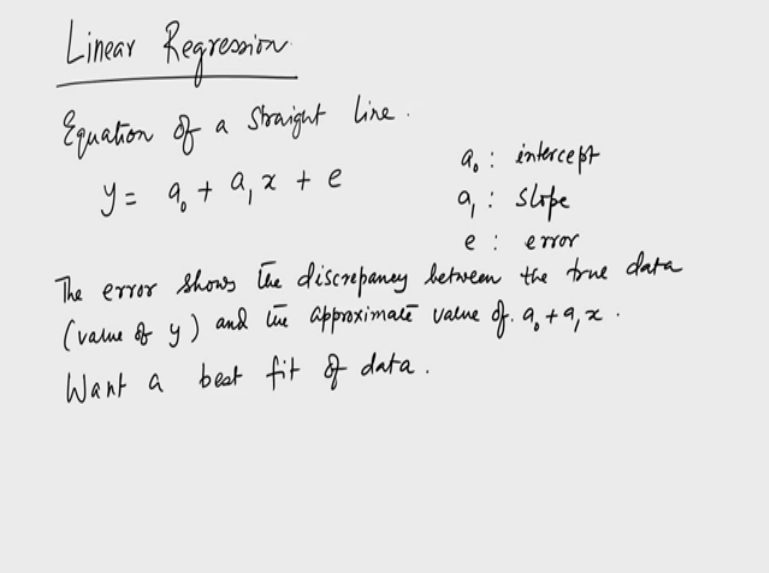
Numerical Methods and Simulation
Techniques for Scientists And Engineers
Saurabh Basu
Department of Physics
Indian Institute of Technology- Guwahati

Lecture 07
Linear and Polynomial regression

So, we have discussed fitting of data using 2 methods minimal methods one is called as a Lagrange's polynomial method and the other one is a Newton's polynomial method. And now we are going to learn the regression method which is purely based on you know the statistics that we have learnt in the previous lecture.

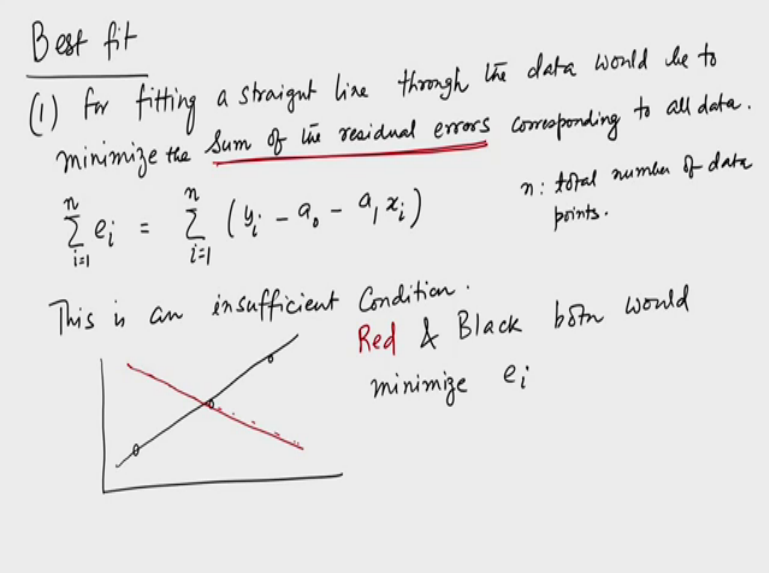
(Refer Slide Time: 01:00)



So, we are talking about linear regression and let us understand that what it means so we know that the equation of a straight line is nothing but y equal to a 0 + a 1 x + some e we do not write it with e usually we write it simply a 0 + a 1 x where a 0 is the intercept in the on the y axis and a 1 is the slope. But then this e is included as an error or the discrepancy to the true data. So, you have a table of data which x i and y i are given where i runs from 1 to n, n are the number of data points and you want to fit it to a straight line and let us see how do we do that.

And this e actually ah ah corresponds to the error or the deviation from each one of those data points given by x i and y i. So, needless to say that your a 0 is the intercept a 1 is slope and e is the error of these line from the data points. So, the error which is important here shows the discrepancy between the 2 true data which is value of y say for example that is value of y and the approximate value given by the straight line which is equal to a 0 + a1x which is predicted by the linear line or the linear in this case the straight line okay.

