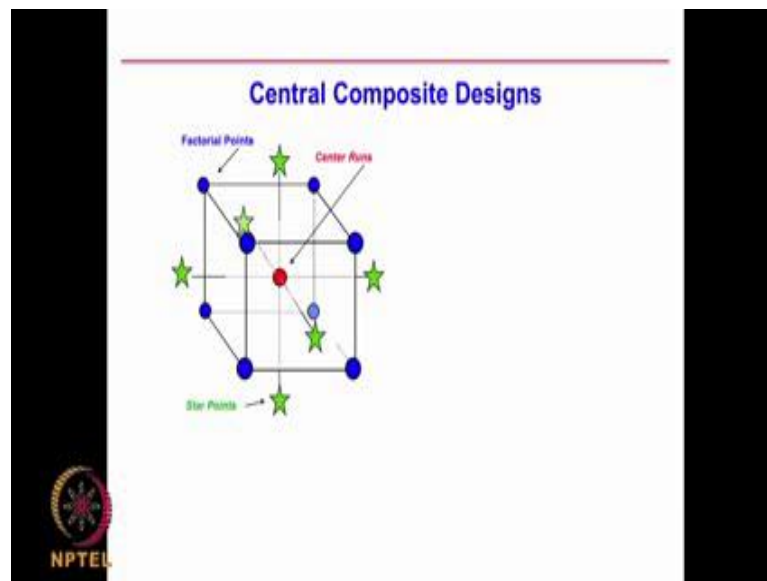


Biostatistics and Design of Experiments
Prof. Mukesh Doble
Department of Biotechnology
Indian Institute of Technology, Madras

Lecture - 38
Second Order Designs

Welcome to the course on Biostatistics and Design of Experiments. We completed what do we want to talk about in screening designs and we started on the second order designs. Second order designs are generally done after you screen all the unwanted factors and come down to a limited number of factors. Second order designs are quadratic in nature. So, we can develop regression models which has second order terms, which has a square terms, and so on, actually. The first one we looked at is called the Central Composite Design.

(Refer Slide Time: 00:46)

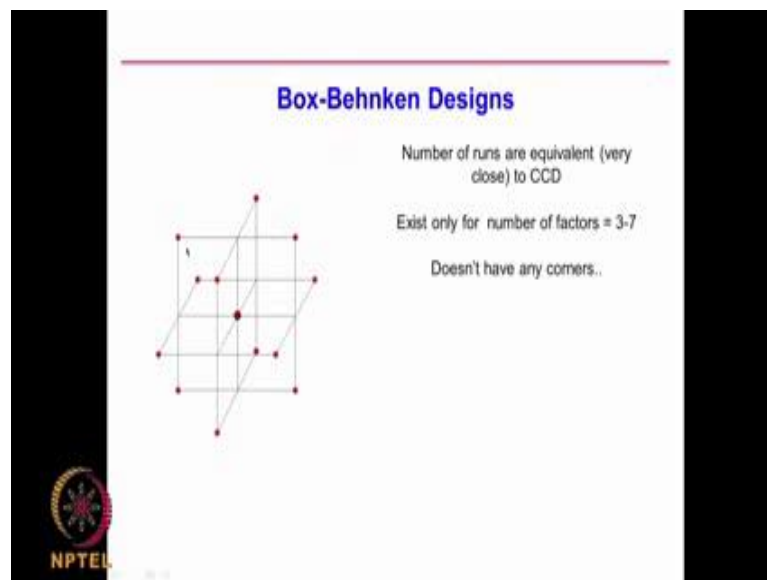


So, in the Central Composite Design, I mentioned, suppose you take a 2 power 3 type of a design, that means, 3 factors, 2, 2 levels, 2^3 design. So, what you have, you have a cube, 2^3 is 8 experiments, $2 * 2 * 2$. So, this could be factor 1; this could be factor 2; this could be factor 3. So, we have 8 experiments. Then, we add one extra experiment in the center. Then, we also add 6 experiments away from the face of the cube. This is the face of the cube; so, away from it. So, you have 6 faces for a cube; so, you have 6

experiments. So, this adds up to $8 + 1, 9, + 6$, that gives you 15 experiments. And, by doing 15 experiments, you are changing each of the factor 5 levels. This is better than a 3^3 type of a design, a full factorial design, 3 levels and 3 factors, that will be $3 * 3 * 3$ is 27, whereas with 15 experiments, I am able to change each of the factor 5 levels; whereas a 3^3 design will change each of the factor only 3 levels. So, that way this is extremely powerful.

Then, we have these Box-Behnken Design. This is also similar to that; only thing is, instead of the corner points it takes the points at the edges of the cube, at the edges of the cube, ok; that is the only difference.

