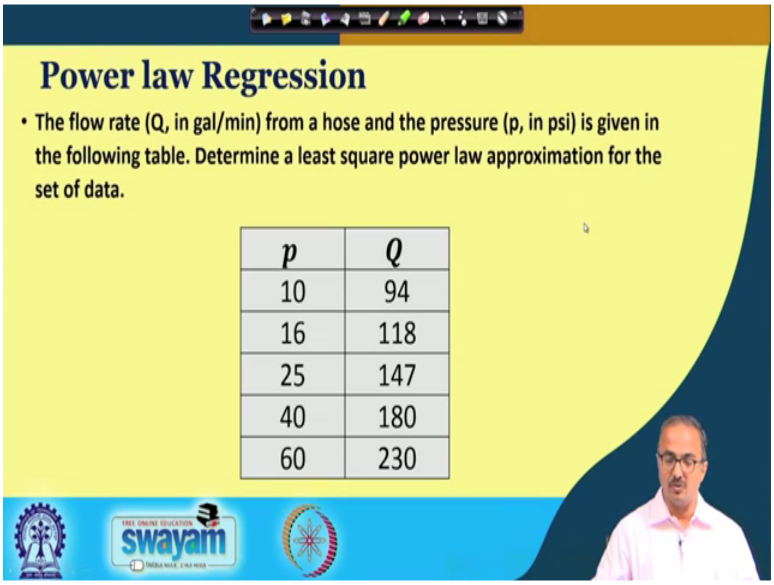
Now, whenever you are getting all these regressed expressions one thing you understand that, they are just simply given giving us the dependency of the dependent variable on the independent variable. But they are not telling us anything about the physics means why they should follow this kind of pattern. These are not there whenever we are going for regression ok, but these just help us whenever suppose you are going for some simulation. So, you can quickly take these values either from some data table regressed them, within the range of your interest and they will help you to quickly do the simulation without going into the physics of the particular process.

(Refer Slide Time: 27:32)



Now, here we find that we again plot the regressed value along with the actual values which are given to us and a like earlier we find out the percentage error. So, here we find the percentage errors and here we also we have see that, this quadratic regression gives a quite good fit and the maximum error which is there is, only about 0.29 percent ok.

(Refer Slide Time: 28:03)



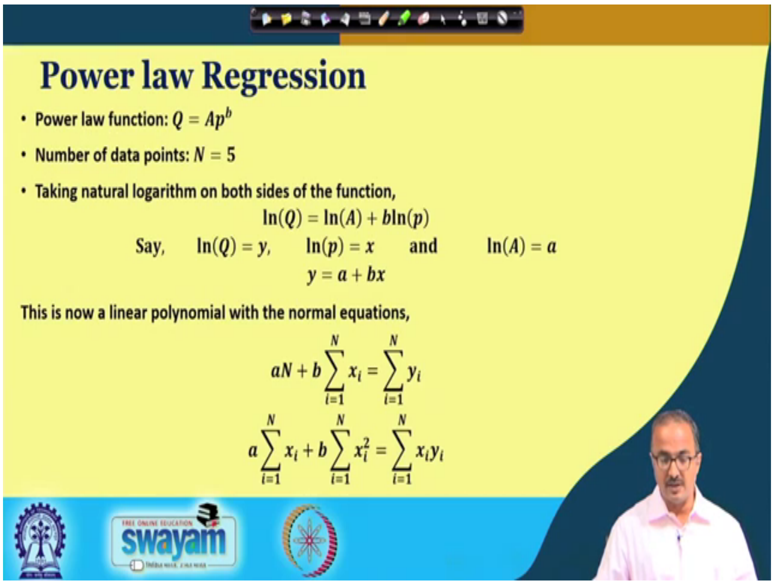
So, that says that the regression is good should I go hello (Refer Time: 28:08) [FL].

[FL].

[FL]. Next we take this particular expression in this problem we have to go for a power law regression. What we mean by this will be clear to you. Now first let us see that kind of data set which have been given to us that is we are given the pressure and at each pressure we are given some flow rate through a hose ok. And we know that if whenever in a hose is a some flow rate is going on now we can find out the flow rate in terms of the pressure drop ok.