(Refer Slide Time: 03:54)



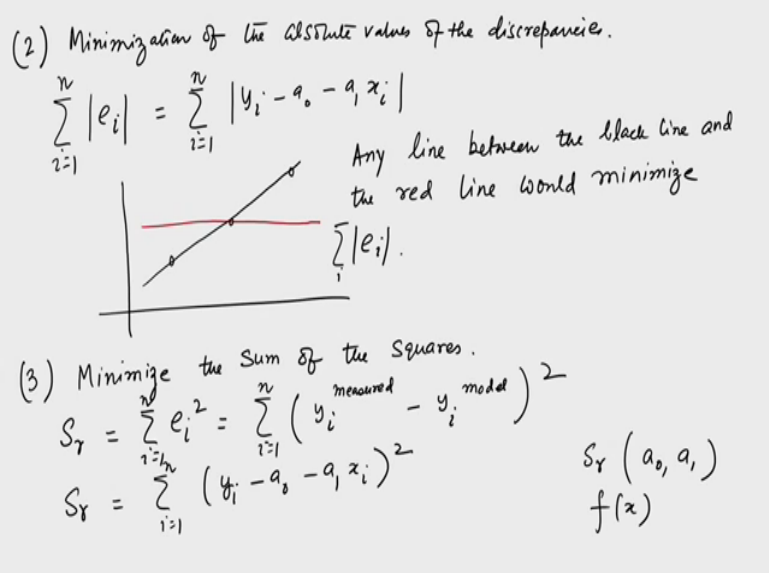
So, we want to have a best fit of this, best fit of data for a given linear relationship as its given here. So, if you want to get a best fit one has to follow certain procedures. Let us see what the procedure is, so one strategy we will call it one. So, this is for fitting a straight line through the data would be to minimize the error or basically we are talking about the some of the error or which is called as some of the residual errors they tell you what it means.

Let us highlight this and corresponding to all data points okay. So, what I mean is that we have to minimize a quantity called as e i sum over i i runs from say 1 to n, n is the total number of data points and which from the linear equation we can write it as y i - a 0 - a 1 x i and so n is the total number of; so what we need to do is that we need to minimize this error so that the data that we have obtained either from a numerical simulation or from an experiment that can be fitted to a straight line.

And if you try to do that you find that this is an insufficient condition and y it is an insufficient condition. Let us take an example of you know there are say three data points and one can actually draw a straight line through this data point all these 3 data points but you see if you want to minimize this even this line that we draw it with a different colour. Even this line which passes through the midpoint of these 3 lines is an equally probable plot though it looks really bad.

But this also this red line and the bold so let us write the red line that is you know write it as a continuous line. So, this red line and so both you know red and and black both would minimize e i okay. So, clearly the red line is not a fit to the curve but it satisfies this criterion, which means that this criterion or this condition is insufficient.

(Refer Slide Time: 07:58)



And if we could look at a slightly better one then we would write it as this is number 2 and in so more logical one it seems like is a minimization of the absolute values of the discrepancies. So, this is like e í and we have i equal to n and we have taken the magnitude ignored the sign of this and this is equal to i equal to 1 to n and a y i - a 0 - a 1 x i so you need to minimize this. But this is also not sufficient because let us draw again these 3 points say for example so this and this and draw a straight line through these points.

And another line which is like this let us draw it with a different colour as we have done for the earlier one. So, any line between the black line and the red line that would also minimize this error or the absolute value of the error, so any line between the black line and the red line would minimize this quantity okay. So, clearly the red line is not a good fit to the data that is given there. So, a more you know accurate and possibly the only method that would do the job that we are looking for.

So, third one is that minimize the sum of the squares. So, importantly we have tried to minimize the error then the magnitude of the error and now we are trying to ah minimize the sum of the squares of this error which means that we are going to minimize this S r and S r is nothing but i equal to 1 to n and e i square which is nothing but i equal to 1 to n and y i measured say for example and y i by this model of linear regression.

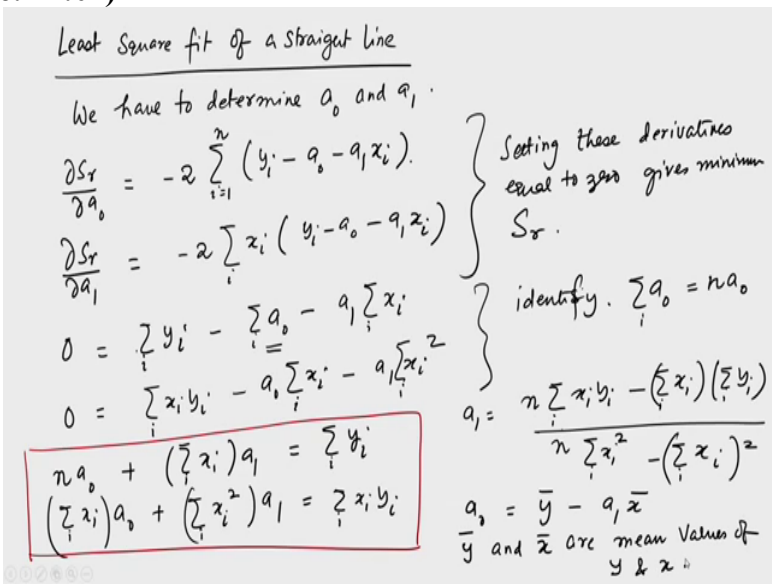
So, this and this is nothing but we need to actually talk about this 1 to n and we have a y i - a 0 - a 1 x i square okay. so, this S r needs to be minimized and if you minimize it turns out that if you minimize S r then you would the job would be done that is that would provide the best fit a linear fit of the data that we are trying to fit ok. It is important to note one thing here S r is actually a function of 2 parameters a 0 and a 1.

