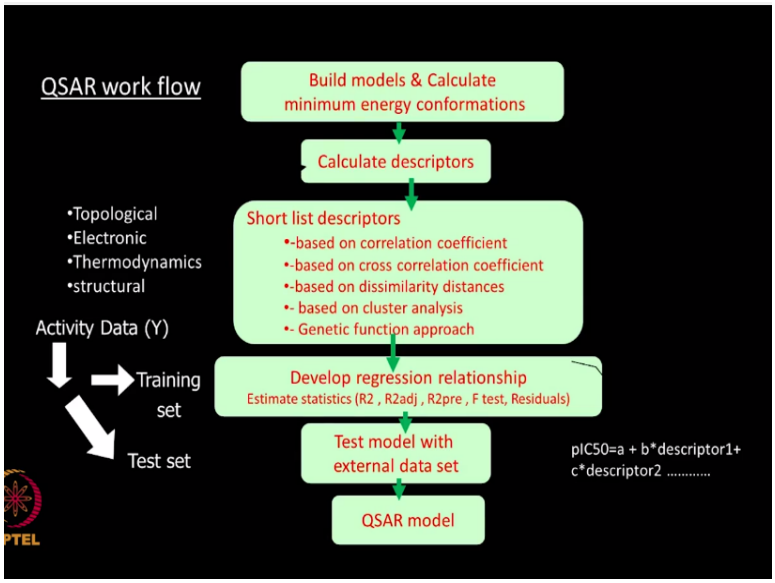
Computer Aided Drug Design
Prof. Mukesh Doble
Department of Biotechnology
Indian Institute of Technology - Madras

Lecture - 29
Quantitative Structure Activity Relationship (QSAR)

Hello everyone, welcome to the course on computer-aided drug design. We will continue on the topic of QSAR because QSAR is very, very important which gives you the structure activity relationship mainly between the descriptors or structural features and the activity.

(Refer Slide Time: 00:33)



I showed you this in the previous class. So you have certain molecules okay whose activities are known and their new energy conformations are known so we use different softwares. I showed you different softwares where you can calculate descriptors. I will show you some more softwares later. Now you have large number of descriptors okay. So you can shortlist descriptors based on the correlation coefficient between a particular descriptor and the activity okay.

And please remember if you have 5 data points you can think of having only one descriptor model okay. So if you have 5 data points, you should not go beyond one descriptor model. If you have 10 then you can have 2 descriptors or 2 independent variables that is x1 and x2. Then, those descriptors should not be cross correlated that is very, very important because some descriptors like molecular weight and size of the molecule maybe related to each other.

So they should not be cross correlated. They should be very dissimilar between each other okay. So we can do different approaches that you select the best descriptors, which are far apart, not correlated and so on. This is valid only when you are talking about more than one x okay. If you are thinking of only one x then these things do not matter. Once you do that we develop regression relationship okay, we estimate lot of statistics.

These are statistics, we will talk about in the course of next one or two lectures and then see how good if it is okay, so you develop a regression relation but parallelly you divide the data set into training set and test set okay. So if you have say 10 data points you do not take all the 10 data points and develop the regression model, around 20% of the data you call it test set, keep 80% of the data for the training and then you develop the regression relation using that 80% data okay.

Then, once you have a good model you try to predict for those 20% data which you have kept aside and see how good the fit is okay that is why you divide the data into training and test set, 80% of the data in training, 20% in the test set that is how you need to do remember that. Even if you have 6 or 7 data points, take 5 data points for your training and keep the other two for test.

Once you develop a regression model and then once you are happy with all these, try to predict those two data points which are kept aside, the activity of those two data points and see how good the prediction is because ultimately QSAR is meant for predicting activity of unknown new molecule okay that is very, very important. Once you have done it, you are very happy and you have a QSAR.

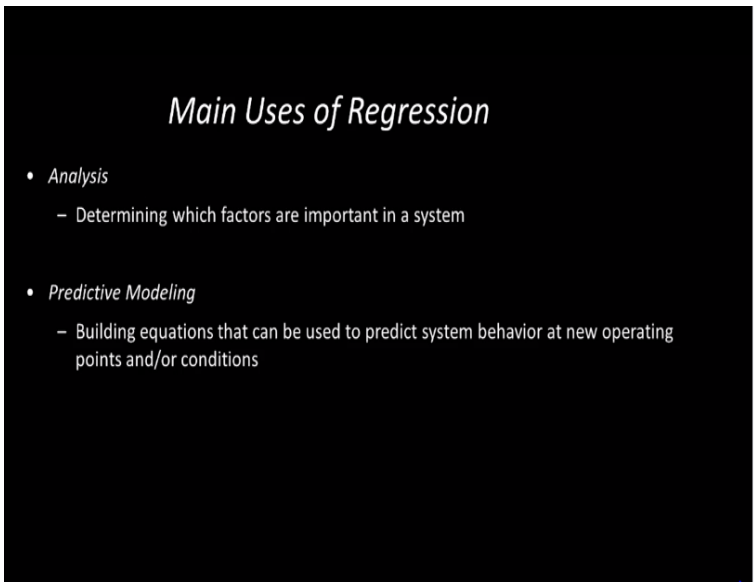
You can use this relationship for predicting new molecules, activity of new molecules and so on actually. This is how the work flow happens. So the main challenge is to reduce the number of descriptors and take those descriptors, which you think are very good like I showed you yesterday a software called DRAGON, which predicts almost 2000 descriptors and you do not know which one to select, which one is the best for you okay that is very, very important.

That is why that is the challenge in QSAR, sometimes we may end up selecting a wrong descriptors or descriptor, we may miss out the important ones, so those things are there. So

descriptors, Topological, electronic, thermodynamics, structural, zero dimensional, 1-dimensional, 2-dimensional, 3-dimensional properties and so on actually okay. So the QSAR the main challenge is shortlisting the descriptors that is very, very important.

