

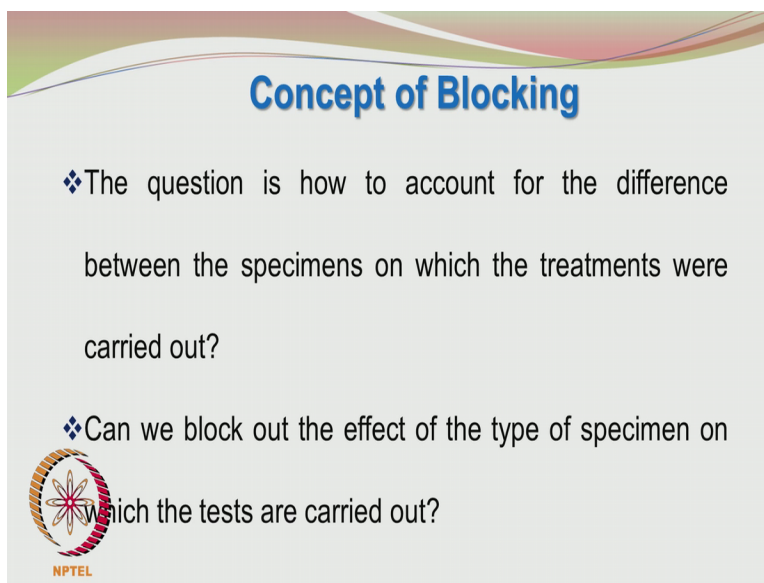
Statistics for Experimentalists
Prof. Kannan. A
Department of Chemical Engineering
Indian Institute of Technology - Madras

Lecture - 54
Statistics for Experimentalists - Summary Part B

Okay, continuing with our discussion on blocking, this is a rarely used concept but has lot of practical value. So it is not possible for you to carry out repeats on a single specimen, but you do it on different ones. For example, if you are trying to look at the effect of different fertilizers in the plot of land, so you put one plot of land and then you put different fertilizers in that plot and monitor the growth of the crops.


Obviously, if you are talking of repeats then you have to wait until the first crop is ready before you put the fertilizers the second time, but if you want to repeats in parallel then you have to put different fertilizers in a second field.

(Refer Slide Time: 01:15)



Concept of Blocking

- ❖ The question is how to account for the difference between the specimens on which the treatments were carried out?
- ❖ Can we block out the effect of the type of specimen on which the tests are carried out?

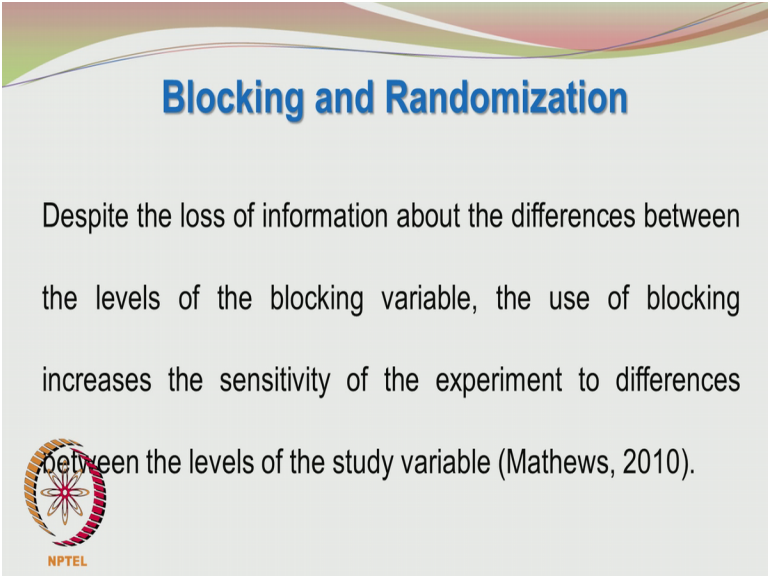

NPTEL

And the second field is different from the first field or it may be different from the first field, so the first field and second field are called as blocks. So the experiments are conducted on different blocks, and we have to account for the blocking influence also, and if you account for the blocking influence, you lose on the degrees of freedom. And what happens is that is a negative point, but on the other hand the sensitivity of your test will increased when you do blocking.

So in blocking we tried to answer the question, how to account for the difference between the specimens on which the treatments were carried out? So what we are doing is we are carrying out different tests on a particular specimen, now that specimen is used up, then we do the same test on the same second specimen. But the first specimen is different from the second specimen or it may be different it may be identical, but usually it will be different from the first specimen.


So the first specimen is a block, the second specimen is also a block, so here we are accounting for the variability due to the 2 different specimens. So we are blocking out the effect of the type of specimen on which the tests were carried out.

(Refer Slide Time: 02:53)



Blocking and Randomization

Despite the loss of information about the differences between the levels of the blocking variable, the use of blocking increases the sensitivity of the experiment to differences between the levels of the study variable (Mathews, 2010).

 NPTEL

So there is some loss of information because of blocking, actually the blocking helps to increase the sensitivity of the experiment to differences between the levels of study variables okay. So please look at the ANOVA table given for blocking, and see the degrees of freedom eaten up by the blocking, and on the other hand how the tests became more sensitive due to the blocking effect.

(Refer Slide Time: 03:25)

Advantages of Factorial Design

- ❖ scientific interpretation of results
- ❖ optimization approaches: Response Surface Methodology
- ❖ quantitative and qualitative factors may be analyzed together



compulsory for industrial competitiveness

Next we move on to factorial design, here we are talking about not a single factor, but we are going to talk about more than one factor. So it can be multiple factors and those factors may be set at 2 levels. In factorial designs of level 2, we can go for any number of factors 3, 4, 5, but each factor will be set at only 2 levels, one lower level, and the other a higher level, so we call it as -1 setting and +1 setting.

What are the advantages of factorial design? It helps to analyze and interpret your results in a scientific manner, you can carry out the response surface methodology, and qualitative and quantitative factors may be analyzed together. For example, you can have temperature, pressure and the type of catalyst carried out in your analysis. Temperature, pressure will be having continuous range of values, whereas the type of catalyst maybe catalyst A, catalyst B and so on.

So it is a discontinuous variable or a qualitative variable, but using design of experiments and factorial design you can account for all the 3 factors simultaneously. And this factorial design is compulsory for industrial competitiveness.

(Refer Slide Time: 04:51)

Advantages of Factorial Design

The design is **orthogonal** as the different effects and their interactions contribute to the sum of squares independently

(The sum of squares is a measure of variability due to experimental error and factors)



And the advantages of factorial design are manifold, the design is orthogonal as the different effects and their interactions contribute to the sum of squares independently. So when you have an orthogonal design, each factor contributes the response in its own way, so the variability in the response is contributed independently by the different factors.

(Refer Slide Time: 05:24)

Information Content of Data

❖ Enables us to extract required information from the experiments even in the face of distractions created from **unpreventable** random variability

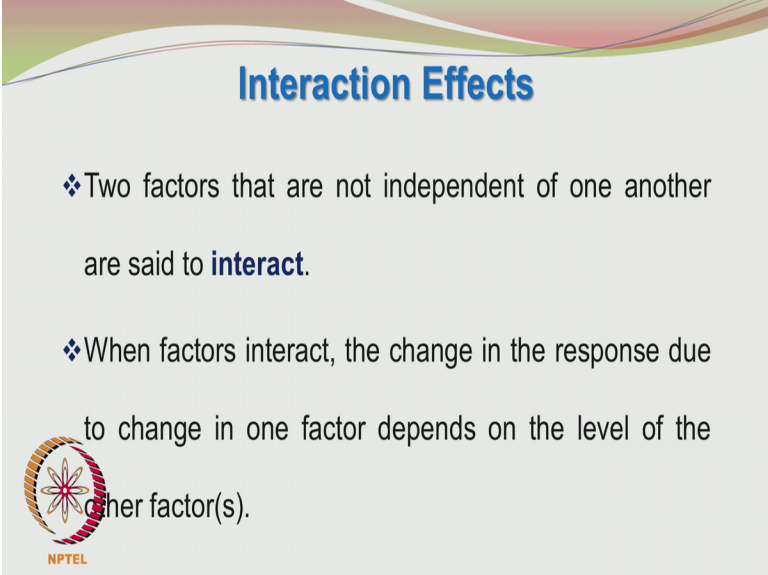
❖ DOE enables us to extract rich informative content from data using limited number of experiments



And factorial design of experiments enables us to extract required information from the experiments, even the face of distractions created from unpreventable random variations. So there are going to be random variations throughout the course of our experiments, despite the distractions from that the design of experiments especially the factorial design will help us to find or identify the main effects and their interactions.