(Refer Slide Time: 02:25)



Box-Behnken Designs

Number of runs are equivalent (very close) to CCD

Exist only for number of factors = 3-7

Doesn't have any corners..

NPTEL

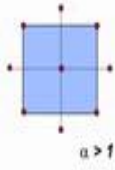
The slide features a 3D wireframe cube with red dots at the midpoints of each of its 12 edges. The text on the right side of the slide provides key characteristics of Box-Behnken designs, including their equivalence to CCD in terms of the number of runs, their applicability to 3-7 factors, and their lack of corner points. The NPTEL logo is visible in the bottom left corner.

So, it does not have the corners; it takes the edges of the cube.


(Refer Slide Time: 02:33)

Star Points Outside the Region of Interest

to explore some area outside it.



$\alpha > 1$

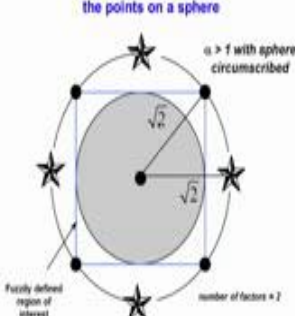


So, let us go back. So, suppose if you have a 2 by 2 system, that is 2^2 , 2 factors only, you have 4 experiments at the corners as the factorial points; then, one experiment in the center, and then, you pickup 4 experiments outside this square. This, we call it α , $\alpha > 1$. If $\alpha = 1$, it will be lying exactly over the edges. Then, that is like your $+1, 0, -1$ type of experiment. Whereas, in the α case, what we are doing is, we are having $+1, 0, -1$; in addition, we have $+\alpha$ and $-\alpha$, for each of the parameters. Do you understand? Now the question is, how do I decide on the alpha?

(Refer Slide Time: 03:33)

The Optimal Value of α

the points on a sphere




For three-dimensional case, it is $\sqrt{3}$.

For six dimensional case, it is $\sqrt{6} = 2.44$.

fully defined region of interest

number of factors = 2

$\alpha > 1$ with sphere circumscribed



So, it has been found that, alpha, square root of the number of parameters is way, is most optimum; that means, if it is a 2, 2^2 , then, α should be 1.414. So, if your square has side of 1, then, alpha should be 1.414; that will be a point. So, the experiments here is +1, 0, -1, +1.414, -1.414, for each of the parameters; that is how it look, or α you call it. So, as I say, as I said, generally, it has been found, square root of number of parameters gives you the best optimum results. So, if it is a 3 parameter problem, you will do square root of 3; that is 1.732; if it is a 6 dimension, square root of 6, that is 2.44. So, this is a good number to have for α .


(Refer Slide Time: 04:25)

CCD Vs 3k Factorial Designs

- CCD has a lower number of runs than the 3k factorial design

# of factors	3k Factorial	CCD
2	9	9
3	27	15
4	81	25
5	243	43

- CCD has lower variance for the regression coefficients than the 3k factorial designs

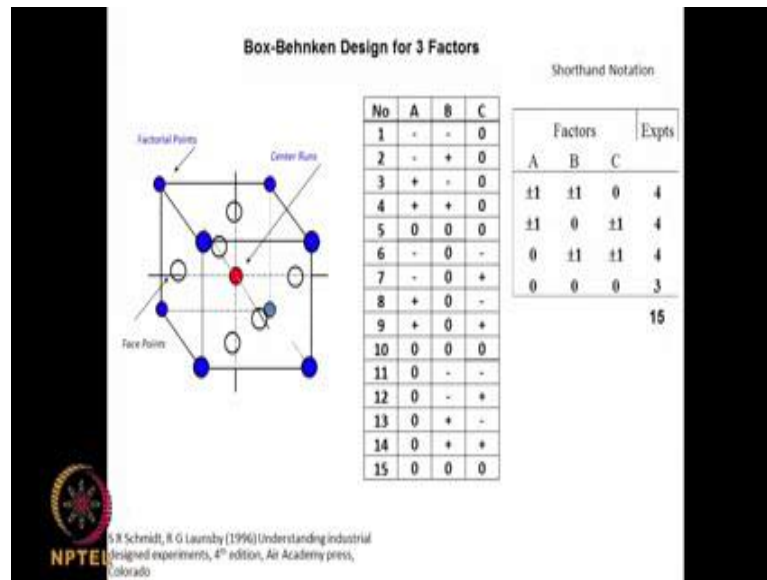


So, if you compare it with your 3k factorial design... So, for a 2 factor, 3 level, $3^2, 3 * 3$ is 9, CCD also gives, will give you 9, that is for 2 factorial; 1, 2, 3, 4, 5, 6, 7, 8, 9, right. So, the number of experiments is the same, but like last time, I told you, if I am doing a factorial, 2 raised to the power 3, I am looking at each factor only at 3 levels, right. Whereas, when I use a CCD, I am looking at each factor at 5 levels. So, I will get better non linear relationship.