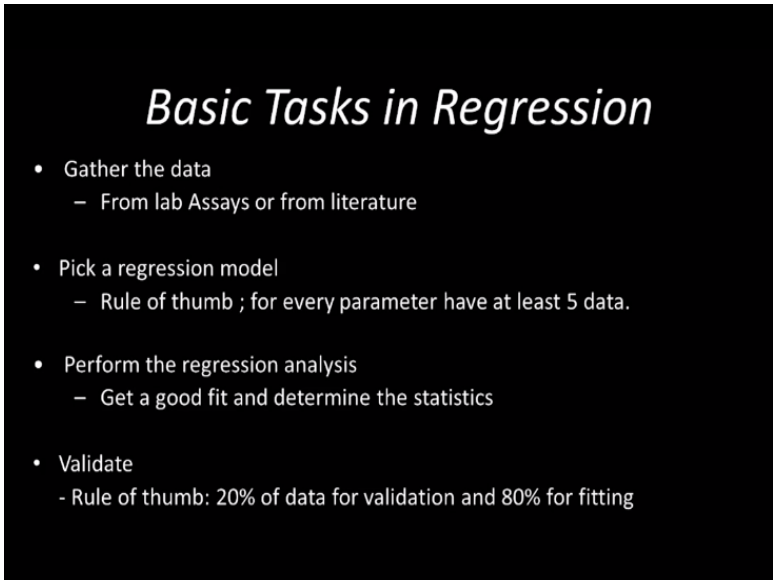
And also I said we need to have some experimental data activity for some molecules. So we can go to literature and try to collect activity so before I start synthesizing new molecule I may go and look into the literature and see whether activities have been reported for molecules of similar structural features and then collect them and develop your QSAR and then proceed further okay.

(Refer Slide Time: 04:52)



So what is the main use of this regression? QSAR is nothing but regression determining which factors are important in a system okay. Building equations that can be used to predict behaviour for new molecules, new operating points or new conditions okay that is the advantage. So I can use QSAR, QSTR, QSPR either for property or toxicity and so on okay.

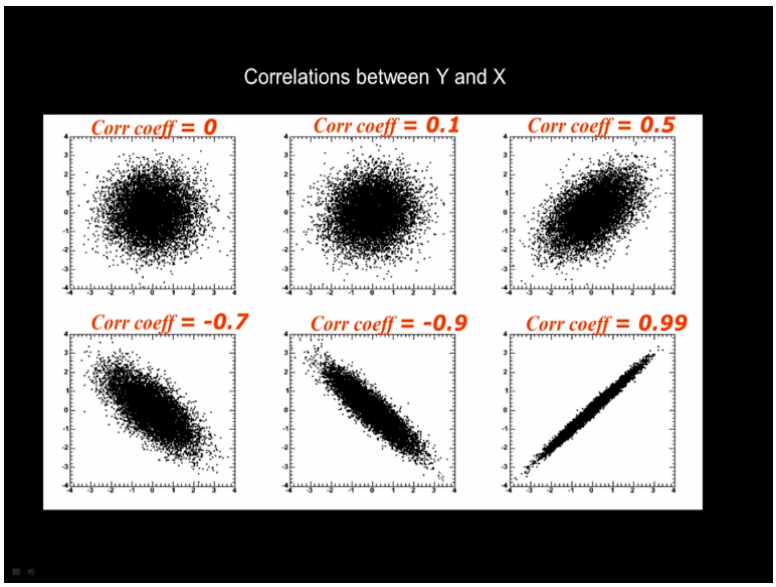
(Refer Slide Time: 05:16)



So initially we may need to have data from literature, from lab Assays and then develop the regression model. So like I said you must have at least 5 data points for one parameter model or one descriptor model, perform the regression, see whether the fit is good and determine the statistics like R-squared, R-squared adjusted. We will talk about what the definitions of all these.

The rule of thumb is 20% of the data for validation or test, 80% for training or fitting the data that is very important.

(Refer Slide Time: 05:52)



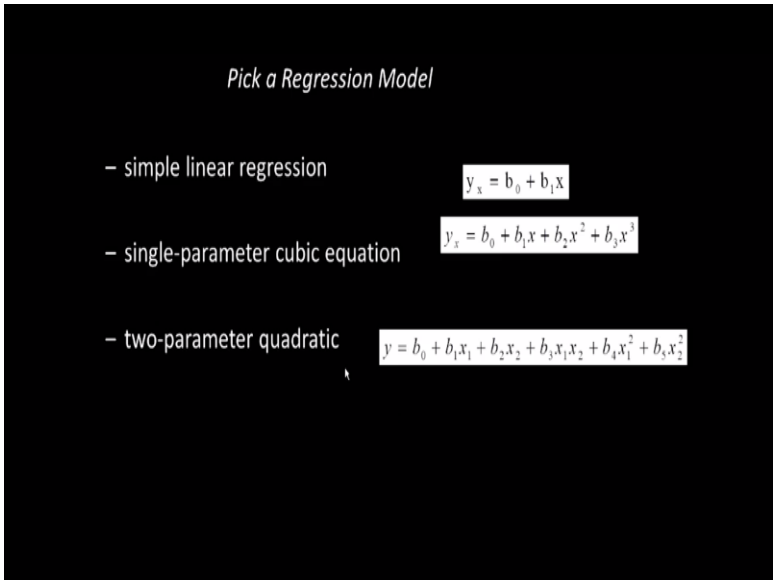
So you can select the x which seems to have good correlation with y, y is your activity, x is your descriptor. So if there is no correlation between y and x that is y axis is y, x axis is x that is activity and descriptor on the x axis and if they are not correlated it may look like this,

slight correlation, reasonably good correlation, very good correlation okay. So as x descriptor value increases activity also increases.

You can also have a negative very good correlation like this, so as the descriptor value increases activity maybe going down. For example, lipophilic systems may not work on hydrophilic acids okay so as the lipophilicity increases or Log P increases, the activity maybe going down. So you can have both negative as well as positive value. This is the ultimate, the best okay that you can have it between like this okay.

So you can find out whether initially the correlation coefficient between a descriptor and the activity falling between -1 to +1 and larger closer towards these 1 or -1 you can see it is fantastic even 0.5 and above 0.6 is also good. So we can select as the descriptors, which appear to have good correlation coefficient okay.