One more important thing especially in industry is design of experiments help us to extract rich informative content from data using limited number of experiments.

(Refer Slide Time: 06:09)



Interaction Effects

- ❖ Two factors that are not independent of one another are said to **interact**.
- ❖ When factors interact, the change in the response due to change in one factor depends on the level of the other factor(s).



Another important concept in factorial design is the interaction between factors, what is really meant by interaction between factors? What will happen is the role of factor A will depend upon the level of factor B, at 1 level of factor B A may behave in 1 manner, and that another level of factor B, A may behave in another manner. In which case the 2 factors are said to interact, the 2 factors that are not independent of one another are said to interact.

When factors interact, the change in response due to change in one factor depends on the level of the other factors.

(Refer Slide Time: 06:55)

Interaction Effects

If the change in level of the first factor causes a certain change in output response at one level of the second factor, an identical change in the first factor level at the second level of the second factor will produce a **markedly different** output response.



So if the change in level of the first factor causes the certain change in output response at 1 level of the second factor, and identical change in the first factor level at the second level of the second factor will produce a markedly different output response. So for this you please look at the example I had given on cricket scores from a batsman depending on whether he had taken tea or beer before coming out to play.

(Refer Slide Time: 07:26)

ANOVA

| Source of Variation | Sum of Squares | Degrees of Freedom | Mean Squares | F _o |
|---------------------|----------------|--------------------|----------------------|----------------|
| A Treatments | SS_A | $a-1$ | $SS_A/(a-1)$ | MS_A/MS_E |
| B Treatments | SS_B | $b-1$ | $SS_B/(b-1)$ | MS_B/MS_E |
| Interaction | SS_{AB} | $(a-1)(b-1)$ | $SS_{AB}/(a-1)(b-1)$ | MS_{AB}/MS_E |
| Error | SS_E | $ab(n-1)$ | $SS_E/ab(n-1)$ | |
| Total | SS_T | $abn-1$ | | |



Now if you look at the typical analysis of variance table in factorial design of experiments, you have a source of variability due to A treatments and B treatments, and then you have the interaction between A and B, then you have the contribution from the error. And so you have sum

of squares of factor A, sum of squares of factor B, sum of squares of factor AB, sum of squares of error.

And again you have the degrees of freedom $a-1$ for A and $b-1$ for B, and interaction has $a-1*b-1$ degrees of freedom, and error has $ab*n-1$. We calculate the mean squares as usual by dividing the sum of squares by the degrees of freedom the respective degrees of freedom, and so this is what you have. So to find F_0 for A, we find the mean square A by mean square error. For F_0 for B, we find mean square B by mean square error.

For finding out the F_0 for interaction between A and B we take means square interaction and divided with the mean square error. So when you have these 3 F values you compare it with the F alpha numerator and denominator degrees of freedom, so the numerator degrees of freedom would be corresponding to the different factors, and denominator degrees of freedom would be corresponding to the error degrees of freedom.

If the computed value of F_0 for the different factors and the interaction or higher than F alpha numerator, denominator degrees of freedom, then those F values are lying in the rejection region. So the F alpha numerator and denominator degrees of freedom would define the critical F value, and if that critically F value is exceeded by one or more of these 3 statistics, then those particular factors are said to lie in the rejection region, and we can reject those appropriate hypotheses.

The hypothesis here would have been the treatment A is having no effect at all, $\mu_A = \mu$ or $\tau_A = 0$, the effect of factor A=0 so $\tau_A = 0$. Similarly, for factor B we say that $\mu_B = \mu$, the overall average or $\mu_B = \mu + \tau_B$, the null hypothesis says that $\tau_B = 0$. Similarly, for the interaction, if your F statistic lies in the rejection region, you reject the null hypothesis and say that factor A is important or factor B is important or factor AB, the interaction between AB is important depending upon whether the F value is lying in the rejection region or not.

(Refer Slide Time: 10:49)

Multiple Regression Model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

β_0 : Intercept

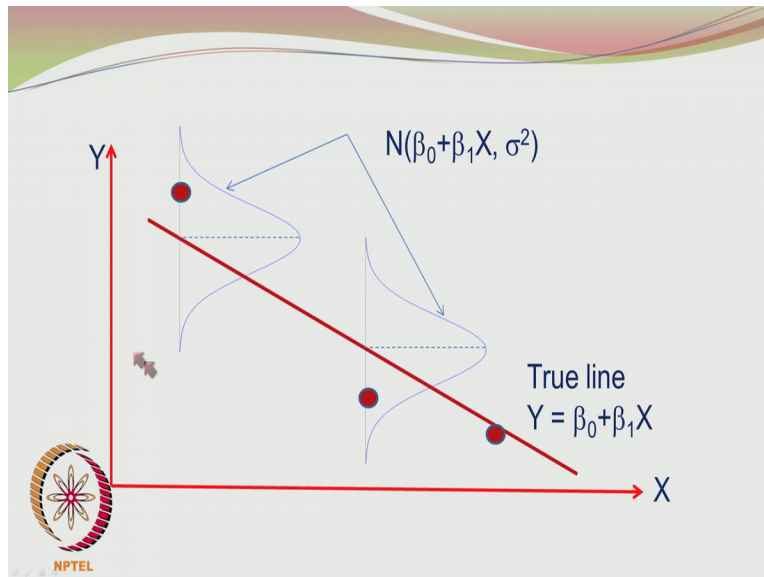
β_1 : partial regression coefficient 1

β_2 : partial regression coefficient 2



Then we move on to multiple regression, where we have the experimental response in terms of $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \text{random experimental error}$, so the response Y is not only strictly determined by factors X_1 and X_2 , but also by a random error component, so β_1 and β_2 are called as partial regression coefficient 1 and partial regression coefficient 2 respectively.

(Refer Slide Time: 11:21)

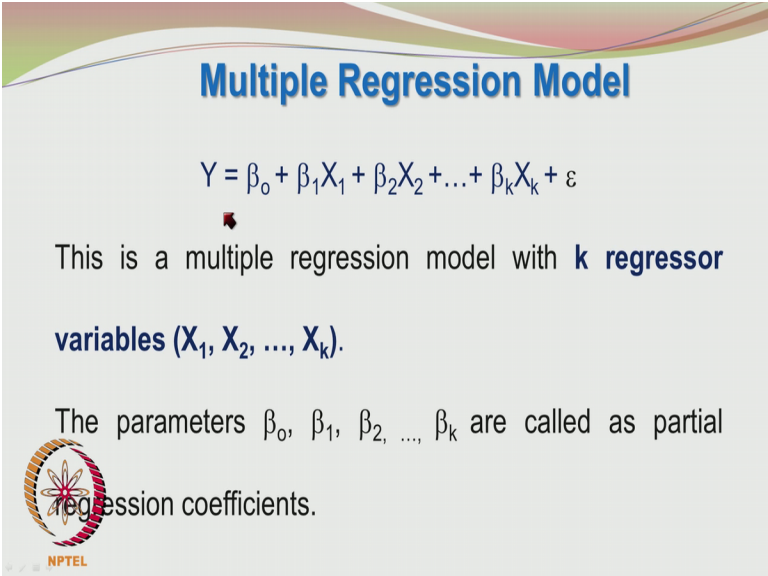


So this is a very interesting diagram, here we plot the response versus X . And we find the response is scattered in the 2 dimensional plane, and we try to fit a line which passes in the best possible manner through the points, we try to balance the line through these points, there can be more than 3 such points. What is of importance is the regression line represents the true value, and the experiments are showing deviations from the true value because of random fluctuations.