Now, let us go to 3 factor problem; 3 levels, $3^3, 3^3$ is $3 * 3 * 3$; that will come out to be 27 experiments; whereas, with CCD, with CCD, I can do 15 experiments; one, this is the $2^3, 8$ experiment, then center, and then, 6 star points, 15 experiments. You see, I have reduced the experiments dramatically. Now, go to 4; $3^4, 81$ experiments; CCD, I just do 25 experiments. 5 factors, $3^5, 243$ experiments, whereas, I get 43 experiments in CCD.

So, number of experiments goes down dramatically. It has got lower variance, because I am looking at different levels of each of these factors. Of course, I can go to fractional factorial of this, we will look at a number of experiments also as we go along, when we do a fraction factorial of $3k$.

(Refer Slide Time: 06:06)



The Box-Behnken Design for 3 factors is given like this. So, it is at 3 levels. So, we have a -, 0 and +. So, as you can see here, we have a minus, minus 0, minus plus 0 and so on. So, there are 15 experiments in this particular set. Basically, what you are doing is, you are doing experiments at these places, these 4 corners, plus, you have a central run. We are doing 3 central runs, and then, we are also doing experiments at these faces. So, there will be 6 faces, actually. So, 8 corners, 6 faces, and then, we are doing at the center.

It can be represented in this short form, A as plus or minus, B as plus or minus and C in 0. That will make up 4 experiments, because A at +, and then A at -, B at +, and B at -. Then, we do A + or -, C + or -, and B will be in 0; that means, A + or -, C + or -; so, that is again, 4 experiments. Then, A will be in 0, and B and C plus or -, + or -; that will be 4 experiments, and 3 experiments at the central point. So, they all add up to 15. This is called a Shorthand Notation. This is the normal for which we are used to. So, when we talk about, say + or -, + or -, + or -, for A and B, and C as 0, this is exactly this. The first 4 experiments are A going both + or -, B going both + or -, and C going 0, and the 5th

experiment is 0, 0, 0; that is, the central point. So, we have 3 places where we have the experiments done at the central point. So, this adds up to 15 experiments.

(Refer Slide Time: 08:09)

Box-Behnken Design for 4 Factors

Shorthand Notation				
Factors				Expts
A	B	C	D	
±1	±1	0	0	4
0	0	±1	±1	4
0	0	0	0	1
±1	0	0	±1	4
0	±1	±1	0	4
0	0	0	0	1
±1	0	±1	0	4
0	±1	0	±1	4
0	0	0	0	1

No	A	B	C	D
1	-	-	0	0
2	-	+	0	0
3	+	-	0	0
4	+	+	0	0
5	0	0	-	-
6	0	0	-	+
7	0	0	+	-
8	0	0	+	+
9	0	0	0	0
10	-	0	0	-
11	-	0	0	+
12	+	0	0	-
13	+	0	0	+
14	0	-	-	0
15	0	-	+	0
16	0	+	-	0
17	0	+	+	0
18	0	0	0	0
19	-	0	-	0
20	-	0	+	0
21	+	0	-	0
22	+	0	+	0
23	0	-	0	-
24	0	-	0	+
25	0	+	0	-
26	0	+	0	+
27	0	0	0	0

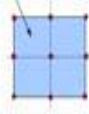
27

Suppose, we have 4 factors, the Box-Behnken Design, 4 factors namely A, B, C, D, again, the same approach. We have A + or -, B + or -, C and D are maintained at 0; so, that adds up to 4 experiments. Then, we have A and B at 0, C and D on + or -, + or -, that adds up to 4 experiments. And then, this is at the center point. So, that comes to one experiment, and then, we do A and D at + or -, B and C at 0; that is 4 experiments. And then, B and C at + or -, and A and D are maintained at 0, that is 4 experiments. And then, one center point, and then, we do A at + or -, C at + or -, that comes to 4 experiments. Then, B at + or -, D at + or -, that comes to 4 experiments. And again, another center point, that is one experiment. So, they all add up to 27. So, this is called the shorthand notation and this is the longhand notation. This is the normal, we are used to it. And so, when you are doing a Box-Behnken Design for 4 factors, we have to do 27 experiments. So, when we are doing a Box-Behnken Design for 3 factors, we do 15 experiments.