Most of the times in physics and in other engineering disciplines or even in other science disciplines one comes across a function which is only a function of a single variable. And when it is a function of single variable and we need to find the position of rather the value of x for which it is minimum okay it could be minimum or maximum which finally would have to be cross-checked with the double derivative whether it has a positive sign or a negative sign.

Because for a positive sign one actually talks about a minima and for negative sign one support a maximum because the double derivative actually measures the curvature of a function at that point. So if the curvature is positive that means we are at the minimum position and if the curvature is negative then we are at the maximum position. In any case most of the time we disregard that subtlety and we try when we try to look for a minimum condition we take a derivative of that function and put that function to 0 and find out at what value of x and that function has a minimum value.

Here importantly we have 2 variables which needs to be minimized in order to see the minimal minimum of these S r. So, we have a parameter space which is you know bigger than one variable it is a 2 variable parameter space but that really is not a problem.

(Refer Slide Time: 14:02)



Because we want to we can do a derivative we can take a perch partial derivative with respect to each one of those a 0 and a 1. And individually put them equal to 0 so that we know that the S r is equal to 0. So, but before that so let me claim almost a priori that this yields this third strategy yields a unique line for a given set of data points. So, let us look at the least square fit of a straight line. So, we have to determine a 0 and a 1 so the way we do that is that we take

partial derivatives of this S r with respect to a 0 which simply gives -2 sum over i equal to 1 to n and y i - a 0 - a 1 this is not i this is a 1 which is the slope and x i and so on.

And if you do it take a partial derivative with respect to a 1 then it is a -2 and there is a x i and it is a y i - a 0 - a 1 x i okay. So, we can set these setting these ah derivatives equal to 0 gives minimum S r okay. And if that is strategy then of course it is easy to do so we will put this derivative is equal to 0 so 0 equal to my; well I mean this 2 has no meaning then so you have a i and a y i - a sum over i a 0 - a 1 sum over i x i okay.

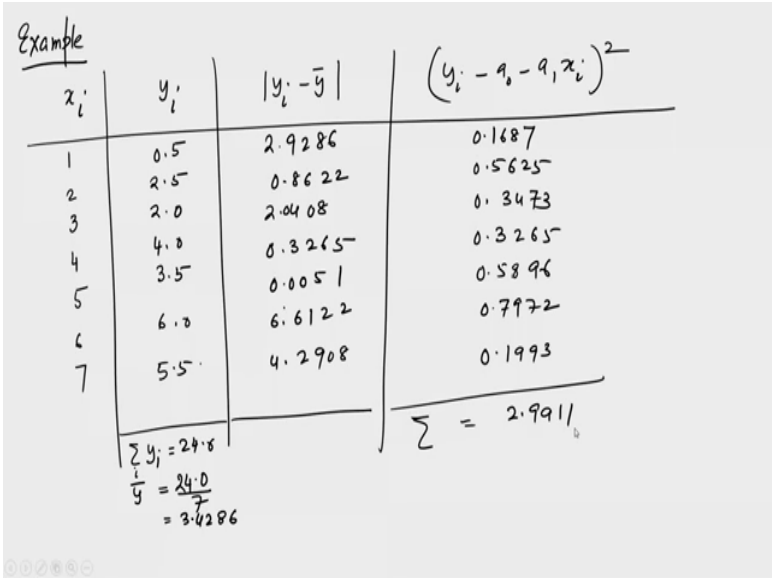
And then again I will put the left the second one equal to 0 so this is equal to sum over x i y i and the sum over i and - a 0 sum over x i and - a1 x i square and so on. Now identify that that this quantity that is some over a 0 sum over i is nothing but n a 0 where n is the total number of data points. So, then I can write down this slightly reorganize the equation and can write it as n a 0 + sum over i x i a 1, so this is i a 1 equal to sum over y i okay.

And there is a sum over i x i with a 0 + a sum over x i square and a 1, so if I miss these i's please understand that these mean that it is sum over i going from 1 to n and this is equal to nothing but x i y i okay. so, these are the 2 equations that we need to solve for the a 0 and a 1 which are the important ingredients to the line that we are trying to fit our data. So, these 2 things if you solve these 2 equations for the 2 variables a 0 and a 1 one gets a 1 equal to n sum over i x i y i and - sum over i x i sum over i y i divided by n x i square and x i square this square.

So, please make sure that you do not get confused so this is a sum over i x i y i so that is a product of each of the data points and then you sum over and this is you sum over all the data points and the I mean basically the corresponding values and then take a product so these 2 are not same. And in the similar way we are taking the product or rather the square of these sum of the squares and this is square of the sum okay.

So these are not same and the a 0 is actually nothing but a y - a 1 x where a y bar and x bar are mean values of y and x all right ok. So, this page or rather this slide summarizes everything that we have taken the square of these error and minimized it with respect to the 2 unknown parameters which are a 0 and a 1 and put them equal to 0 and there from there we have solved for a 0 and a 1 which comes in terms of the data points like this all right.