So that distribution of the fluctuations from the true regression value is described by a normal distribution centered around the true line value, and the variance of this distribution is sigma squared, the sigma squared is also called as the error variance. Because of the random fluctuation or random errors only the data points are deviating from this straight line. So the mean of these normal distributions correspond to the regression line value $Y = \beta_0 + \beta_1 X$.

But there are deviation from this because of random error contributions, and when we do the experiments next time you may get the data point lying somewhere here, or it may be line somewhere here, because it is a random phenomena.

(Refer Slide Time: 13:08)




Multiple Regression Model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \varepsilon$$

This is a multiple regression model with **k regressor variables** (X_1, X_2, \dots, X_k).

The parameters $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ are called as partial regression coefficients.

 NPTEL

We can also do multiple regression model especially with linear algebra in a very swift manner, so we describe a general multiple regression model as $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \varepsilon$, and this has k regressor variables, this is a multiple regression model with k regressor variables, so they are also called as factors or regressor variables X_1, X_2 so on to X_k . The parameters $\beta_0, \beta_1, \beta_2$ so on to β_k are called as partial regression coefficients.

(Refer Slide Time: 13:48)

Matrix Approach to Multiple Regression

Let there be k regressor variables with n observations

$$(X_{i1}, X_{i2}, X_{i3}, \dots, X_{ik}, Y_i), i=1, 2, \dots, n$$

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik} + \varepsilon_i \quad i=1, 2, \dots, n$$

Note: $n > k$



this may be represented in a matrix notation as $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

Now we can have a matrix approach to multiple linear regression, so if there are k regressor variables X_1, X_2 so on to X_k and with n observations, here the index i represents the run number. You can have n runs performed, so X_{i1}, X_{i2}, X_{i3} and X_{ik} are the X values corresponding to the i th run for factors 1, 2, 3, 4 so on to k . And the model is given by $\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik} + \varepsilon_i$, where ε_i is the random error component.

And usually the number of experimental settings should be greater than the number of regression parameters, so this may be represented in matrix notation as $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.

(Refer Slide Time: 14:42)

Matrix Form of the Regression Equations

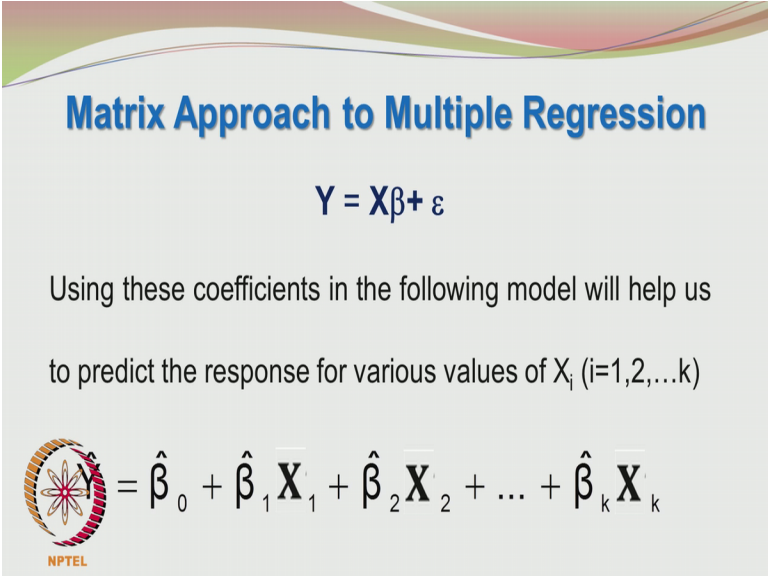
$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & X_{13} & \dots & X_{1k} \\ 1 & X_{21} & X_{22} & X_{23} & \dots & X_{2k} \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & X_{n1} & X_{n2} & X_{n3} & \dots & X_{nk} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$



So you have the column vector here $Y=Y_1, Y_2$ so on to Y_n , $X=1, X_{11}, X_{12}, X_{13}$, so on to X_{1k} . This perhaps maybe main factor A, main factor B, X_{13} may have been interaction between the 2 factors and so on. You can even have quadratic terms like X_{11} squared or X_{22} squared and so on, and so you have for n experimental settings. And then you have the beta column vector comprising of β_0, β_1 so on to β_k .

Epsilon is the column vector corresponding to random error component ϵ_1, ϵ_2 so on to ϵ_n .

(Refer Slide Time: 15:31)



Matrix Approach to Multiple Regression

$$Y = X\beta + \epsilon$$

Using these coefficients in the following model will help us to predict the response for various values of X_i ($i=1,2,\dots,k$)

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \dots + \hat{\beta}_k X_k$$

NPTL

So when we defined the matrix approach to multiple regression $Y=X\beta + \epsilon$, using these coefficients in the following model will help us to predict the response of the various values of X_i . So we were knocking off the error component, because this is the prediction, we say that $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \dots + \hat{\beta}_k X_k$, note that we are having k regression parameters $\hat{\beta}_1, \hat{\beta}_2$, so on to $\hat{\beta}_k$, $\hat{\beta}_0$ is the intercept in the multi-dimensional space.

(Refer Slide Time: 16:13)

Least Squares Estimators of β

This leads to the solving of the following system of equations

$$X'X\hat{\beta} = X'Y$$

These are the **least squares model equations** in matrix form. Their solution is given by

$$\hat{\beta} = (X'X)^{-1}X'Y$$



So when we want to find the least square estimators for beta, we can solve this equation for beta hat $X'X^{-1}X'Y$ gives you beta hat.

(Refer Slide Time: 16:27)

Variance-Covariance Matrix

The variances of the least square estimators are the elements of the $(X'X)^{-1}$ matrix multiplied by the variance σ^2 .

$$C = (X'X)^{-1}\sigma^2 = \begin{bmatrix} C_{00} & C_{01} & C_{02} \\ C_{10} & C_{11} & C_{12} \\ C_{20} & C_{21} & C_{22} \end{bmatrix} \sigma^2$$



So next we go on to the variance-covariance matrix which is a very important one, and we want to look at the variances of the estimated parameters, if the variances are small then those parameters are being estimated quite precisely. So you have the $X'X^{-1}$ matrix multiplied by sigma squared, and this is what you have here.

(Refer Slide Time: 17:00)

Resolution of Error Sum of Squares

$$SS_E = (Y'Y - \hat{\beta}'X'Y)$$

$$SS_E = (Y'Y - \frac{(\sum_{i=1}^n Y_i)^2}{n}) - (\hat{\beta}'X'Y - \frac{(\sum_{i=1}^n Y_i)^2}{n})$$

$$SS_{\text{error}} = SS_{\text{total}} - SS_{\text{regression}}$$



By definition of the total sum of squares, the second term becomes the regression sum of squares

The sum of squares of the error is given by $Y'Y - \hat{\beta}'X'Y$, there is a typo let me just correct it. So we have sum of squares of error as $Y'Y - \frac{(\sum_{i=1}^n Y_i)^2}{n} - \hat{\beta}'X'Y + \frac{(\sum_{i=1}^n Y_i)^2}{n}$. The sum of squares of error is written as the sum of squares of total-sum of squares of regression, so this term here represents the regression sum of squares.

Here, we are knocking off $\frac{(\sum_{i=1}^n Y_i)^2}{n}$ to correspond to the sum of squares given by the intercept $\hat{\beta}_0$.