(Refer Slide Time: 09:17)

Box-Behnken Design

For 5 parameters




A	B	C	D	E	No.
±1	±1	0	0	0	4
0	0	±1	±1	0	4
0	±1	0	0	±1	4
±1	0	±1	0	0	4
0	0	0	±1	±1	4
0	0	0	0	0	3
0	±1	±1	0	0	4
±1	0	0	±1	0	4
0	0	±1	0	±1	4
±1	0	0	0	±1	4
0	±1	0	±1	0	4
0	0	0	0	0	3
					46

NPTEL
S. R. Schmidt, R. G. Laursby (1996) Understanding industrial designed experiments, 4th edition, Air Academy press, Colorado

So, if you are doing a 5 parameters, that means, A or 5 factors A, B, C, D, E. Then, we can develop in the same way, the shorthand notation is given and we end up with 46 experiments. So, we have A, B **+ or -**, C, D, E are maintained at 0; like that, you keep on varying, and then, you also have these center points. So, you do totally 6 experiments for center points; so, this comes to 46 experiments. So, you need to do 46 experiments, if you want to do the Box-Behnken type of design.

(Refer Slide Time: 09:54)

BOX WILSON - CCD



CCD, n=3 using 2³ factorial

Run	A	B	C
1	-	-	-
2	-	-	+
3	-	+	-
4	-	+	+
5	+	-	-
6	+	-	+
7	+	+	-
8	+	+	+
9	0	0	0
10	0	0	0
11	0	0	0
12	0	0	0
13	0	0	0
14	0	0	0
15	α	0	0
16	-α	0	0
17	0	α	0
18	0	-α	0
19	0	0	α
20	0	0	-α

Factorial
Centre
Axial points

CCD, n=3 using 2³⁻¹ factorial

Run	A	B	C
1	-	-	-
2	-	+	+
3	+	-	+
4	+	+	-
5	0	0	0
6	0	0	0
7	0	0	0
8	α	0	0
9	-α	0	0
10	0	α	0
11	0	-α	0
12	0	0	α
13	0	0	-α

Factorial
Centre
Axial points

$n_c = 4 \sqrt{[n_c+1]} - 2n$

NPTEL

Now, let us do the CCD, which is called the Box Wilson design. The CCD is called the Box Wilson design. In CCD, what do you do? You do the factorial, and then, we do the center points, and then, we do the star points, or we call it α points, they call it α point. So, if you are doing the factorial could be at 2^{-3} , or it could be at 2^{-3-1} .

Let us look at the full factorial. So, we have the 8 experiments for the factorial. This is the first 5, 8 experiments of the factorial. So, we have all the combinations for A, B, C, that is 8, and then, we do; these are the center points. So, we will be doing 6 experiments for the center points. We will... I will tell you why you need to do 6 experiments. And then, we do the star points, or the α points; that means, A at $+\alpha$ and $-\alpha$, B at $+\alpha$ and $-\alpha$, C at $+\alpha$ and $-\alpha$; that comes to 20 experiments. Why do you need to do 6 at the center point? There is a formula which is like this, n_c , that is number of center points


$$n_c = 4 \sqrt{[n_F+1]} - 2n$$

where n_F is the number of factorial experiments.

So, in this particular case, we are doing 8 experiments; that is, 2^{-3} full factorial. So, that comes to $8 + 1$, 9; square root of 9 is 3; so, $12 - 2$, n is the number of factors, number of variables or factors; the... So, here we have a 12, and here we have $2 * 3$, 6; so, $12 - 6$ is 6; so, we are doing 6 experiments. We can do the same Box Wilson, but instead of factorial, we can do a fractional factorial, that means, $2, 3 - 1$, that means, we are doing only 4 experiments for the factorial region, and then, the remaining is the center points, and then, $+\alpha, -\alpha, +\alpha, -\alpha, +\alpha, -\alpha$. Here, we are doing only 3 center points, because according to this formula, 4 square root of, n_F here is 4; so, $4 + 1$ is 5; square root of 5 is 2.236; multiply by 4; so, that is about 10; and then, -, the number of n is 3 here; so, you will end up with 3 center point.

So, depending upon the number of runs you are doing in the factorial, the formula is given, where you substitute that. So, if it is a 2^{-3} full factorial, we will put n_F as 8; if it is a 2^{-3-1} , fractional factorial, we will put n_F as 4; that will give you an idea about the center point. This is how you build up the Box Wilson-CCD, ok.

(Refer Slide Time: 12:48)



Summary of Objectives for 3-Level Designs

Design	Objective
3^k	-Used for qualitative or quantitative factors -Estimate all linear and quadratic effects and, when higher order resolutions are desired, you can also estimate linear and quadratic 2-way interactions -Typically the 3^k is a good choice if you have qualitative factors with few or no interactions. This is not a good choice for factors that are quantitative.
Box-Behnken	-Used only for quantitative factors; however, for some k you can have one qualitative blocking factor -Estimate all linear, quadratic and 2-way linear interactions plus experimental error.
CCD	-Used primarily for quantitative factors; however, if you only have 1 quantitative factor, the CCD can still be useful -Estimate all linear effects, selected quadratics, and selected 2-way linear interactions plus experimental error -This is typically your best choice for quantitative factors, each as 3 levels
Full factorial	-Used for qualitative or quantitative factors -Estimate all linear and quadratic effects plus all possible simple and higher order interactions.