(Refer Slide Time: 21:29)



So, take a table so let us take an example it is a book example but you can take any example for that matter take an example that you have performed in your lab maybe the steam table for a for in your thermodynamics course or maybe a physics course simple pendulum experiment or for that matter in a chemistry lab when you do all these experiments of say measuring something with respect to time and so on.

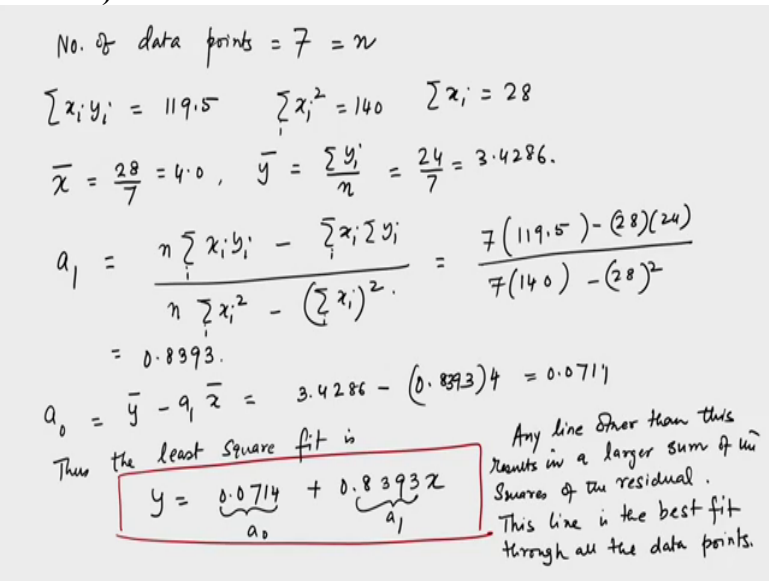
So you can take any of these data we are just taking an illustrative you know table. So, x i y i and because I would need it so I am computing this and then also I am computing the error okay. So, we have 1,2,3,4,5,6,7 and you have these corresponding values for the Y i 0.5, 2.5, 2.0, 4.0, 3.5, 6.0, 5.5 and so on. So, these are corresponding data points and I can easily calculate this y i so there is a sum over y i which comes out as 24.0 if you can add all these numbers it will come out as that.

And we can also get this y bar which is 24.0 divided by because there are 7 data points so we divided by 7 so these numbers so this is like a three .42686 will keep only 4 decimal places so 4286 okay so one can you can check that. Now each one of them I will subtract from each one of these y i values I will subtract this three .4286 and take a absolute value of that which is the mod that is shown here. So, this is like a 8.5765 so this is y i and minus of this well, so this has to be done so maybe this data we written is so it is a .5 and - 3.42.

But then you have to take this thing there so it should be so .5 and minus- this would be so .5 - 3.4286 would give me so this is like a 0.5 here 0.500 so this is like a 6 and 8 and 2 and so this is like a 9 and this is like a 2 this is a 2 and so on. And then of course it is negative but we will take this value so it is 0.9286 please see these things 2.9286 then we have this as .8622 and this is like 2.4408 0.3265 and .0051 and well a 1.0051 0.0051 and 6.6122, 4.2908 and so on.

So, I please see these data and should be fine even if I made some mistakes you should not repeat that to do it yourself. So, this y i - y bar and if you calculate these data so it is 0.1687 0.5625, 0.3473, 0.3265, 0.55896 and 0.7972 and 0.1993. And if you take the sum of these then it becomes equal to 2.9911 okay.

(Refer Slide Time: 27:09)



So, the reason that we did is following will show that, so here let us analyze what we have done and how we can calculate a best fit line of this. So, number of data points equal to 7 which is your value of n sum over x i y i equal to 119.5 and sum over i x i square equal to 140 and sum over x i equal to 28 these the values that are added from 1 to 7. So, is these values that are added ok we need all these x i y i and x i square and x i sum over all that.

Also we need because a 0 demands the knowledge of mean x so which is equal to 28 over 7 which is equal to 4 or we can write it as 4.0, y bar is nothing but sum over y i over n which is equal to 24 over 7 which has already been calculated 4286 and so on. So, this these are our preliminary calculations of all these average values of x i y i and x i Square and things like that, so if you plug in into the formula for a 1 which is the slope which is here in this formula if we plug in these things this formula that we plug in and this formula that we plug in then of course your a 1 becomes equal to n sum over i x i y i and sum over x i I have written that once again for your reference n x i square - i x i square okay.