(Refer Slide Time: 18:02)

Analysis of Variance (ANOVA)

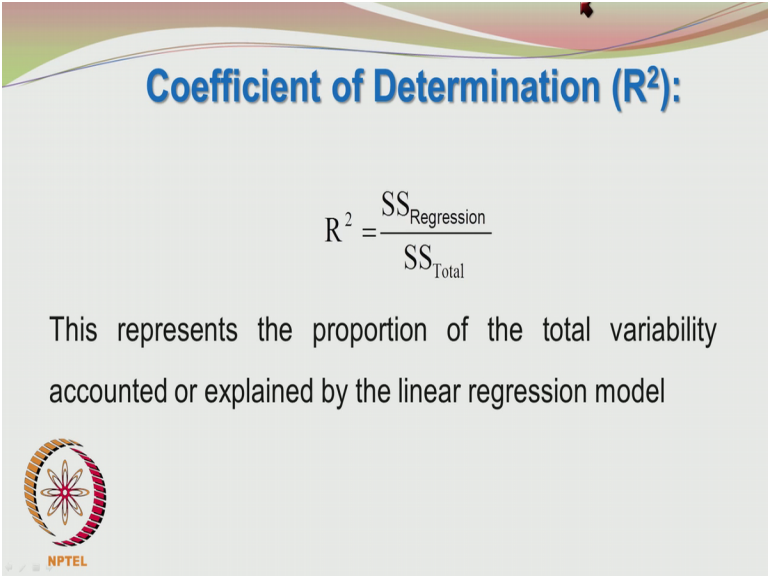
| Source of Variation | Sum of Squares | Degrees of Freedom | Mean Square | F_0 |
|---------------------|----------------|--------------------|-------------|---------------------|
| Regression | SS_R | k | MS_R | $\frac{MS_R}{MS_E}$ |
| Error or Residual | SS_E | $n-p$ | MS_E | |
| Total | SS_T | $n-1$ | | |



So we again have the analysis of variance table source of variation due to regression, and errors are residual sum of squares of error. So you have k and n-p degrees of freedom, and total sum of squares is having a degree of freedom of n-1. And again we find the mean square regression to mean square error, so we get sum of squares of regression by k which is mean square regression, sum of squares of error is given by is divided by n-p degrees of freedom to give mean square error.

The ratio of mean square regression to mean square error will give you the appropriate F0 value, which you can test to see whether this F0 value is lying in the rejection region or in the acceptance region.


(Refer Slide Time: 18:47)



Coefficient of Determination (R^2):

$$R^2 = \frac{SS_{\text{Regression}}}{SS_{\text{Total}}}$$

This represents the proportion of the total variability accounted or explained by the linear regression model



NPTEL

And then we also talk about R squared which is sum of squares regression/total sum of squares, this represents the proportion of the total variability accounted or explained by the linear regression model.

(Refer Slide Time: 19:00)

Adjusted R²:

Here Mean Squares of the error and model/total sum of squares are used.

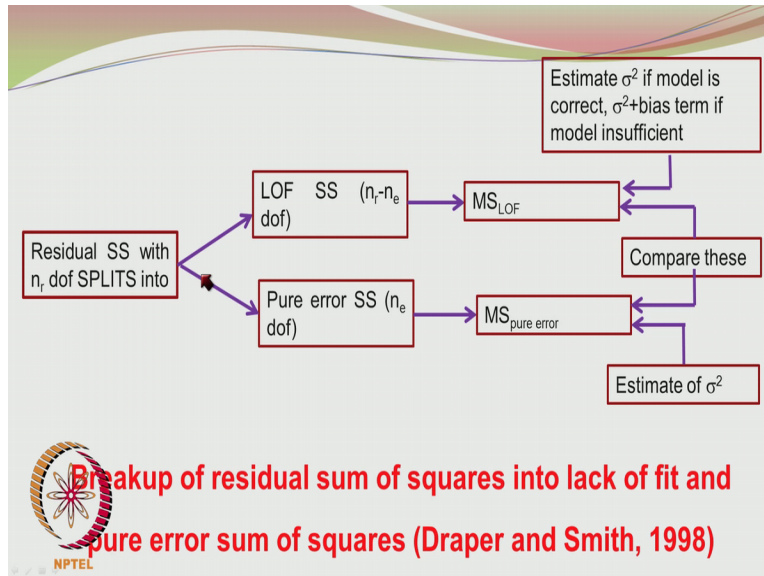
$$R^2_{\text{adj}} = 1 - \frac{\text{SS}_{\text{Error}} / n - p}{\text{SS}_{\text{Total}} / n - 1}$$



Here, an adjusted R square we have mean square of the error and model total sum of squares are used, and here we penalize the model for having too many parameters, R squared adjusted = 1 - sum of squares of error/n-p / sum of squares of total/n-1. Here, we are using the mean square of the error and the model sum of squares, so here we are dividing by n-p when the number of parameters increases, then n-p will decrease, 1/n-p will increase, so this term will increase and the R-squared value will go down.

So the suggested procedure is keep adding more parameters to your model until the R-squared adjusted starts to decrease, so it starts to penalize the model for having too many parameters. So if the adjusted R squared value also increases along with R squared upon adding of the parameter, then that particular parameter you have added to the model is making a effective contribution.

(Refer Slide Time: 20:16)



So now how to analyze for lack of fit, so what we do is we take the residual sum of squares. What is the residual sum of squares? It is the residual is defined as the balanced or leftover when you subtract the experimental response with the model prediction, so the balance is called as the residue, and the residual sum of squares are obtained by summing the square of these residues. So the residual sum of square is split into lack of fit sum of squares and the pure error sum of squares.

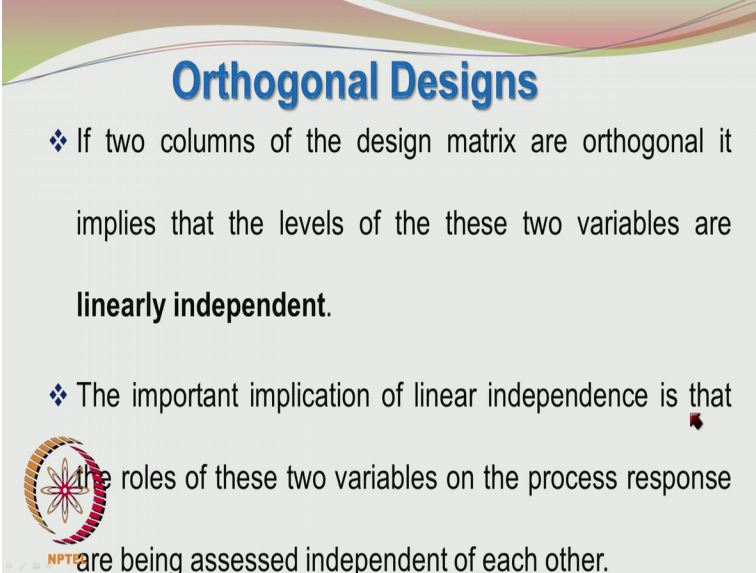
So the degrees of freedom of residual sum of squares n_r , and that is split into lack of fit sum of squares and pure error sum of squares, pure error sum of squares is having degrees of freedom n_e , and lack of fit sum of squares is having $n_r - n_e$. And when we divide the lack of fit sum of squares by $n_r - n_e$ we get mean square lack of fit, when we divide pure error sum of squares with n_e , we get mean squared pure error.

Now what we do is we compare the mean square lack of fit with mean squared pure error using $n_r - n_e$ numerator degrees of freedom and n_e denominator degrees of freedom. If the f test says that these 2 are comparable, and it does not lie in the rejection region, then we can say that the model does not have any lack of fit. Because the lack of fit sum of squares are comparable to the pure error sum of squares, there is no further incentive to develop the model further.

But on the other hand, if the mean square lack of fit is considerably higher from the mean square pure error, then the f statistic would be lying in the rejection region, then you have to conclude that there is sufficient scope for model expansion, and the lack of fit is significant. So we have to consider the addition of more terms in your equation. How do you get the pure error? The pure error is obtained by carrying out genuine repeats in your experimental runs.


You fix all the factors at certain value, and then repeat at the same value more than once, then you chose some other set of values for the factors repeat this experiments more than once at such factor settings, like this if you do you will be able to get genuine repeats which will help you to find the pure error sum of squares, and then you will get the mean square pure error. So this is a lack of fit test is a very important in linear regression analysis and it helps you to stop at a particular stage of model development.

(Refer Slide Time: 23:39)



Orthogonal Designs

- ❖ If two columns of the design matrix are orthogonal it implies that the levels of the these two variables are **linearly independent**.
- ❖ The important implication of linear independence is that the roles of these two variables on the process response are being assessed independent of each other.

 NPT

So the next important concept we discussed is about orthogonal designs, we touched upon the advantages of orthogonal designs a few slides back. And if the 2 columns of the design matrix are orthogonal, it implies the levels of these 2 factors are linearly independent. The important implication of linear independent is that the roles of the 2 variables on the process response are being assessed independent of each other.