(Refer Slide Time: 07:18)



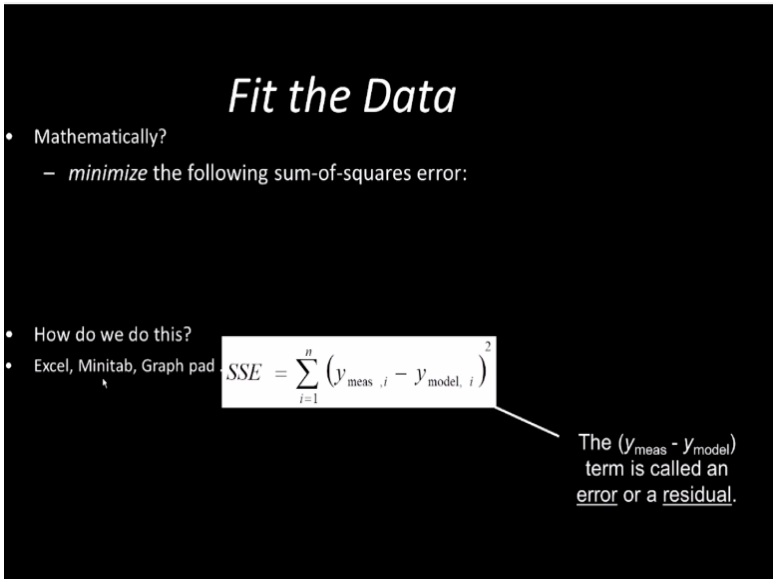
So we can start with a simple linear regression okay. So we can have say activity on this side, one constant+b1*x like I showed you one example yesterday with x could be your molecular volume or dipole movement okay so it could be molecular weight and so on actually okay or Log P. You can also have single parameter square, Log P square, Log P cube. Generally, we will not have cube and all but we can have square in certain situations.

Of course, remember you may need to have more data points so that you have enough degrees of freedom. If you want to have square type of model because there are 3 unknowns here. We can also have 2 parameter model, we can have Log P and say electrostatic okay

sigma coming here so again as you can see the more number of parameters more complicated the quadratic.

You have so many unknowns okay so you need to have a lot of data points so that the degree of freedom is satisfied.

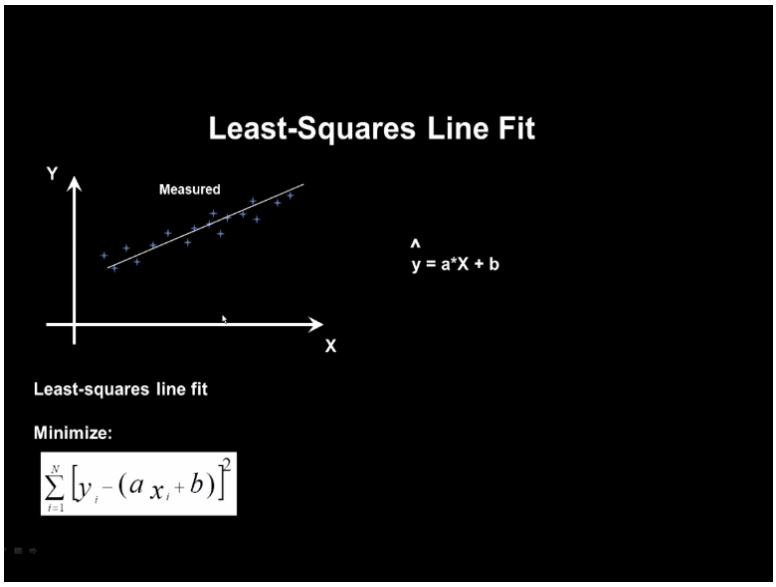
(Refer Slide Time: 08:20)



So basically little bit of statistics, so you are trying to minimize the sum of squares, so y is the activity as measured that is experimental y model is the prediction of the regression. So this is squared, you are taking sum and then you are trying to reduce this, this is what it is all about actually, you are trying to minimize the sum of squares and this is called error or residual because this is measured activity and this is as predicted by the model.

So this should be as low as possible, you are squaring it so that – or + it will always be positive and then you are taking a summation okay. This is sum of squares of error or residual that is called okay so there are lot of softwares, which can do that.

(Refer Slide Time: 09:08)

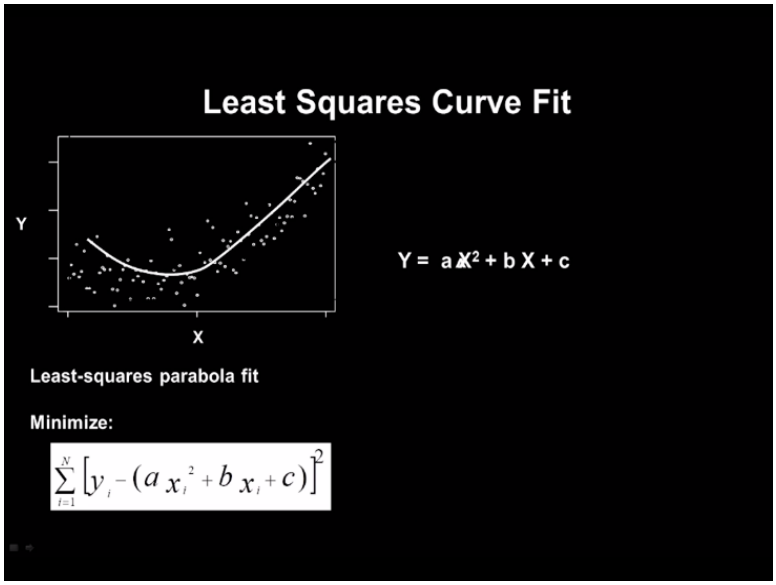


So when you are doing a one parameter model or one descriptor model, you have y I have put hat because this is called prediction normally you put a hat y=a*x+b, x is your descriptor, Log P or molecular volume or molecular weight or dipole movement whatever it is, so you are trying to get a least square line fit. It is called as least square because you are trying to minimize the square here right that is why it is called least square.

So if your model is ax+b then you are trying to minimize yi is the prediction, ax+b is the model, so we are squaring it up, difference you are squaring it up, taking a summation you are trying to minimize this okay. This is what you are doing. So if you want to minimize a function what do we do? We take a differential with respect to a equate it to 0, then take a differential with respect to b equate it to 0.

And then we have 2 equations, 2 unknowns, we will calculate for a and b that is what you do if you remember minimization and your calculus.