So, we have looked at different types of 3 level designs, which are also called non linear designs. So, we have the $3^k - q$. This is exactly like your $2^n - q$ type of thing. It can be used for qualitative or quantitative factors. We can use it for linear and quadratic effects. You can also estimate linear and quadratic interaction; that means, I can put a A^2 or $A * B$ type of terms in my regression. So, this can be done, actually.


Now, we looked at Box-Behnken Design. So, we can use it for quantitative factors. Of course, we can also use it for linear, quadratic, two-way linear interactions and so on. Then, we came to CCD. The advantage of CCD is, it can look at almost 5 different levels, because we are having star, (Refer Time: 14:47) $\alpha +$ and $\alpha -$, unlike any other, other designs we are talking about. We can estimate linear effects, quadratic effects, two-way linear interactions and so on. Then, the full factorial, 3^k . So, we are going to have a large number experiments here. We can use it for qualitative or quantitative; we can use it for linear, quadratic effects; also, all types of interactions, two-way, 2 level interactions, 3 level interactions and. so on, actually. This is a full factorial design, ok.

(Refer Slide Time: 14:17)

Comparison of Design Types for Quantitative Factor Scenarios

Scenario	3^k	Box-Behnken	CCD**		Full factorial
k=3 with no interactions	9 (0)	15 (3)	13 (3)*		27 (1)
k=4 with 3 2-way linear interactions	27 (0)	27 (3)	20 (4)*		81 (1)
k=5 with 1 2-way linear interactions	27 (0)	46 (6)	20 (2)*		243 (1)
k=5 with all 2-way linear interactions	81 (0)	46 (6)	33 (7)*		243 (1)
k=7 with 7 2-way linear interactions	81 (0)	62 (6)	33 (3)*		2187 (1)

* Number of centre points
 ** The total number of runs for the CCD includes 2k axial points. In many cases all the factors will not necessarily have a significant second-order term.



So, this is a comparative table; full factorial, if am having k is equal to 3, that means 3 factors, or 3 variables, 3^3 , $3 * 3 * 3$ is 27; and then, if it is 4, sorry, $3 * 3 * 3 * 3$, 81; so, it grows up exponentially. If you come to CCD, if you assume a fractional factorial, for the central portion, and then, we consider the center points as well as the star points, you will have 13 experiments. If you remember this, if you remember this, we have 13 experiments, right. So, we can have 13, whereas Box-Behnken will give you 15 experiments. If you have 4 variables, then we can have 27 experiment; that is, we can build up one-third factorial type of design; Box-Behnken gives you 27. This tells you how many center points, and CCD, fractional factorial for the central portion will give you 20 experiments, and there will be 4 center points. Like that, you know. So, as you can see, CCD appears to be very efficient when compared to any of these design - Box-Behnken, or 3^{k-q} , or full factorial design, actually. And, also you can see, we can get center points; that is very very important. Both these designs give enough number of center points which will help you to estimate the errors.

So, CCD appears to be the best, if you are looking at a quadratic type of model or higher order models. So, we looked at a large number of design strategies. We looked at the, the screening design strategy; then, we looked at the second order design strategy. Here, we have the Box-Behnken; we have the central composite designs; we have the, of course, the full factorial 3 level design; then, we have the fractional factorial 3^{k-q} type of design and so on. So, once we have these designs, what do we do? We can perform an ANOVA.


Once we perform an ANOVA, we will know which factors are important. We can even look at whether any interactions are important. Then, we can go into regression; we can develop regression relationships, and then, get a model. If it is a second order model, that means, if it is a quadratic model, we can optimize that model. By doing an optimization, we can find out the best set of parameters at which, say my yield is maximum. Then, I can go and test it out in my actual reactor, bio reactor, or whatever system I am looking at to see whether I got the same answer as predicted.

So, ultimately, you want to optimize, or maximize your productivity, maximize the yield of the metabolite, maximize the yield of biomass. So, obviously, once you perform your design calculation, you shortlist the best set of parameters, you identify which parameters are going to play a important role, are significant, eliminate the other parameters; then, you develop a regression relationship; and with the regression relationship, you optimize your product yield; that means, you change the parameter values theoretically, using the regression relationship, you change parameter value and find out the best set of parameter values, which give you the maximum yield. And then, you can go and test it out in your lab, and you can say, this is the best type of conditions at which I should operate my bio reactor. So, this is how the sequence of events take place, when you start looking at the experimental design strategy. So, that means, I need to have a good knowledge about, about my regression.