So, here we are given that pressure drop you can say as if and each pressure drop we are given the flow and again you please take care of the units it is gallons per meter and pressure is in psi that is pounds per square inch. And we have to find a least square power law approximation.

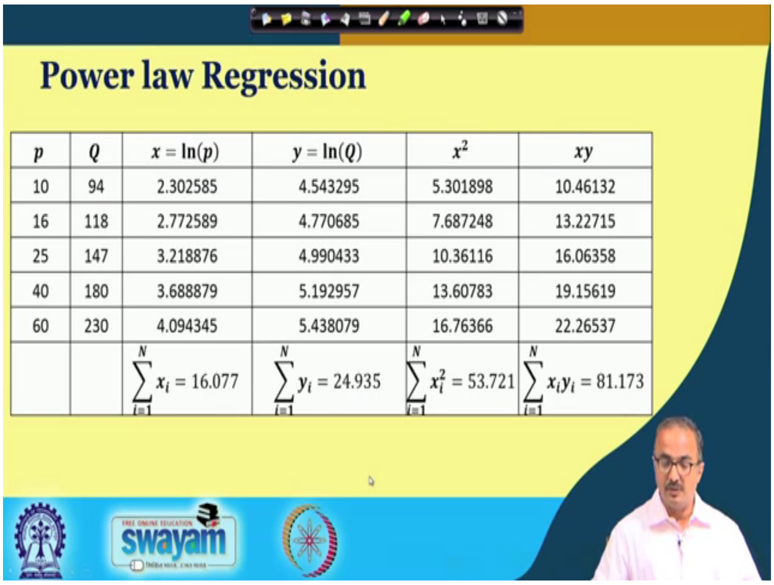
(Refer Slide Time: 29:24)



So, what we mean and how to express this power law approximation for that let us see this. That we put this is power law that we put Q that is the flow rate as function of A equal to p to the power b and here we have this a N equal to 5 ok. Now to put this in terms of our a square so, that we can solve it in a linear fashion what we do? We take the logarithm of this whole thing from both sides.

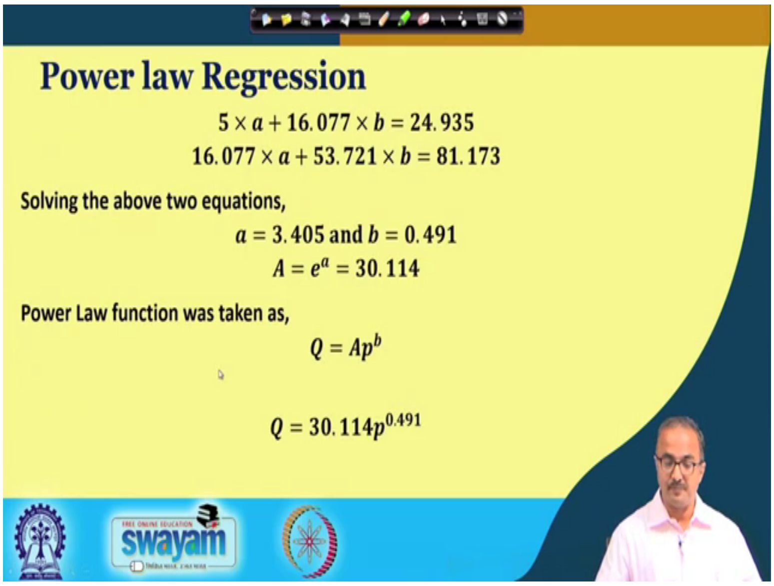
So, we take natural logarithm. So, here if we find that ln Q equal to ln of A equal to b plus ln of p ok. Now you see that we replace this ln of the; that means, natural log of the flow rate as y and then logarithmic of A we take as a and logarithmic p we take as x. So, we have this equivalent expression that is y is equal to a plus b x. Now whatever we have done earlier to do a linear curve fitting we apply the same method only thing is this from the given table first we have to find out the logarithm of q and logarithm of p and we shall be regressing between log q and log p. So, here this x is representing log p and y representing log q.