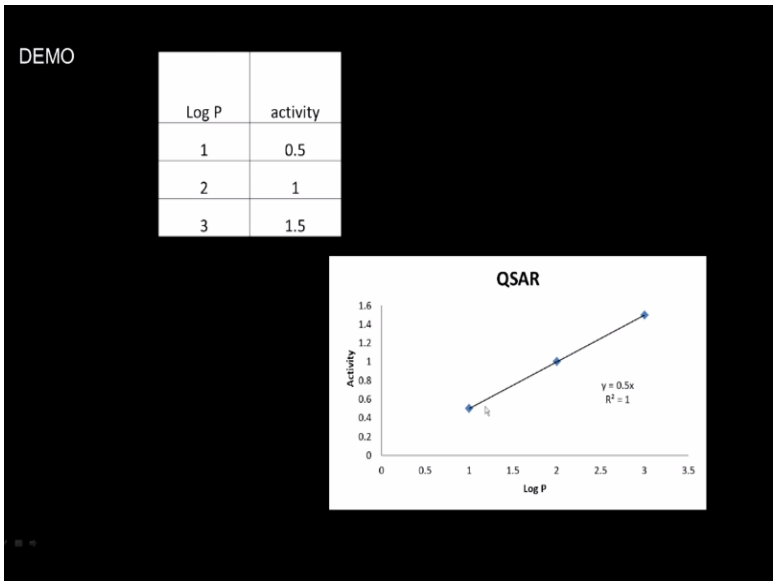
(Refer Slide Time: 10:24)



Similarly, you can have a square okay y=ax square+bx+c so x could be Log P so you could be having Log P square and Log P here so we have 3 parameters. Again, you are trying to minimize, yi is your actual experimental data, this is your prediction okay so difference, squaring it up, summation you are trying to minimize this. So if you want to minimize this what do we do?

We take a differential with respect to a equate it to 0, differential with respect to b equate it to 0, differential with respect to c equate it to 0 then estimate a, b, c because we have 3 equations, 3 unknowns okay. Basically, this is how it may look like okay.

(Refer Slide Time: 11:08)

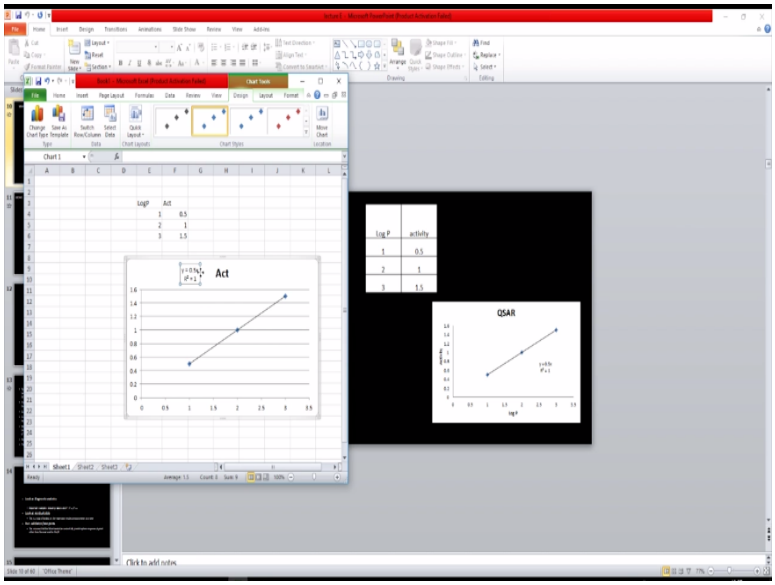


So for example if I have yesterday the previous class also I showed you how to calculate if it is a simple, it is very, very easy to do okay using excel. We do not have to sweat it out at all.

So we do this, so this is Log P for example we call it Log P 1, 2, 3 so we say activity then 0.5, 1, 1.5 okay so this is very, very simple. We can use the graph here. So we got it here right what is shown here so we can get the regression relationship, add a trend line then we can say display equation and R square as you can see the slope is 0, it passes right through the origin and y activity=0.5x, R square is 1.

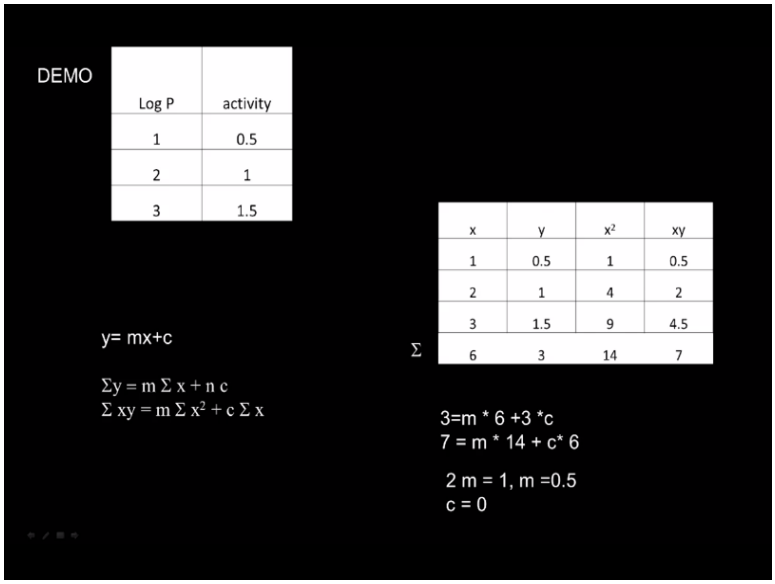
I have given a very simple looking data so that is why R square comes to be 1 but in real life R square maybe 0.7, 0.8 and 0.9. It will never become 1 actually okay.

(Refer Slide Time: 12:48)



So simple, if you have a single descriptor model anybody can do it using excel, there are many softwares also for us to do okay.

(Refer Slide Time: 13:01)

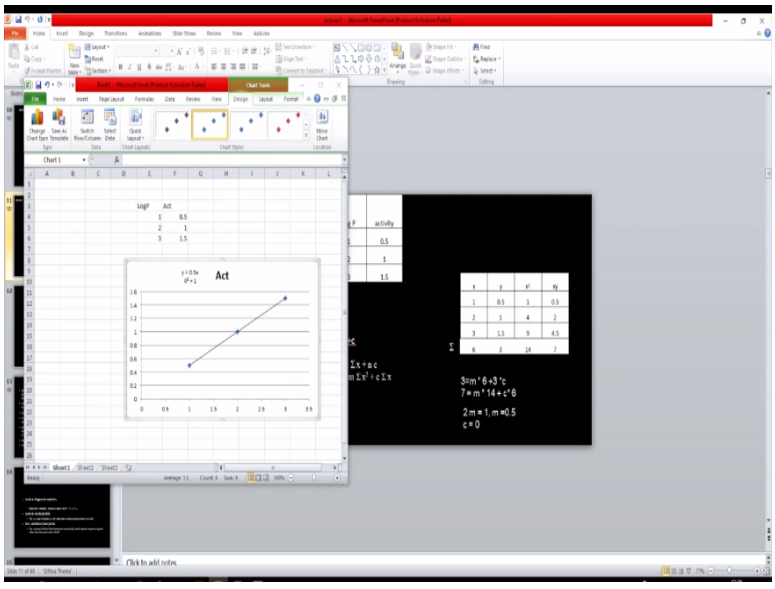


So y=mx+c that is the equation you are trying to solve. So the two equations if I use calculus like I mentioned and minimization we get summation of y=m*summation of x+n*c, n is the number of data points, here it will be 3. Other equation will be summation of xy=m*summation of x square+c summation of x. Basically, you are taking this equation, doing the summation all over.