(Refer Slide Time: 18:34)

Summary of Objectives for 3-Level Designs

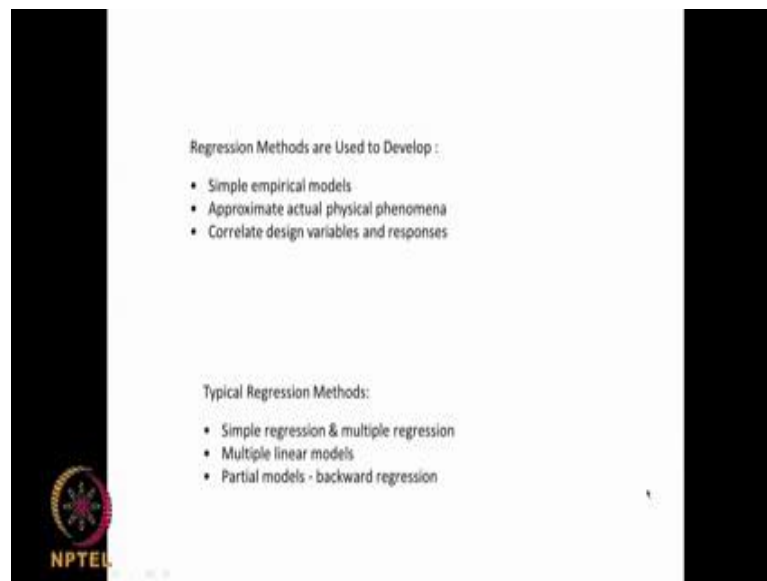
Design	Objective
3^{k-1}	<ul style="list-style-type: none"> -Used for qualitative or quantitative factors -Estimate all linear and quadratic effects and, when higher order resolutions are desired, you can also estimate linear and quadratic 2-way interactions -Typically the 3^{k-1} is a good choice if you have qualitative factors with few or no interactions. This is not a good choice for factors that are quantitative.
Box-Behnken	<ul style="list-style-type: none"> -Used only for quantitative factors; however, for some k you can have one qualitative blocking factor -Estimate all linear, quadratic and 2-way linear interactions plus experimental error.
CCD	<ul style="list-style-type: none"> -Used primarily for quantitative factors; however, if you only have 1 quantitative factor, the CCD can still be useful -Estimate all linear effects, selected quadratics, and selected 2-way linear interactions plus experimental error. -This is typically your best choice for quantitative factors, each as 3 levels
Full factorial	<ul style="list-style-type: none"> -Used for qualitative or quantitative factors -Estimate all linear and quadratic effects plus all possible simple and higher order interactions



What type of regressions to use, and what type of models I need to use, and so on, actually.

So, let us look at some of those regression approaches by which one goes about doing the modeling. So, what is a regression analysis? So, regression methods are used to develop simple empirical models.

(Refer Slide Time: 19:12)



So, we can get $y = a + b x_1 + c x_2$, where x_1, x_2 are my parameters. If I have a second order model, I can use it for $a = b x_1 + c x_1^2 + d x_1 x_2$ and so on. We can get some actual understanding of the physical phenomena. We can correlate design variables. So, we can do so many things with regression models. And, typical regression methods use a simple regression method, multiple regression method, multiple linear models, partial models, backward regression and non linear models, second order models, and so on, actually. So, regression, we can do so many things with the regression models.

(Refer Slide Time: 20:08)



So, regression can help us to understand which factors are important. Of course, we can do it through it ANOVA also, as I have been talking about it in the past so many classes. So, it can... of course, the regression also tells you, depending upon the parameter which you get, which ones are good. We can use it for predictive. Like I said, I can optimize the conditions, and then, I can say, if I operate at this temperature, this pH, this carbon amount, I will get maximum yield. How do I do that? I can take the regression model, I can modify that, the conditions and see whether I am able to increase the productivity; that is called simulation. And, it is, that is called prediction. So, you come up with the new operating points. We can use it for estimating the coefficients, what will be the constant that gets multiplied with temperature; what is the constant that gets multiplied with pH. The constants also tell you something very very important. It tells you, what is the effect of each of these, what is the quantum of effect, or the weightage of each of these parameters on my total yield, for example; if that is what is my desire.

So, regression has lot of uses. It not only developing an equation, it can be used for later optimization, simulation, model predictions, new scenario generation, looking at which factors are important, which factors are not important, whether some factors are positively correlated, whether some factors are negatively correlated, and so on, actually. So, regression analysis is extremely important, if one wants to do a proper analysis.

So, what are the approaches, we pick a regression model. So, do I go for a linear regression model, do I go for a non-linear regression model, do you have $a X_1^2$, do I have $X_1 * X_2$, that is interactions; X_1^2 means it is a quadratic; X_1 and X_2 means it is an interaction.