When you are having factorial design, they are orthogonal based designs, and when you have factor A, factor B you can see that A factor is treated independent of the B factor, the contribution brought in by the A factor is independent of whether you are considering the B factor or not. Then we talked about AB interaction, how the experimental response goes from one value to another value upon changing A from lower level to upper level depends upon whether B was at a lower level or B was at a higher level.

Then the 2 factors A and B are said to interact, but we are talking about orthogonal designs and we say that A effect is found independent of the B effect, so when you have AB interaction the AB interaction effect is also found independent of the A effect and the B effect. So when you develop a model and you do not consider the interaction term between A and B, the factor A would still have a particular value.


Let us say the effect brought in by factor A is 20, if you consider interaction between A and B also, the effect of A would be till 20, so it does not matter whether AB is present in the model or not, the effect of A computed to be the same. Similarly, the effect of B maybe 10, the effect of B is 10 independent of whether A is present in the model or AB is present in the model. Suppose you have a full-fledged model and we are considering AB and AB.

Then the factor B would have an effect of let us say 10 units, if you do not have factor A and factor AB, we would still have 10 units, so this is very important. And the best way to understand this is to actually do a problem, develop a model for the orthogonal case, find the effects first you have only A, then you have B, then you have AB. Then you will find that A effect, B effect and AB effect are found independent of each other.

(Refer Slide Time: 26:15)

| | | | | | | |
|-------|---|-------|-------|-----------|---------|---------|
| $X =$ | 1 | X_1 | X_2 | $X_1 X_2$ | X_1^2 | X_2^2 |
| | 1 | -1 | -1 | 1 | 1 | 1 |
| | 1 | -1 | 1 | -1 | 1 | 1 |
| | 1 | 1 | -1 | -1 | 1 | 1 |
| | 1 | 1 | 1 | 1 | 1 | 1 |
| | 1 | 1 | 1 | 1 | 1 | 1 |

| | |
|--|---|
| Model | Greedy Model |
| $\hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_{12} X_1 X_2$ | $\hat{\beta}_{11} X_1^2 + \hat{\beta}_{22} X_2^2$ |



And in our model development we have to be careful, suppose we are having only 4 runs corresponding to the 2 power 2 factorial design, so we are having only 4 runs this is the column vector of 1's, this is the column vector corresponding to X_1 , corresponding to X_2 and $X_1 X_2$. So you are having all these so 4 independent runs and so you can estimate 4 independent parameters $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_{12}$.

However, if you want to estimate more parameters from your experimental design, for example you want to consider the quadratic terms also in your design like putting $\hat{\beta}_{11}$ and $\hat{\beta}_{22}$, you will find that the $\hat{\beta}_{11}$ corresponding to X_1 squared and $\hat{\beta}_{22}$ corresponding to X_2 square, X_1 squared is having all 1's and X_2 squared is also having all 1's. And if you look at the X matrix the 1's are corresponding with the column vector of 1's.

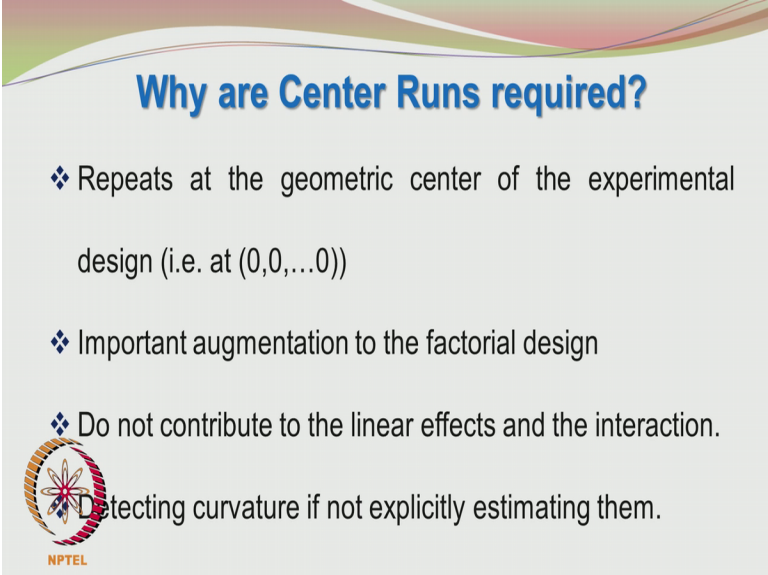
The columns are not linearly independent anymore, you can say that there are linearly dependent columns, there are 3 columns which are not linearly independent, and that would lead to all kinds of difficulties in your estimation of the parameters. So it is important that you are restrained yourself depending upon the size of the run, you do not try to fit too many parameters in your model, do not go for the greedy model.

Look at number of parameters you want to estimate and number of experimental runs available, you would think that number of experimental runs and number of parameters can be the same,

then it is no longer a regression analysis but procedure for solving A equations in A unknowns, so you will get the exact fit to the experimental data. Normally, we conduct experiments in such a way that we conduct a large number of experimental runs and then we estimate only a few parameters.

So that we have sufficient scope for accounting for experimental error and also for lack of fit test.


(Refer Slide Time: 28:28)



Why are Center Runs required?

- ❖ Repeats at the geometric center of the experimental design (i.e. at $(0,0,\dots,0)$)
- ❖ Important augmentation to the factorial design
- ❖ Do not contribute to the linear effects and the interaction.

Detecting curvature if not explicitly estimating them.



In certain experimental design strategies, we require center runs, we can do factorial design and you can repeat the experiments at factorial point factorial design points, but that may probably lead to large number of repetition of the runs and maybe also expensive. So rather than doing that you may want to carry out experiments at the center of your experimental design space, so that repeats are only conducted at the center of the experimental design space which is midway from all the experimental settings.

So that center runs are very important, because you are able to get an idea about the pure error. And it is an important augmentation to the factorial design, and it does not contribute to the linear effects and the interaction terms, and it helps to see qualitatively whether curvature is there or not.

(Refer Slide Time: 29:24)

Scaled Prediction Variance

- ❖ It is evident that the variance of prediction varies from point to point in the design space.
- ❖ It is also a function of $(X'X)^{-1}$ and hence the experimental design as well.
- ❖ It is a measure of how well one predicts with the model.



This is often used as a criterion for comparing different design strategies.

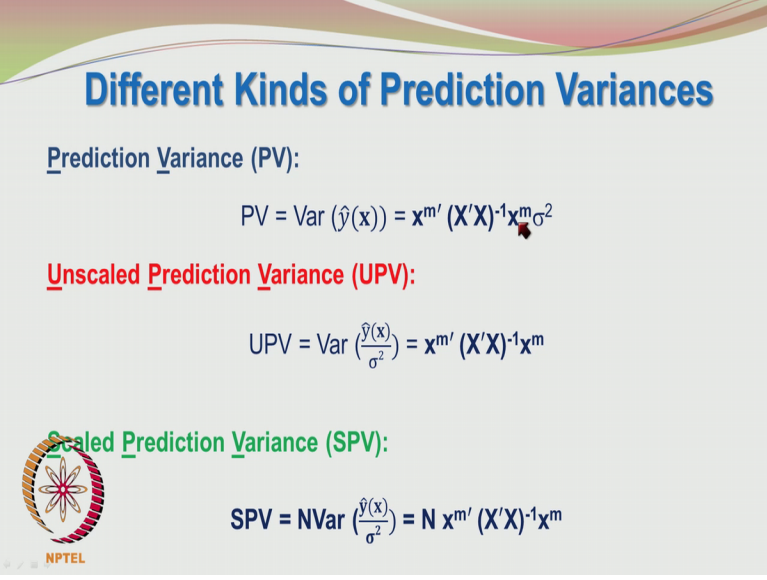
Another important parameter which people do not really understand or use in their design of experiments is the scaled prediction variance, there is a bit of linear algebra associated with it and that may be the reason why people do not really appreciate and utilize it. It is associated with the prediction nature of your model, so your model is going to predict certain values in the experimental designs space. And how good, and what is the quality of the model predictions?