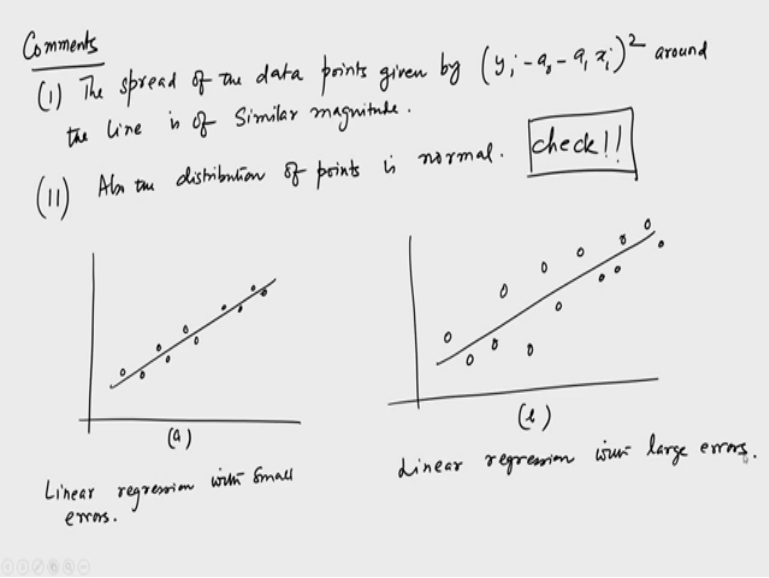
So, if you put in all the values that we have obtained 119.5 because n ah ah is equal to 7 and x i y i has a value 119.5, so this is equal to 28 x 24 and this is like 7 into 140 ah n - 28 square which is the x i and all that so this taking the square of that and if you simplify this, this comes out as 0 .8393 and a 0 is simply equal to 3.42 so a 0 is if you want me to write down its y bar - a

1 x bar so this is equal to 3.4286 - 8393 9 multiplied by 4 which is equal to .0 714 please see these calculations.

So, then does the least square fit y equal to 0.0714+ 0.8393 sorry okay, so this is equal to your a 0 and this is equal to your slope a 1 and this is the best fit line for that particular table of data that we have seen. But how do we know that how do we know that this is the best fit line this is the intercept and this is the slope that will fit the data that is that are given and that fits it I mean this line fits those data the best.

So, of course I mean goes without saying that any other line any line other than these results in a larger sum of the squares of the residual, so the calculated this line is the best fit through all the data points. So, this is the best fit line through all the data points.

(Refer Slide Time: 33:06)



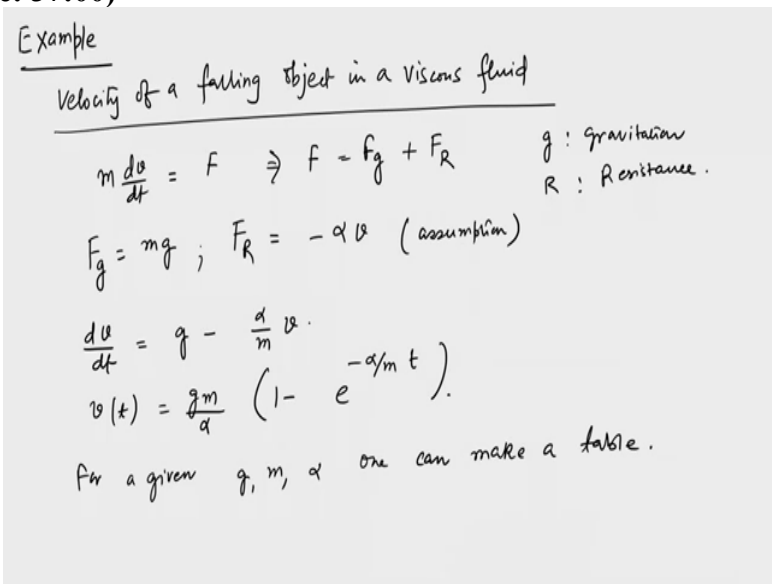
how do we say that we will see it in a while but let me make these comments a priori one is that so these are comments, one is that the spread of the data points given by y i - a 0 - a 1 x i square around the line around the line is of similar magnitude. Let me see what I mean by that what I mean is the following that if you look at this last column, let us this laser pointer if you look at this last column you see that all are between 0 and 1 and rather it is you know centered around some values which are like 0.3 and so there are some values like .79 to some 1.16 and all that.

But there is no number which is like of the order of 1 and the other is of the order of 10 so they are of the same order. And this is a feature that we have been able to minimize the residual error some of the residual error okay so they are similar in magnitude. And the second thing is that we shall not elaborate here but we will do another example in which will show this also the distribution of points is normal.

So it corresponds to a normal distribution which is what we have discussed earlier. So, this is a check that you should do but we will give you another check. If you look at 2 plots, so you have a line and you have data point like this and so on. So, this is called let us call this as a this called as a linear regression with small errors. And take another one these are freehand drawing so I just wanted to make the point here.

So, this is like and so on so this is b this is again the linear regression with relatively large errors, I mean larger than the plot in a. Let us see a physical problem where you can apply it okay.

(Refer Slide Time: 37:00)



So, a velocity of a falling object in a viscous fluid okay. This could be just in a parachutist falling down or maybe a ball falling down through paraffin or honey. So, it could be anything so we write down the Newton's equation of motion as m dv dt it is equal to F. And the F is a force that acts this is equal to due to gravity so we will write it with a g and our resistance force because I have said it is a viscous fluid so we will write it with R.