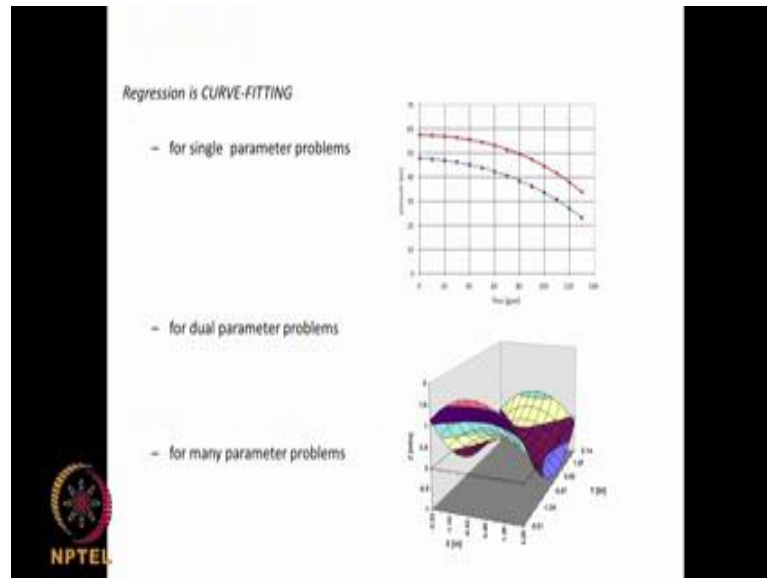
(Refer Slide Time: 22:21)



So, I need to decide, do I go for a linear model, nonlinear model. Then, you gather the data, of course; you collect enough data. You know which parameters are important after performing a screening design. Then, you perform the regression analysis. That means, you actually fit your data to the model you have selected, and then, estimate those constants, A, B, C, whatever it is, the constants of this regression model. So, you need to decide on which model to consider; that is very, very important. So, a priori you say I will take a linear, multi linear model; that means, $a + b X_1 + c X_2 + d X_3$, like that; that is a multi linear model; or, I will say, I will take a single parameter quadratic model, $a + b X_1 + c X_1^2$; or, I will take a 2 parameter model with quadratic terms; that means, $a + b X_1 + c X_2 + d X_1 X_2 + e X_1^2 + f X_2^2$; like that, you know. So, you need to a priori decide. But, later on, when you do the regression analysis, depending upon your A, B, C, D, you may neglect some of the parameters; if the parameter values are very, very small, you may do that. But, you need to anyway consider a priori, what type of model you will do; that is very, very important. Then, you perform the regression analysis, and you select the best model for your... based on the r

square, based on the ANOVA. So, like that, finally you end up with the best regression model for your system.

(Refer Slide Time: 24:10)



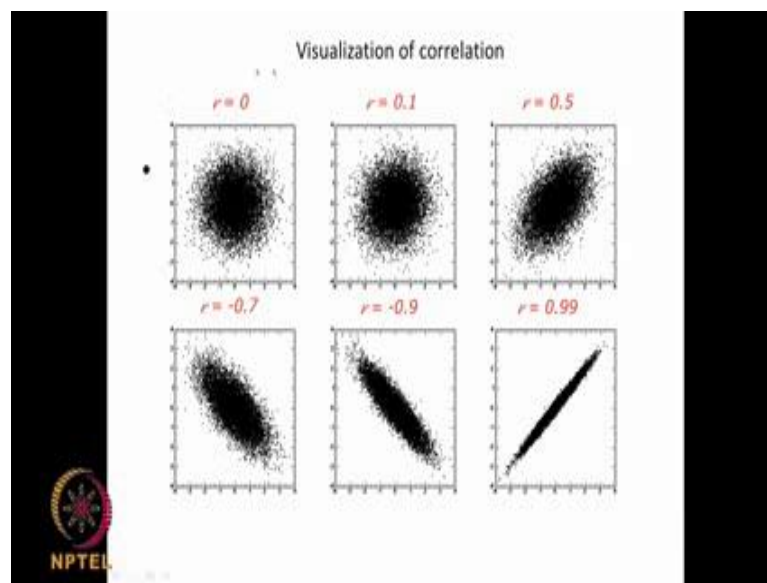
So, regression is nothing but a curve fitting for a single parameter; that means, if I am measuring flow as a, sorry, I am measuring pressure as function of flow, flow is your independent variable; your pressure is a dependent variable. So, if it is a linear model, I will have $y = A + B \cdot \text{flow}$, that is, pressure is equal to $A + B \cdot \text{flow}$. If it is a second order model, I may say $\text{pressure} = A + B \cdot \text{flow} + C \cdot \text{flow squared}$. So, when I fit the data, I can estimate A, B, C.