If the model predictions have wide variance associated with them, then those predictions cannot be really relied upon, so we want to have a controlled scaled prediction of variance scaled prediction variance, for which we have to focus on the experimental design. For finding out the scaled prediction variance, we do not have to conduct the experiments as such, we have to choose upon the suitable experimental design strategy.

And by looking at the X matrix, we calculate X' , and then we calculate $X'X$ inverse, and then we identify a certain set of coordinates in the experimental design space, and then we compute the scaled prediction variance. And once we have the scaled prediction variance, we look at different points in the domain, and see whether the scaled production variance is kept under check in most of these points.

If there are certain points in the experimental design space where the scaled prediction variance suits up, then that particular design is not to be recommended.

(Refer Slide Time: 30:55)



Different Kinds of Prediction Variances

Prediction Variance (PV):


$$PV = \text{Var}(\hat{y}(x)) = \mathbf{x}^m{}' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}^m \sigma^2$$

Unscaled Prediction Variance (UPV):

$$UPV = \text{Var}\left(\frac{\hat{y}(x)}{\sigma^2}\right) = \mathbf{x}^m{}' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}^m$$

Scaled Prediction Variance (SPV):

$$SPV = N \text{Var}\left(\frac{\hat{y}(x)}{\sigma^2}\right) = N \mathbf{x}^m{}' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}^m$$

 NPTEL

So you can have different definitions for the prediction variance, in the first definition you have sigma squared and $\mathbf{x}^m{}' \mathbf{X}' \mathbf{X}^{-1} \mathbf{x}^m$ * sigma squared, \mathbf{x}^m is the co-ordinate point expanded into the model space. Please look at the appropriate slide for this definition. Then we also have the unscaled prediction variance, where you divide the prediction variance by sigma squared and you get $\mathbf{x}^m{}' \mathbf{X}' \mathbf{X}^{-1} \mathbf{x}^m$.

And then you also have the scaled prediction variance, where you multiply the unscaled prediction variance by the size of the run, you cannot officially make your scaled prediction variance as small as possible by increasing the number of runs, if you want to compare different designs then you have to put them on a common bases, and to do that you multiply by n which is the size of the run. And the scaled prediction variance is the very important parameter in statistical design comparisons, we have $SPV = n \mathbf{x}^m{}' \mathbf{X}' \mathbf{X}^{-1} \mathbf{x}^m$.

(Refer Slide Time: 32:00)


Estimated Prediction Variance

❖ Estimated Prediction Variance (EPV):

$$= \mathbf{x}^m' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}^m \hat{\sigma}^2$$

$$= \mathbf{x}^m' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}^m \text{MSE}$$

Standard Error of the estimated mean (SE)



$$\text{SE} = \sqrt{\mathbf{x}^m' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}^m (\text{MSE})}$$

Okay, coming to this next slide, we have what is called as estimated prediction variance, please note that we do not know the value of sigma squared the error variance. In such cases we try to replace the sigma squared the suitable error estimate, and that would be the mean square of the residuals, the residual sum of squares is divided by degrees of freedom for the residual sum of squares.

If you have n experimental points, and then you have p parameters including the intercept β_0 , β_1 so on. So you have p parameters note that $p=k+1$, where k is the number of regression coefficients okay, so you have $n-p$ as the degrees of freedom for the residual sum of squares, so you can divide the residual sum of squares by $n-p$ to get the mean square error, and that can be used instead of sigma squared.

We call it as sigma hat square to the note that it is an estimated one, and once you get the estimated prediction variance we can find the square root to get the standard error.

(Refer Slide Time: 33:13)

Why so much fuss on II order Models?

- ❖ Experimental design space (Response Surface) is no longer planar but may be marked by peaks and/or valleys
- ❖ II order models are required to estimate this response and enable the identification of optimum solution (if any).



We talked a lot about second order models, we looked at many research papers, we also talked about second order models. Why should there be so much fuss about second order models? Because experimental design space may no longer be planar, but or have only simple interactions, it may also be characterized by peaks and or valleys. And second order models are required to estimate this response and enabled identification of an optimal solution if any.

(Refer Slide Time: 33:48)

Why so much fuss on II order Models?

II order models are of the form

$$y = \beta_0 + \sum_{i=1}^k \beta_i x_i + \sum_{i < j} \sum_{j=2}^k \beta_{ij} x_i x_j + \sum_{i=1}^k \beta_{ii} x_i^2 + \varepsilon$$

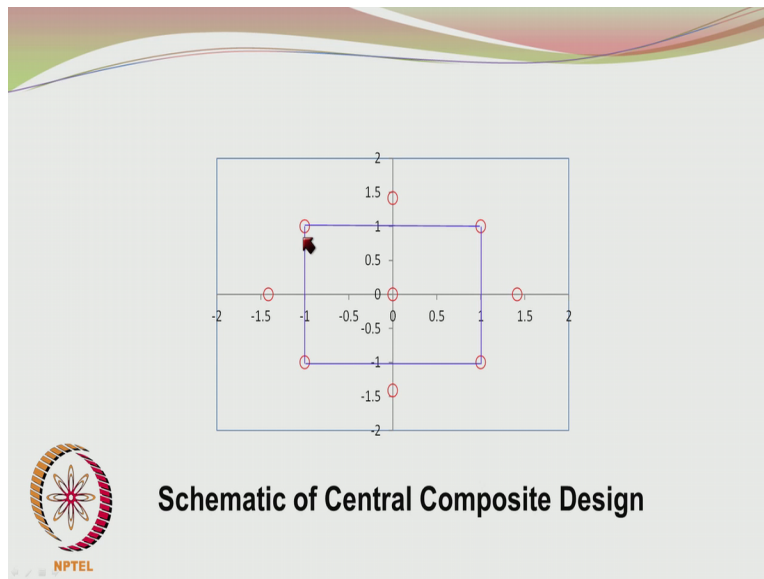
This equations requires estimation of



$$1 + (k) + {}^k C_2 + k = 1 + 2k + k(k-1)/2 \text{ parameters.}$$

So second order models are of the form $Y = \beta_0 + \sum_{i=1}^k \beta_i x_i + \sum_{i < j} \beta_{ij} x_i x_j + \sum_{i=1}^k \beta_{ii} x_i^2 + \varepsilon$, so you are accounting for the main affects, you are accounting for binary interactions and you are accounting for the quadratic terms. And this would require an estimation of totally $1 + k + {}^k C_2 + k$ which is $1 + 2k + k(k-1)/2$ parameters.

(Refer Slide Time: 34:18)



And one important second order design is the central composite design, where you have the regular factorial design and it is augmented by axial points. I am showing it for a design involving only 2 factors, so that I can represent it on a 2 dimensional diagram, so you have the regular factorial points and then you have the center points which help you to find the pure error and also tell qualitatively whether the curvature effects are there.

In addition to the center and factorial points you also have the axial points, and these axial points are important augmentation for the central composite design.

(Refer Slide Time: 35:03)

Roles played by center points in CCD

- ❖ help in the detection of II order or curvature effects
 $(\beta_{11} + \beta_{22})$ but not in their individual estimation.
- ❖ number of the central points decides the distribution of
S² in the region of interest


NPTEL

So the role played by center points in the central composite design are it helps in the detection of second order or curvature effects $\beta_{11} + \beta_{22}$, but not in their individual estimation. The number of central points decides the distribution of scaled prediction variance in the region of interest, so what I am trying to say is the prediction capability of your model may also depend upon the number of center points considered.

(Refer Slide Time: 35:31)

Axial points of Central Composite Design

- ❖ The axial terms contribute to the estimation of the individual pure quadratic effects significance.
- ❖ If axial points were not present, only the sum of the quadratic terms significance ($\sum_{i=1}^k \beta_{ii}$) could be gauged using the center points




The actual terms help in the contribution or help in the estimation rather of the individual pure quadratic effects, and if the axial points were not present, only the sum of the quadratic term significance $\beta_{11} + \beta_{22}$ could have been gauged using the center points.

(Refer Slide Time: 35:50)

Roles played by the axial points of Central Composite Design

- ❖ The axial points do not contribute to the estimation of the interaction effects
- ❖ The center points and the axial points contribute to the flexibility of the CCD.