So, g corresponds to gravitation and R corresponds to resistance. so F g equal to mg, F R equal to - alpha V so this is an assumption, this is the most you know simple assumption that the resistive force that the body experience experiences while you know are traversing through the viscous fluid it is proportional to the velocity it could be velocity square or even you know larger power of the velocity.

But for a small viscosity that is when we talk about normal fluids we usually take it as a single power of V this V is the velocity. And so we can write down as dv dt this is equal to a g - alpha by m into V and so V of t can be gm over alpha and 1 - exponential - alpha by mt okay. We

certainly do not have a linear equation we have an equation which is has a very non linear dependence see as a function of t.

So if you want to measure the velocity of the particle at you know discrete time steps and make a table out of it and want to fit it with a linear regression you could see what happens. I am not doing this problem but posing a question to you that is linear regression at all a good approximation for this for this the data that you would be generating from this equation or it is completely it does not make sense at all to have a linear fit of this.

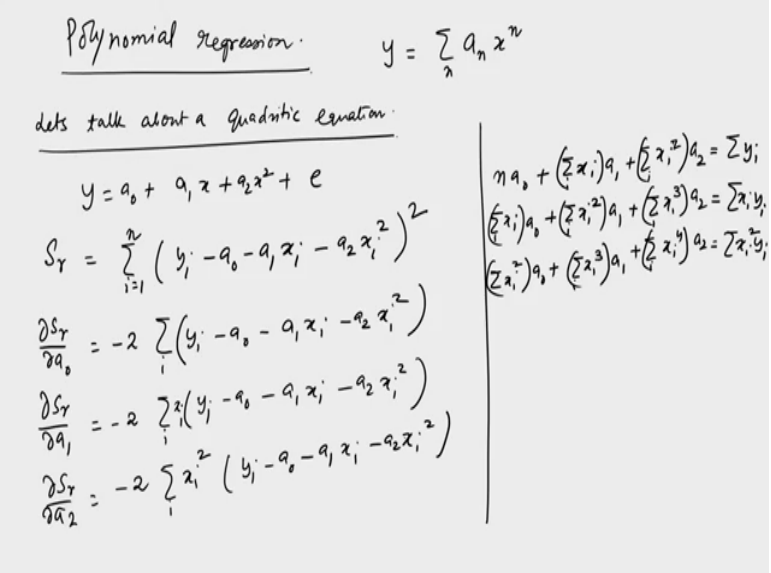
Or it makes sense in some part of the plot that it generates it only makes sense in that part of the plot and does not make sense in other parts of the plot please think over this and so on. So, but you can you can calculate this or rather put some discrete time steps let t equal to 1 second, 3 seconds, 5 seconds, 7 seconds and all that and calculate V by putting and knowing all these values of g and g's of course the acceleration due to gravity which is 9.8 meter per second Square and alpha is the coefficient of the viscosity of the viscous coefficient which would but for a particular fluid it will have a value.

And mass m is the mass of the particle which is of course would be given to you. So, for given g, m, alpha one can make a table okay. So, of course it looks like to you that maybe when these alpha over m to be a very small quantity then a linear approximation would hold that is if the mass is small and the Alpha is very large then of course or rather it is lesson talk about the values per say.

But let us say that we can make a linear approximation out of these things then we can have this of course if alpha is very large the motion would stop almost immediately but if you expand this exponential. And keep only the linear term then you will have a linear equation. So, the question that I am posing to you is that that in what regime these linear fit would hold and after that your linear fit would no longer be a good approximation to fit the data or good model to fit your data.

So, this makes us go to the next level of fitting that is we can do a instead of a linear regression we can do a polynomial regression that is invoke larger power of x. And as we will see that the problem is not too difficult it is an in fact a trivial problem of getting or rather raising the difficulty to a polynomial regression.

(Refer Slide Time: 42:58)



So, this is a polynomial regression, so now and polynomial is as you know that is equal to it involves all powers of x that is a 0 + a 1 x + a 2 x square + a 3 x cube + a 4 x to the power 5 and so on. So, this is like y equal to something like n a n x n so n will vary from you know 0 to some power that you are interested in but let us just talk about a quadratic equation okay. So, we have y equal to a 0 + a 1 x + a 2 x square and + e that is the sum of the residuals.

And we are going to take the same route as we have taken earlier that is we are going to minimize the sum of the residuals that is we are going to minimize this quantity. Now we have a quadratic term that is extra and that of course raises the problem to one order more or rather it increases the size of the problem to an order more. So, it is x i and - a 2 x i square and this has to be minimized with respect to so this is the square of this has to be minimized with respect to.