And then you are multiplying by x, again taking summation. So you have 2 equations here, 2 unknowns m and c, so we can calculate and so what do we do x=1, 2, 3; y is 0.5, 1, 1.5. So we need x square right so x square is 1*1=1, 2*2=4, 3*3=9 here. We need this here xy, xy=1*0.5 is 0.5, 2*1 is 2, 3*1.5 is 4.5 so we are trying to sum it. Summation of x=1+2+3 that is 6, summation of y=3, summation of x square is 14, summation of xy=7.

So now we can substitute here. So this will become 3=m*6+3*c and this is 7=m*14+c*6. So we will have 2 equations okay that is what this is. We will have 2 equations for m and c. We can solve this okay. So what do you get? We get m=0.5 and c=0. In fact, that is what we got in our excel also right.

(Refer Slide Time: 14:50)

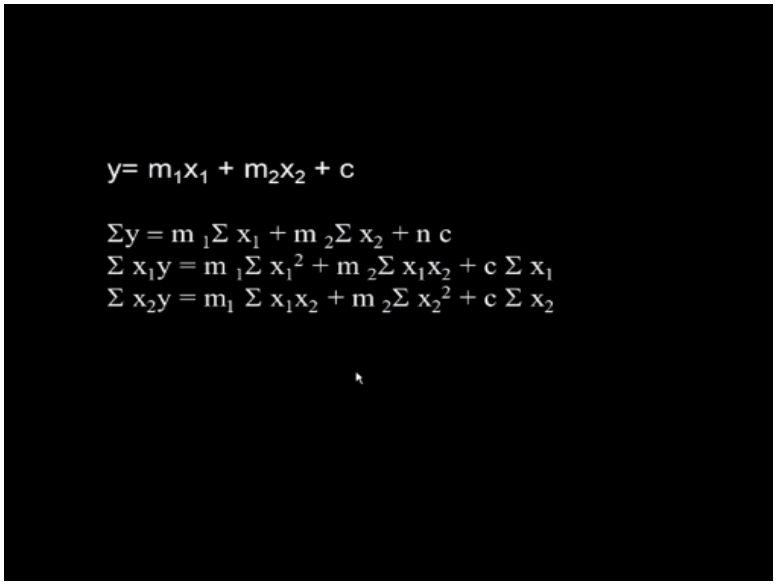


That is what we got in our excel also okay m=0.5, c=0 so this is the principle by which this particular thing works okay. So if you do not have excel, you can do it this is the basic by which it works actually so summation of y=m*summation of x+nc, summation of xy=m*summation of x square+c summation of x. Basically, you are taking this and getting this equation and you are multiplying by x and then you are getting this equation.

So summation of x is nothing but adding up all the x values, summation of y is adding up all the y values, summation of xy is nothing but calculate x square and then adding all of them, summation xy=x*y then adding all these, so you develop 2 equations, 2 unknowns m and c, you can solve it and you end up with c=0, m=0.5 and that is what you got in your excel also correct.

So this is the principle by which you get the m and c. So if you have one parameter, it is very simple with excel you can draw the graph and then calculate the regression equation and R square. If you do not have excel you can use this formula to calculate the m and c okay. This is the principle by which it works okay.

(Refer Slide Time: 16:32)



So what happens if you have 2 variables that means 2 descriptors x1 and x2. So you have m1, m2 two slopes and c, so you have 3 parameters. So basically you need 3 equations so how do you get the 3 equations? This is the first equation, summation of y=m1 summation of x1+m2 summation of x2+nc, the other one will be multiply by x1 that is you get this equation x1y=m1x1 square+m2 summation of x1x2+c summation x1.

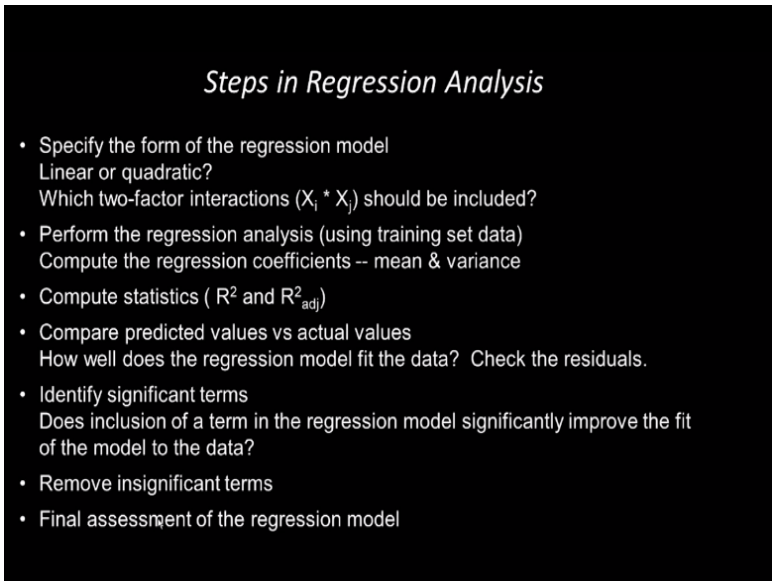
And then multiply with x2 so you get summation of x2y=m1 summation x1x2+m2 summation x2 square+cx2 so you see you have now 3 equations and 3 unknowns, m1, m2 and c. So we can solve this and get these 3 constants. So then your QSAR will be this, this is a 2 descriptor problem okay. So descriptor one could be Log P and descriptor 2 could be sigma or some electrostatic parameter here okay.

This is your QSAR equation. Of course, when you have this type of model when compared to one descriptor model you should have at least 10 data points. So as per rule of thumb 5 data points you should stick to only one descriptor model, approximately 10 data points you can go for 2 descriptor model okay at least 8 data points you should have.