And the axial points also do not help in the estimation of the interaction effects, and that is obtained from the factorial points, the center points and the axial points contribute to the flexibility of the central composite design.

(Refer Slide Time: 36:06)

Box-Behnken Design (BBD)

- ❖ A creative approach to planned experimentation involving relatively smaller number of runs.
- ❖ An important alternative to central composite designs (CCD).
- ❖ Box-Behnken design involves balanced incomplete block design.

An example of an balanced incomplete block design for 3 treatments are given below




An important alternative to the CCD is the BBD, an alternative to the central composite design as the Box-Behnken design. It is a creative approach to planned experimentation involves relatively smaller number of runs, it is an involves balanced incomplete block design.

(Refer Slide Time: 36:28)

BBD (k=3) MINITAB® Design with 3 center points

| A | B | C |
|----|----|----|
| -1 | -1 | 0 |
| 1 | -1 | 0 |
| -1 | 1 | 0 |
| 1 | 1 | 0 |
| -1 | 0 | -1 |
| 1 | 0 | -1 |
| -1 | 0 | 1 |
| 1 | 0 | 1 |
| 0 | -1 | -1 |
| 0 | 1 | -1 |
| 0 | -1 | 1 |
| 0 | 1 | 1 |
| 0 | 0 | 0 |
| 0 | 0 | 0 |
| 0 | 0 | 0 |



so for 3 factors you have the Box-Behnken design, here you have a regular 2 power 2 factorial for factor AB, and C is kept at the center point. In the next phase you leave B at the center point

and then construct a 2 power 2 factorial for A and C. For the next phase you consider B and C and then you have A at the center. After you have exhausted all the 3 combinations, you then have a set of center points defined here.

(Refer Slide Time: 36:59)

Box-Behnken Design for k=6 factors

| | x_1 | x_2 | x_3 | x_4 | x_5 | x_6 |
|-------|---------|---------|---------|---------|---------|---------|
| $D =$ | ± 1 | ± 1 | 0 | ± 1 | 0 | 0 |
| | 0 | ± 1 | ± 1 | 0 | ± 1 | 0 |
| | 0 | 0 | ± 1 | ± 1 | 0 | ± 1 |
| | ± 1 | 0 | 0 | ± 1 | ± 1 | 0 |
| | 0 | ± 1 | 0 | 0 | ± 1 | ± 1 |
| | ± 1 | 0 | ± 1 | 0 | 0 | ± 1 |
| | 0 | 0 | 0 | 0 | 0 | 0 |

Each row in the above design matrix (except the last one) refers to 8 possible combinations of a 2^3 design. Unlike the previous cases (k=3, 4 and 5) the factorial structure now involves the 2^3 design.

So when you have large number of factors let us say 6 factors instead of going for a 2 power to design for a certain subset of factors you go for a 2 power 3 design. So first block or first phase you consider x_2 x_3 and x_5 , here you construct 2 power 3 design involving these factors. Then after doing out the 8 settings corresponding to x_1 x_2 and x_4 , you go to x_2 x_3 and x_4 carry out the 2 power3 design. In such a case all the remaining factors would be at the center values

Similarly, you go around taking 3 factors at a time, and construct all your 2 power 3 designs out of these factors. So the important thing to notice you are considering 3 factors out of the 6 at any given time, so those 3 factors would constitute a 2 power 3 factorial design, whereas the remaining factors be at the center values. Finally, after exhausting all the combinations, you come to the center runs, where all the factors are kept at the center values.

(Refer Slide Time: 38:12)

Response Surface Methodology

In any experimental work, an important objective could be to identify optimum levels of the various factors which will maximize/minimize a suitable objective for e.g. reaction yield, conversion, process time, energy consumed etc.



And once we know how to do CCD and BBD, we can then do the response surface methodology where the objective is to find the optimum value.

(Refer Slide Time: 38:25)

Response Surface Methodology

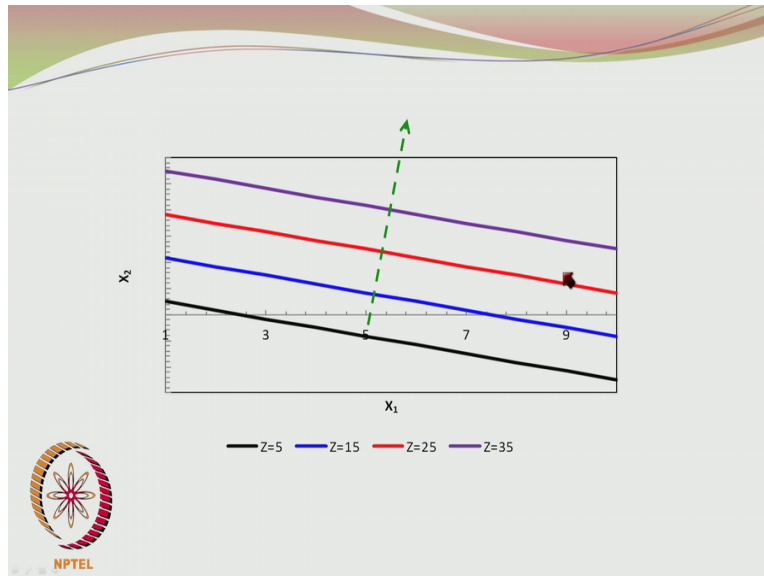
❖ The current level of operation may usually be far away from the optimum and we cannot afford to wander in the wilderness of n -dimensional experimental variable space hoping to eventually reach the optimum.



Response Surface Methodology deals with identifying optimum settings of the factors in a systematic manner.

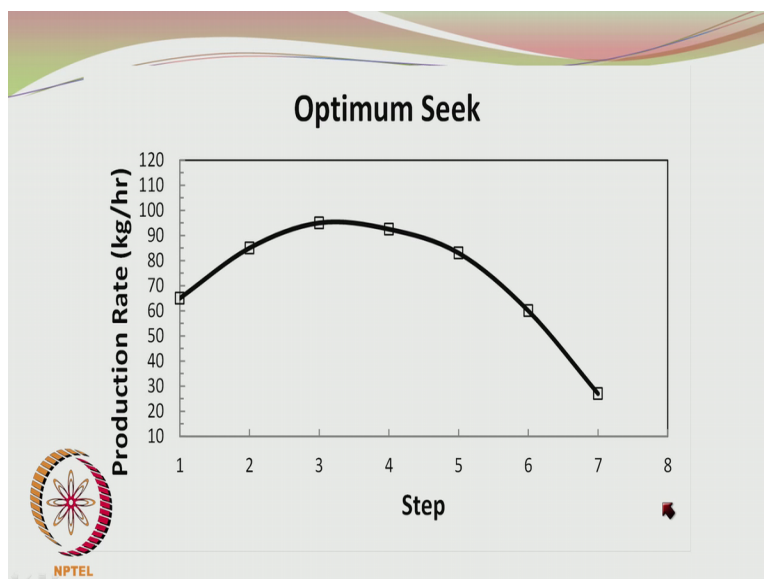
So the current level of operation may be very far away from the optimum, and we cannot afford to wander around in the n dimensional space wasting resources, manpower and time. So we need a structured and we need a well-thought-out procedure to quickly progress towards the optimal solution, for this we use the method of steepest ascent.

(Refer Slide Time: 38:50)



So I am demonstrating this method for 2 variables, here you are having X_1 X_2 , and this is responses are obtained from 2 power 2 factorial design, and we are showing no interactions the contours are not curved, but they are linear. And to proceed along the direction of steepest ascent you go in the direction perpendicular as shown here perpendicular to the contour lines. These are response lines; we are going in direction perpendicular to them.


(Refer Slide Time: 39:21)



So once we keep doing experiments, please remember that we cannot use the developed model to identify the values or outcomes along the direction of steepest ascent, but we actually do experiments out of this design space, and we keep doing the experiments until we reach a stage where we find the values passing through an optimum. Here, we construct a central composite

design around this optimal point, and then evaluate all the parameters in the model. And we also see whether the optimum is minimum or maximum or a saddle point.