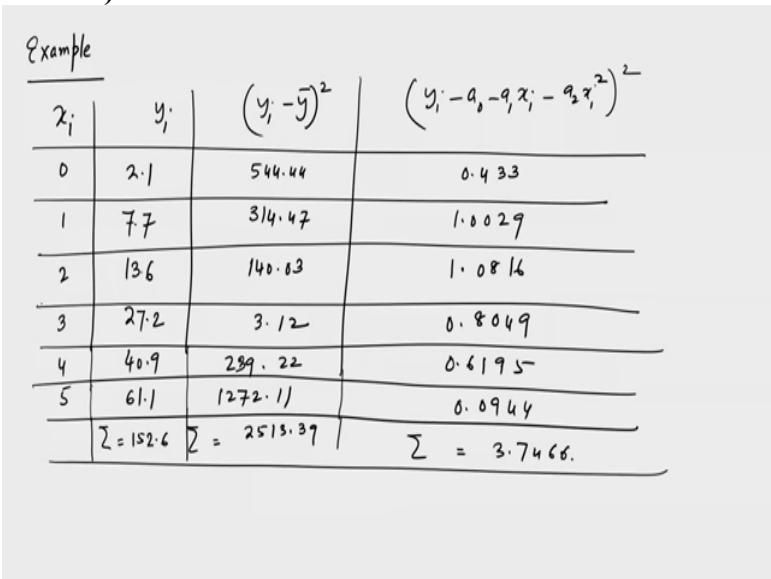
Now 3 parameters namely the a 0, a 1 and a 2 okay so we have three equations so again following the same procedure of minimizing them it is this sum over i and a y i is a 0 a 1 x i - a 2 x i square ah del S r del a 1 it is equal to - 2 sum over i x i y i – a 0 – a 1 x i - a 2 x i square del S r del a 2 it is equal to - this x i square y i - a 0 a 1 x i - a 2 x i square, so, this is the 3 equations that you have to solve.

And if you set you know your sum over a 0 equal to n a 0 so we get 3 equations which are like n a0 + sum over i x i a 1 + sum over i x i square a 2 this is equal to sum over i y i and this is sum over i x i a 0, so this sum over i means the sum goes from 1 to n + x i square a 1 and a + x i cube a 2 this is equal to nothing but x i y i and we have x i square a 0 + x i cube a 1 + x i for a - it is equal to x i square y i okay.

So these are you know these are simple you can easily figure out these equations excise where y i and so on. So, these are again 3 equations and with 3 unknowns and usually we do not try to

solve it by this elimination method or rather this substitution method what we can do is that we can write down sort of code for Gauss elimination which we will see later. And then calculate these a 0, a 1 and a 2 and take a quick example of that here.

(Refer Slide Time: 48:09)

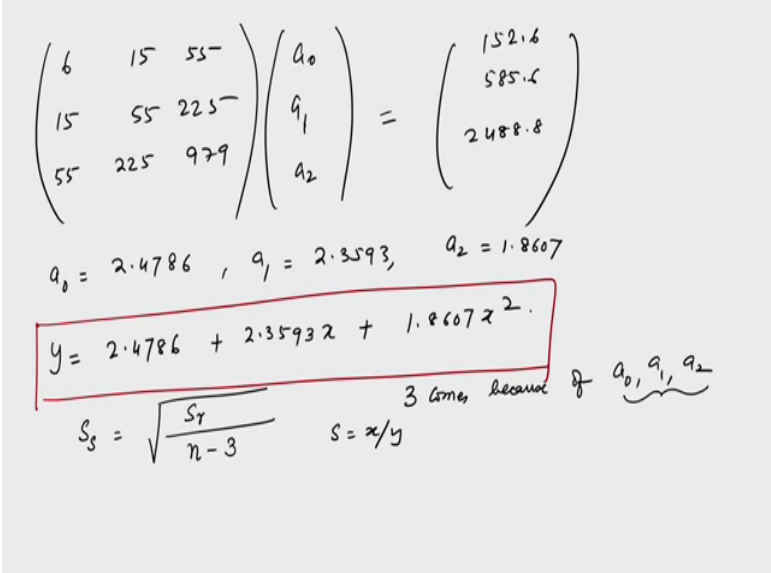


So, we will convert this equation into a matrix equation and solve it using Gauss elimination method and as I said earlier that will see this usage and applicability of this method later. So, an example so we have a x i we have a yi we have a yi - y Square and we have a yi - a 0 a 1 x i - a 2 x i squared and the whole square of that so then and maybe we have just 5 data points. So, 1, 2, 3, 4, 5 and here we write down the results of that.

So, this is a 0, 1, 2 okay we have 6 data points 3, 4, 5 and we need a little more space here in order to calculate these sums and all that so y i is like 2.1, 7.7some experiments have given this 13.6, 27.2, 20.9, 61.1 okay. So, if you sum this comes out as 152.6, so this is like 544.44, 314.47 and 140.03 there is 3.12, 239.22 and 1272.11 and so this sum of this is equal to 2513.9 now this is 0.4331.0029 and 1.0816 and .8049 and these are .6195 there is a .0944 and if you take the sum of that this becomes 3.7466.

So, using these 3 equations that we have written down in the previous slide let us try to solve this so these are these excise and and all that.