Of course, that is not correct because if you want to divide the data into training and test set so even if you have 7 data points and 2 of the data points we can put it as test set so 5 data points we can develop a single model. So if you have 10 data points so 2 of them you can keep it for test set so with 8 data points we can may be develop a 2 descriptor model and then try to predict for the other 2 data points the activity value.

So this is how it is very close, so ideally you should go to 10, 11, 12 like that so that we can have a 2 descriptor model. Remember that we should have enough data points so that we have enough number of data points in the test set as well as the training set.

(Refer Slide Time: 18:57)

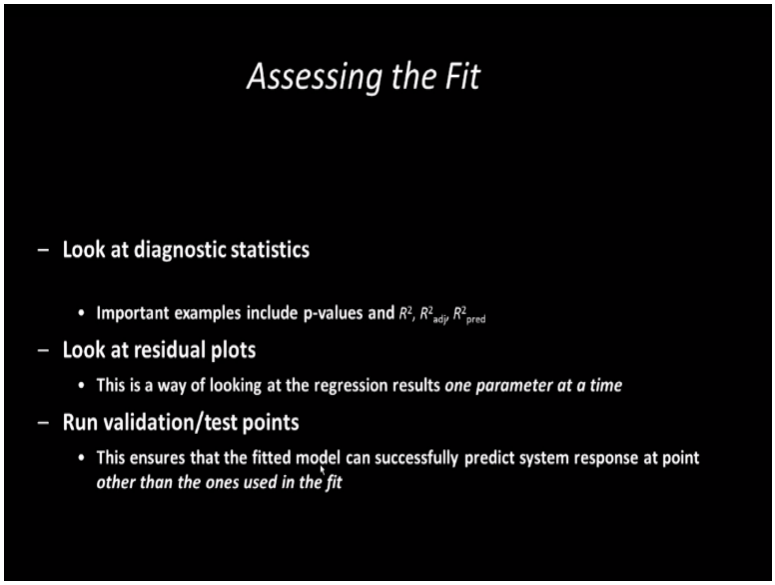


So what are the steps in regression which I showed you? Specify the form the regression model linear or quadratic, Xi to Xj which one to include so they should not be cross correlated that is very, very important okay. Perform the regression analysis using training set data, compute so many parameters, mean, variance, R square, R square adjusted, R square predicted, Q square we are going to do that.

Compare the predicted versus actual values, how well the regression fits, check the residuals. Residual is that error y experiment-y predicted that is called residuals and see how they look

like. Identify the important ones, there may be some descriptors which may not be very important. Then, we can remove this okay, we can take only those which are important and then again fit the data like that we can do actually. Remove insignificant terms okay. Then look at the final regression model.

(Refer Slide Time: 20:02)

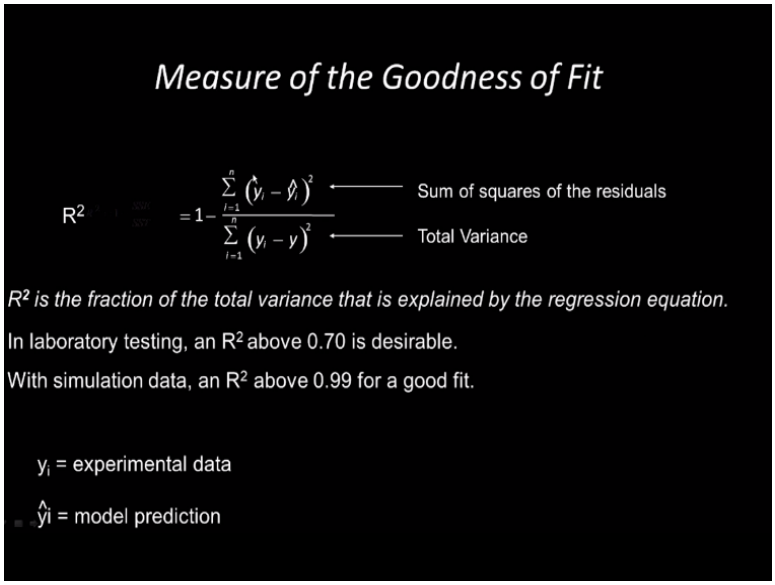


This is how it goes, so these diagnostic statistics that is R square, R square adjusted, R square predicted sometimes some people use Q square so many things. Look at residual plots then run the test data. These 3 are very important when you develop a QSAR. Do not think QSAR is just a mathematical fit using a software and that is it. It is not correct, you must look at these p-values, R square, R square adjusted, R square predicted, Q square and so on.

Look at the residuals, see whether there is some pattern, there is some problem in the residuals. Then, run the test set or validation set and see whether the model is able to predict the data, then only you can say your QSAR is good until then you cannot trust the QSAR. Please remember QSAR is not fitting some points using a regression line or quadratic model and then that is it.

We need to spend some time, think and see whether there are some hidden mistakes okay. I am going to show many of them where you can go for.

(Refer Slide Time: 21:14)



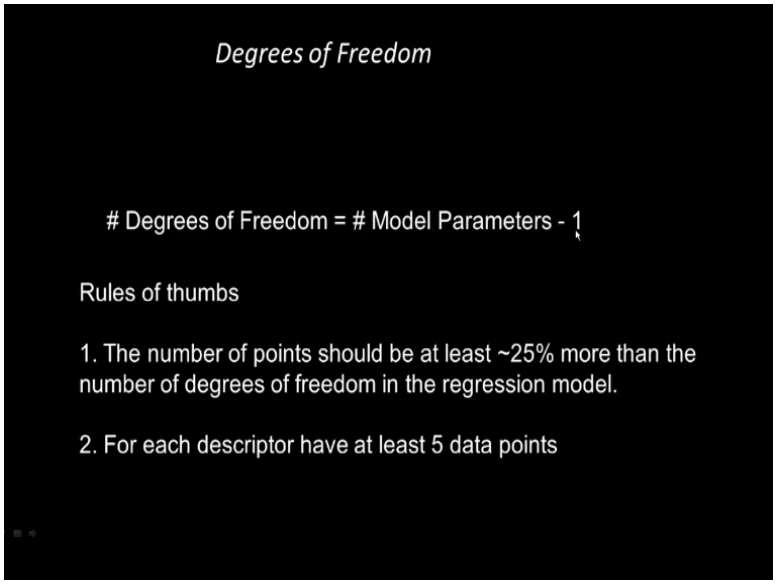
Now let us look at these statistics. We call it goodness of fit R, R we know is the correlation coefficient, there is something called R square is the fraction of the total variance that is expected by the regression equation. So yi-yi hat, yi is your actual experimental, y hat as I said before is you model prediction okay and yi-y that is this y is the average of all the y's okay. This is called total variance and this is sum of squares of the residual.