(Refer Slide Time: 30:48)



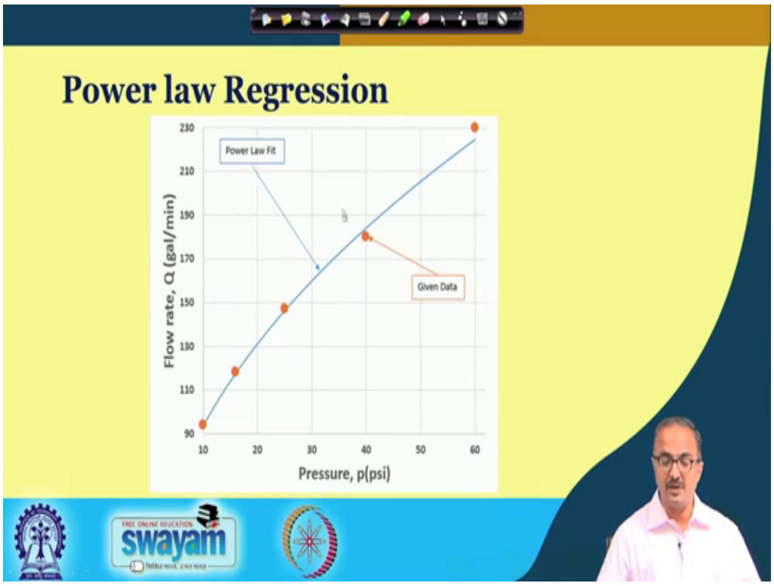
So, here we have the solution that p Q we first write from the given data, then we find a log p value over here then log Q value over here and this is x this is y we find the x square value and the x y value and these are the summations which are needed to find out the values of the a and b.

(Refer Slide Time: 31:07)



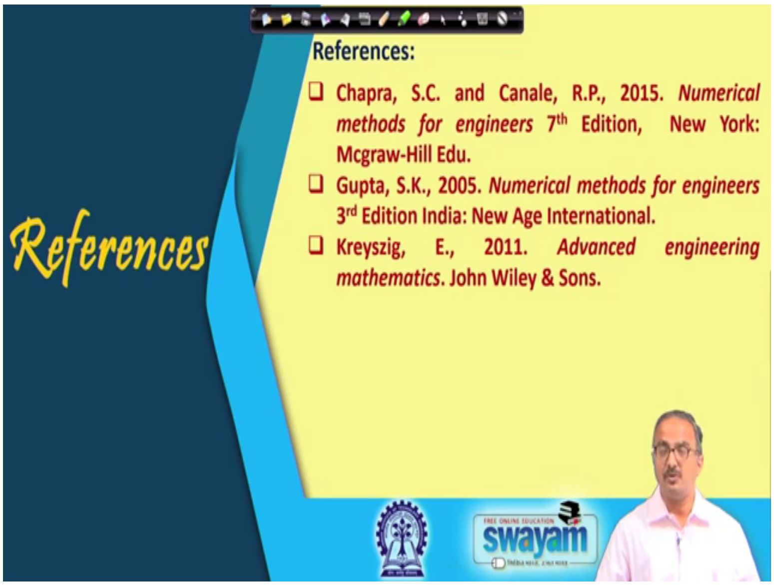
And here we have these 2 equations in terms of a and b and here we get the value of a this and b this. Now A we know that this capital A which is there in the original expression. So, that is equal to exponential of this small a which we have found out and this is coming out like this. So, that is now we find that Q is equal to this a value here and p to the power b that is 0.491.

(Refer Slide Time: 31:39)



So, here we are putting the regressed curve here and these are the data points given and we find that it is not a linear ok. When we are plotting in a linear scale it is not linear, but if we plot this p and Q in your log scale then we shall find that this will come to be a linear curve that is the difference. So, here we find that which is not a linear curve, its the kind of a exponential curve which we.

(Refer Slide Time: 32:11)



So, these are the references which you can look for more detail of this kind of techniques.

Thank you.