So, in this particular case, I may think, it is not a linear, it is a non-linear. So, I will take $A + B \cdot \text{flow} + C \cdot \text{flow squared}$. For a 2 parameter, of course, suppose, I have X here, y here, and z is my dependent variable; so, I will have, linear model will be, $y = A + B \cdot x_1 + C \cdot x_2$; and non-linear model will be $y = A + B \cdot x_1 + C \cdot x_1^2 + D \cdot x_2 + E \cdot x_2^2 + F \cdot x_1 \cdot x_2$ also, if you think there is an interaction, and so on, actually. Then, you estimate the constants A, B, C, D, E, F; by fitting the data, you determine the error. What is error? What is predicted by the model, minus what is actual, square it up, sum it up - that is the error sum of squares. So, from there, you perform an ANOVA, and then you see, whether the model is really predicting the data or not. So, that is quite straight forward, and that is what you do in your regression analysis. So, basically, it is a curve fitting for a single parameter, curve fitting; so, it is easy in a one-dimensional system, that means,

single parameter, or even a two-dimension system to visualize; but, if it becomes higher dimension, we cannot really visualize. We have to look at the error sum of squares and then make a conclusion, whether the error sum of squares is very large for one model. I have another model; I get another error sum of squares. I can compare both of them and say, as they are very different, or no, they are not very different.

But, if I am having a very simple single parameter, or a 2 parameter model, I can even visually plot them, and I can say, yes, this model looks good. But, for larger systems, it becomes more difficult. You will not be able to tell, unless you actually calculate the error sum of squares. So, that is very, very important. And, we will go down, as we go along to... and I will tell you how to calculate these various parameters, and how to estimate the constants of regression, the multiplication factor or the slope of regression line, and so on. Excel also has that option, to determine both the slope, as well as the intercept for the regression line.

(Refer Slide Time: 27:26)



So, if we can visualize the data, there are different relationships that are possible. Suppose, I take a parameter X here, that is the independent variable, and this is my dependent variable Y , I can have different types of relationships; that means Y and X are highly correlated. As you can see, as X increases, Y also increases; whereas here, if you see this data, as X increases, Y is all over the place. So, this is a 0 correlation. This is the best correlation. So, in between, you can have different numbers possible; in between,

you can have different numbers possible, different figures; look at these figures; not at all correlated, little correlated, highly correlated. So, you can have correlation coefficient varying between 0 to 1, and (Refer Time: 29:09) if it is 1, as X increases, Y also increases; if it is -1 , as X increases, Y decreases. So, you can have both the situation. (Refer Time: 29:09) And, the direct correlation, or indirectly related, or positively related, negatively related, and this is called the correlation coefficient. And, I will tell you how to estimate the correlation coefficient also. But, it becomes very easy if you have 1 parameter to calculate the correlation coefficient, ok.

(Refer Slide Time: 29:09)

Covariance is a measure of the strength of the correlation between two or more sets of random variables

$$COV(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$

$$\text{Correlation} = \rho = \frac{\text{cov}(X, Y)}{\sigma_x \sigma_y}$$

Correlation is a scaled version of covariance

There is something called a Covariance. Covariance tells you the strength of the correlation between 2 or more sets of random variables, X and Y. So, how do you calculate? It is given by this formula,

$$COV(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$

covariance x y, summation of, i is varying from 1 to n, x i minus x bar, x bar is the average of all the x s, y i minus y bar, y bar is average of all these y s divided by n minus

1, is called the covariance. It tells you the strength of the correlation, and the correlation coefficient is given by

$$\text{Correlation} = \rho = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

covariance of X Y / sigma x sigma y; sigma x is, what is sigma x? $\sum (x_i - \bar{x})^2$, divided by n - 1. What is sigma y square? $\sum (y_i - \bar{y})^2$, ÷ n minus 1 summation.

So, correlation is the scaled version of covariance. You are scaling it down with respect to the sigma; that is standard deviation X, and standard deviation Y. So, as I said, once you collect the data, we need to perform a regression analysis, and before the regression analysis, we actually look at the correlation coefficient between **X** and the **Y**; that means, the independent variable and the dependent variable. So, if I have many **X**s, that means, I have temperature, pH, carbon amount, and for each one of them, I have the yield data; so, I can perform a correlation analysis. I can calculate the correlation coefficient between temperature and the yield, between pH and the yield, between carbon amount and yield, to see whether my correlation coefficient value is very high or very low; whether they are correlated or not correlated. So, you can do that as your first step to analyze your data. So, we will continue on this regression analysis in the next class also.

Thank you very much.

Key words - Second order designs, Central Composite Designs, CCD, Box-Behnken Designs, CCD Vs 3k Factorial Designs, Box-Behnken Design for 4 Factors, BOX WILSON – CCD, Regression Methods, Main Uses of Regression, regression analysis, Visualization of correlation, Covariance