This is residual because this is the experimental, this is what the model predicts okay so this difference is called the residual. This is sum of squares of the residual, yi-y this y as I said is the average of all the y's is called total variance is called R square. R square generally should lie between 0 to 1, you cannot go to negative because this is square okay and 0 means very, very absolutely no relationship or fit.

Now 1 is best so when you are doing experimental in the labs with drug discovery and all that, you may have R square may be 0.7, 0.8, we can say it is very good, reasonably good but when you are doing some simulations okay modeling then you may get very high R square but experimental you will get only this R square. So this is 1-sum of squares of the residuals that means this is the experimental, this is the model prediction.

This could be say mx+c like okay or m1x1+m2x2+c predicts so this difference square, this is yi experimental-y this is the average of all the y's okay, this is called total variance okay. This is called R square.

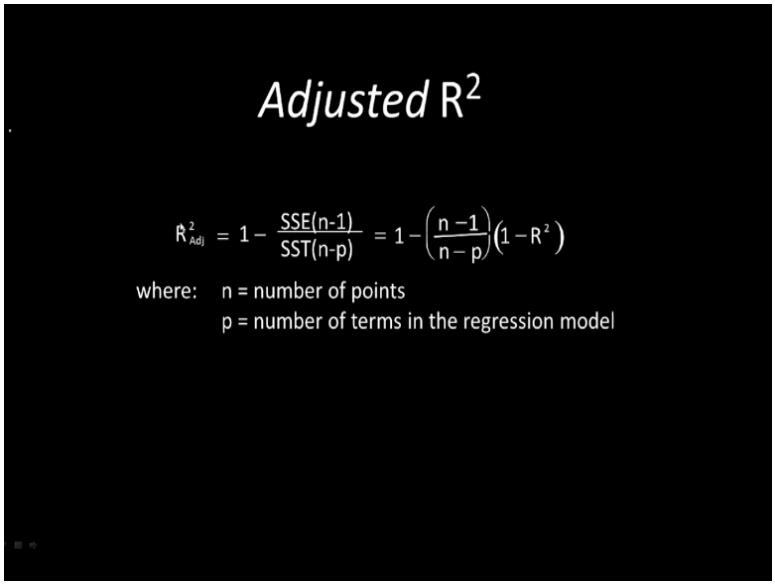
(Refer Slide Time: 23:38)



Degrees of freedom, the number of model parameters-1. The rule of thumb, the number of points should be at least 25% more than the number of degrees of freedom in the regression equation okay. So if you have y=mx+c that is one parameter model m and c are 2 so 25% more so you should have at least 3 data points but we always safe and we should have 5 data points, we are trying to be more careful.

So 3 data point yes you can fit a model but you will not be able to have a test value if you are right so if you have a 5, 6, 6 is good because then one of the points we can use it for test and the remaining 5 we can use it for model fitting okay, this is the rule thumb.

(Refer Slide Time: 24:24)



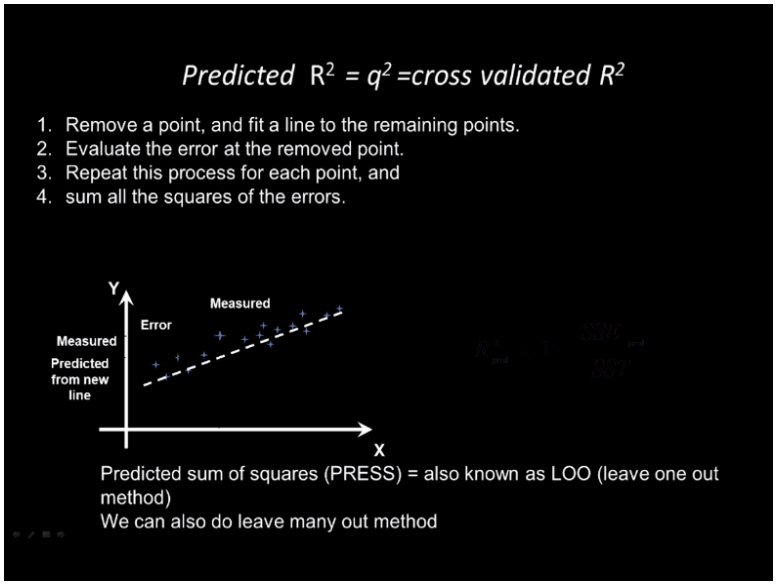
There is something called R square adjusted. So R square adjusted is nothing but 1-sum of squares of the error/sum of squares of the total okay that is this one this is sum of squares of

the total variance n-1/n-p, n is the number of data points, p is the number of terms in the regression model that means y=mx+c means p will become 2 so we can write it as 1-n-1/n-p*1-R square.

So R square adjusted will always be<R square and if you have more number of parameters, terms in the regression model, R square adjusted will come down because you have negative here and here okay. So with more parameters we will be able to fit the model R square will become good but R square adjusted will become bad okay so more number of parameters, R square adjusted will come down.

So the challenge is to keep less number of parameters but still fit your model that is the greatest challenge for you. So R square adjusted will always be<R square.

(Refer Slide Time: 25:38)



Then, there is something called predicted R square. Sometimes, we use R square predicted, some people use capital Q square, this is called cross validated R square okay. This is slightly difficult to understand but let me see. Suppose we have so many data points when we fit the equation we get R square but what we do is for predicting this R square, predicted of Q square, we remove one data point, fit with the -1 set of data that means if you have n data points we use n-1 data point to fit the equation, fit the model.

And then try to predict the value for that point, which was missed out that is evaluate the error at the removed point okay. So then square it, now you keep this point but remove another point again develop a regression model, then try to predict the y value for that point

and then get a residual, square it like that we keep on adding all the sum of squares of the errors or residuals okay.

That is called predicted sum of squares or PRESS, it is also known as LOO leave one out method because every time we remove one data point and then try to develop a model, try to predict the value for the number which we have left out. Once we get the PRESS what we do? We do 1-PRESS/sum of squares of the total that will give you R square predicted or Q square okay.