(Refer Slide Time: 51:15)



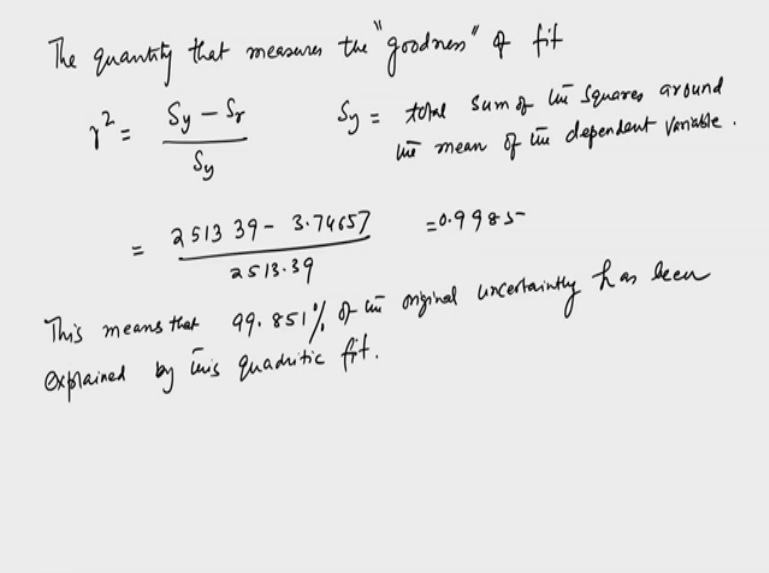
So, it is 6, 15, 55, 15, 55, 225 sorry 55, 225, 979 and this is a 0, a 1 and a 2 and this is equal to 152.6, 585.6 2488.8 so these are these things which are y i, x i, y i and x i square y i so those had to be calculated and so on. So, this is these are the 3 equations that one need to solve and at this moment. I am not getting into the details of the solution but it can be done by a variety of methods and one of them being Gauss elimination.

So one can get a a 0 equal to 2.4786, a 1 equal to 2.3593 and a 2 equal to 1.8607 so what we get is that we get a quadratic fit that is a quadratic line which fits the data is 2.47857 + 2.35 that is a slope 93 so 86 let us call it simply 86 93 x and the + 1.8607 x square so this is the best fit to the line. Now why is this the, best fit to the line? Let us see that so we can calculate the standard deviation of this data.

And the standard deviation I am not going into details of that but the standard deviation for this particular case can be calculated as so it is S s is equal to S r that is the residual error S r that we have defined and it is n – 3, now this 3 comes in the denominator because of a 0, a 1 and a 2 these are 3 in number so we actually lost 3 degrees of freedom and that is y it is n – 3. And this S actually can be x or y depending on which variable is a you know the dependent variable and the independent variable it is rather than the dependent variable.

So, in this particular case S is equal to y so here if we calculate this S r which we have already calculated here that is this one 3.7466 which is the sum of the residual squares, so if I put it here 3.7466 and then n is equal to 6 because it goes from 0 to 5 and it is this and then if you do this calculation then it comes out as 1.12, so that is the standard deviation but that does not say much it only says that the standard deviation is quite small.

(Refer Slide Time: 55:00)



But however what is more important in this particular context is calculating a quantity called as r or which the quantity that measures goodness of data that is how goodness-of-fit rather not data or the data versus the fit is computed by a quantity called as a r squared which is that S y we are writing it explicitly it is actually S s the one that is introduced in the last slide, so S is equal to y here a - S r / S y.

So S y is nothing but the total sum of the squares around the mean of the dependent variable, so this turns out that it is like 2513 we have calculated this already .39- 3.74657 / 2513 .39 so this is equal to 0.99851 which means that this means that 99.851% of the original uncertainty has been explained by this quadratic fit okay. And which means that it is a very good approximation this quadratic approximation of this quadratic regression that we have done is a very good approximation to the data that is shown there.

In fact what you could do is that you could do a linear regression for this just to remember one point here the linear regression would have a n - 2 because this 2 is the number of these parameters that we actually find out and then you use the same formulas here and you can see that the quadratic fit is a much better approximation of the quadratic regression is a much better approximation for the data in this particular case in.

The similar fashion the polynomial regression can be built up every time you increase the power that is if you want to do a cubic fit of this data your order of equations go up by one that is the number of equations that you need to solve go up by one. And in principle is not a problem numerically but then writing it over and over again for a large you know matrix would be problem. But then you need to do that if you really want to fit very sort of scattered data with a polynomial having a large degree.

So, I believe that I have been able to explain that why these linear regressions and quadratic regressions are done for a set of data that you have found either by experiments or by numerical simulations. And you need to know the analytic behaviour of the data. Suppose the data is actually diverging now of course in your linear fit or a quadratic fit it is hard to catch. But then you can do you know these kind of fits in present or rather than in the vicinity of the points where it is showing an analytic behaviour.

And you want to know what is the nature of the analytic behaviour if it is a diverging quantity then how's it diverging what kind of singularity exist is it a removable or a non removable singularity. And all these questions are important both in physics and science in other branches of science and in engineering that we need to understand the behaviour of a function from the collection of data that is the you know that is the motive behind doing all this exercise.