(Refer Slide Time: 39:56)



| A | B | P |
|--------|--------|------|
| -1 | -1 | 85 |
| 1 | -1 | 126 |
| -1 | 1 | 64 |
| 1 | 1 | 92.5 |
| -1.414 | 0 | 77 |
| 1.414 | 0 | 123 |
| 0 | -1.414 | 101 |
| 0 | 1.414 | 66.1 |
| 0 | 0 | 93.5 |
| 0 | 0 | 95 |
| 0 | 0 | 91 |
| 0 | 0 | 96 |

So this central composite design around the optimal point is shown, for example here you have the factorial points, here you have the axial points, finally the repeats.


(Refer Slide Time: 40:09)

Second Order Model

The second order response may now be expressed as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \epsilon$$

The predicted expression is given below

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_{12} X_1 X_2 + \hat{\beta}_{11} X_1^2 + \hat{\beta}_{22} X_2^2$$


So now you can fit a second order model as given by the following equation.

(Refer Slide Time: 40:16)

Location of Stationary Point

The partial derivatives of the predicted response with respect to the variables X_1 and X_2 must be set to zero and solving the resulting equations, the coordinates of the stationary point may be identified.



$$\frac{\partial \hat{Y}}{\partial X_k} = 0 \quad k = 1, 2$$

And once you have identified all the model parameters, then you have to identify the stationary point. You have to first locate where the stationary point is, the stationary point is where the partial derivative of \hat{Y} with respect to X_k , where X_k can be X_1 or X_2 depending on the number of independent factors you have considered that all the partial derivatives should be set to 0. So from the identified model equation you can set all the partial derivatives with respect to the X values to be 0, solve the resulting set of perhaps non-linear algebraic equations, and find the set of stationary conditions.

(Refer Slide Time: 40:56)

Location of Stationary Point in Matrix Notation

Expressing the second order relation in matrix notation as

$$\hat{Y} = \hat{\beta}_0 + \mathbf{X}'\mathbf{b} + \mathbf{X}'\mathbf{B}\mathbf{X}$$



The stationary point is the solution to the equation

$$\frac{\partial \hat{Y}}{\partial \mathbf{X}} = \mathbf{b} + 2\mathbf{B}\mathbf{X} = \mathbf{0}$$




There is another way of doing it that is from the matrix method, here you identify what is called as a small \mathbf{b} matrix and the capital \mathbf{B} matrix. What is the small \mathbf{b} matrix and the capital \mathbf{B} matrix?

(Refer Slide Time: 41:04)

Matrix Notation Explained

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{bmatrix}$$

$$B = \begin{bmatrix} \hat{\beta}_{11} & \frac{\hat{\beta}_{12}}{2} & \dots & \frac{\hat{\beta}_{1k}}{2} \\ & \hat{\beta}_{22} & \dots & \frac{\hat{\beta}_{2k}}{2} \\ & & \ddots & \\ \vdots & \vdots & \vdots & \hat{\beta}_{kk} \end{bmatrix}$$

$$b = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}$$


The small B matrix is the column vector of all the main factor parameters, and then you have the capital B matrix whose structure is interesting, here along the diagonal we have the coefficients for the quadratic terms, and along the off diagonal we have one half of the interaction term. So this matrix is symmetric for example beta 12 location will be the 21 location, so here also you have beta hat 12/2 and in 12 or 21 location also you will have beta hat 12/2.

So this b matrix is a symmetric matrix $b_{ij} = b_{ji}$ that is the matrix being symmetric, in such a case to account for the interaction effect twice we are dividing it by 2. On the other hand, the diagonal terms are all the quadratic coefficients beta hat 11, beta hat 22 and beta hat kk. So once you have the capital B and the small b matrices, evaluating the locating the stationary point is quite straight forward, this is obtained by the solution to the equation $b + 2BX = 0$, remember here we are dealing with matrices.

(Refer Slide Time: 42:34)

Location of Stationary Point in Matrix Notation

Solving the above equation we get

$$\mathbf{X}_S = -\frac{1}{2}\mathbf{B}^{-1}\mathbf{b}.$$

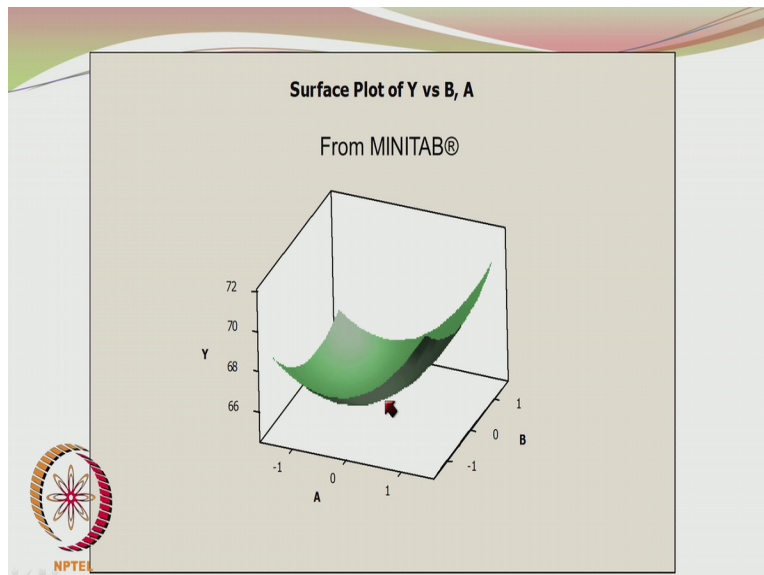
The predicted value of Y at this stationary point is

$$\hat{Y}_S = \hat{\beta}_0 + \frac{1}{2}\mathbf{X}_S'\mathbf{b}$$



So once you solve the above equation, we get the stationary point coordinates as $-\frac{1}{2} \mathbf{B}^{-1} \mathbf{b}$, so the predicted value of Y at the stationary point is $\hat{Y}_S = \hat{\beta}_0 + \frac{1}{2} \mathbf{X}_S' \mathbf{b}$, where \mathbf{X}_S is given by this equation.

(Refer Slide Time: 42:54)



So you can look at the response surface, and it can see for this particular example it passes through a minimum, so we have a minimum solution here.

(Refer Slide Time: 43:04)

Nature of the Optimum Solution

Check the eigenvalues of **B**.

If all the eigenvalues are positive (negative), and the stationary point is within the region of exploration, the stationary point is a minimum (maximum).



So in order to identify whether the identified stationary point is maximum or minimum or a saddle point, we have to check their eigenvalues, if all the eigenvalues are positive then the obtained solution the stationary point corresponds to a minimum, and if all the eigenvalues are negative the identified stationary point is a maximum. And if you have some eigenvalues positive and some eigenvalues negative, then it corresponds to a saddle point another true optimal location.

(Refer Slide Time: 43:33)

A Recap

The **CCD** and **BBD** designs augmented with center runs are extremely popular among practitioners. Recall that the statistically designed experiments are evaluated on the basis of factors such as



So the CCD and BBD runs are augmented with center runs, they are very popular among the practitioners.

(Refer Slide Time: 43:44)

Evaluation of Statistical Designs

- a. Good fit to the data – high value of R^2 , adjusted R^2 , low PRESS etc.
- b. Allowing test for lack of fit i.e. possible model expansion
- c. Allowing sequential construction of models of increasing



And the statistical designed experiments are evaluated based on the factors such as good fit to the data, it should allow for lack of fit and it should allow for sequential construction of models of increasing order or complexity.

(Refer Slide Time: 43:53)

A Recap

- a. Estimate pure error through repeats especially at center points
- b. Robust by being insensitive to the presence of outliers in the data



c. Cost effective i.e. involving less number of runs

d. Provides a good distribution of scaled prediction variance

And you should have enough repeats of the center, to have an estimate of the pure error, and it should be robust to the presence of outliers in the data, and it should be cost effective that means it should involve less number of runs, and it provides us good prediction of scaled prediction variance. So this completes our discussion on experimental design strategies, we have covered quite a lot of ground that should be enough for researchers, who are planning to do designed experiments.

We have covered sufficient theory as well as done significant number of problems. I request you to go through them, and solve as many problems on your own, and then also become familiar with statistical software like MINITAB design expert and so on. So I wish you all the best in your application of this statistic principles we learnt for your experimental program, thanks for your attention.