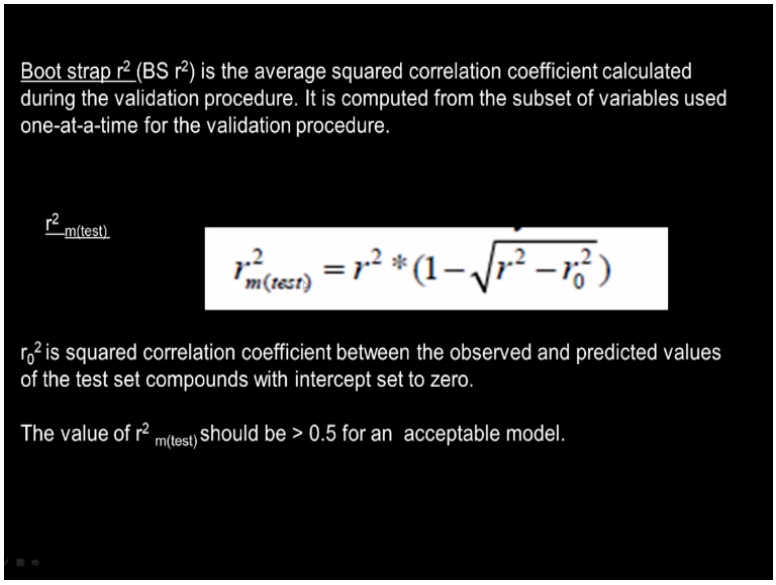
So if you leave out one point at a time that is called leave one out method one point at a time. Do you understand this? So suppose there are n points, we leave one point out so we have n-1 point, we develop a regression model with that n-1 point and try to predict for the nth point, which we have missed out and find out the residual or error, square it, then you leave another point, keep the previously left out point.

So again will have another n-1 points, develop a regression model, predict the value for that point which was left out and get the error, square it, like that you repeat for all the n points, add up all the square of the errors that is called PRESS predicted sum of squares. Then, 1-PRESS/sum of squares of the total will give you R square predicted or Q square okay, that is cross validated R square, people use different terminologies here.

Instead of one point we can also leave many points. We can also to leave many out method, just like leave one out method we can also have leave many out method that means instead of one point we can miss out 2 points and develop a regression relationship and then try to predict the data for those 2 points and repeat the process and get errors. So this is a very stringent R square.

So that is why Q square will always be lesser than R square adjusted, which should be lesser than R square okay. So these are different R squares.

(Refer Slide Time: 29:02)



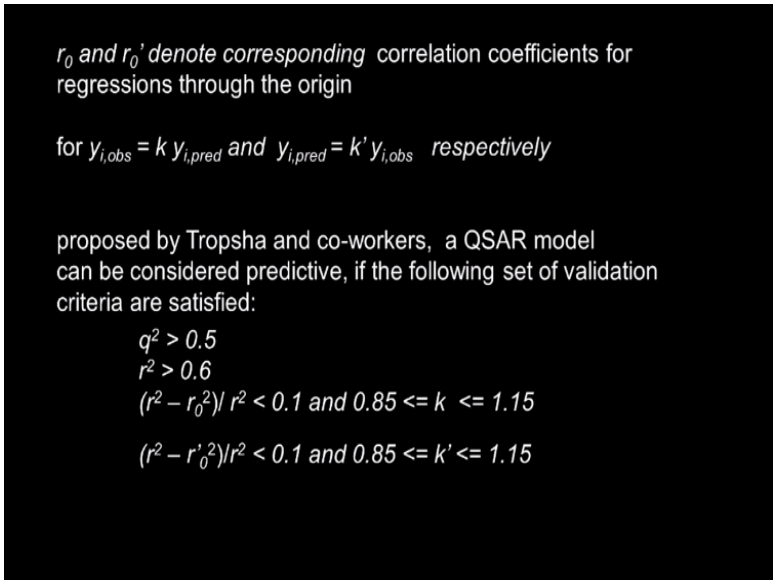
Now more, there is something called boot strap r square is the average of squared correlation coefficient calculated during the validation process. It is computed from the subset of the variables used one at a time for the validation process. So I can leave one point at a time okay and then do for the validation I can take suppose I have 7 points I may take 2 points as a test set, remaining 5 points and then develop R square model.

Then, I can remove some other 2 points and then again develop R square model like that many R squares I will get right. So I will take the average then that is called boot strap r square okay. Then there is also called r square m test okay. This r0 square is the squared correlation coefficient between the observed and predicted values of the test set compounds with the intercept set to zero.

So I also mention that we should have some test set right apart from training set. So with the test set what we do is we do a r square m test using this particular formula where r square 0 is squared correlation coefficient between the observed and predicted values of the test set compound with the intercept set to zero that means you force the intercept to pass through the zero and get an r0 square without allowing it to pass through zero we get an r square.

And then use this formula that is called r square m test. This is more stringent okay so you are forcing the line to pass through the 0 to get this r square, here you are not forcing it so you get this and then see whether the value is>0.5 so that the model can be accepted okay.

(Refer Slide Time: 30:49)



Then you also have other approaches r0 and r0 dash. These are corresponding correlation coefficients for the regression through the origin where you have i observed=k yi predicted or yi predicted=k dash yi observed okay. So we see we fit the model this way, we fit a model this way okay. So this becomes here, this becomes here and then we allow it to pass through the origin okay, you get r0 and r0 dash okay.

Then, you see whether those are quite good above 0.6 okay. So these are different ways of testing it out whether your model has good predicted capability and whether your model has good fitting capability. Tropsha and co-workers also suggested some more approaches for predicting and validation criteria. They said q square should be>0.5, r square should be>0.6, r square-r0 square/r square should be<0.1.

And k here should be between 0.85 and 0.15 that means it should be close to 1. Similarly, the k dash also should be quite high okay. So, so many different things they mentioned okay here, so many things they said r0 as I said r0 square is when you calculate r square when you force it to pass through the origin, here your r dash square when you force it to pass through origin, r dash involves predict, you develop a model between predicted and observed.

Whereas here you observed and predicted developing a model okay so in both the times this k and k dash should be very close to this point 0.85 to 1.15. So there are so many different criteria we need to follow to ensure to ascertain that we have good fit as well as good predicted capability and we are not trying to over fit our data that is very, very important not over fitting our data that is very important.

So these are the various parameters, which we are talking about okay. We have R square, we have adjusted R square, we have Q square or cross validated R square, boot strap r square, r square m, r0 square, then we have k, k dash, r dash 0 square and so on actually okay. Another important point is to see how good the fit is based on the test set data also. We will continue more in the next class okay. Thank you very much for your time.