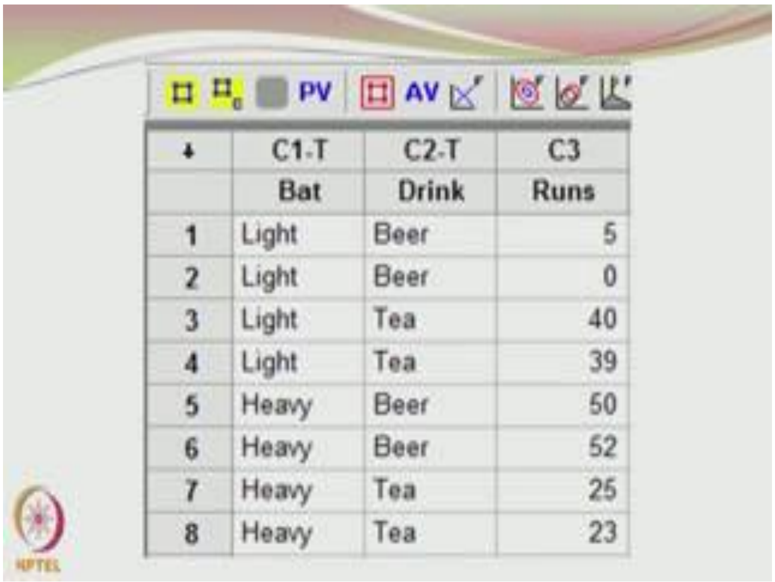
Introduction to Research
Prof. Kannan
Department of Chemical Engineering
Indian Institute of Technology, Madras

Lecture – 04
Introduction to Design of Experiments

Welcome back to this continuation on Introduction to Design of Experiments. We will start with the simple example definitely a fictitious one.

(Refer Slide Time: 00:28)

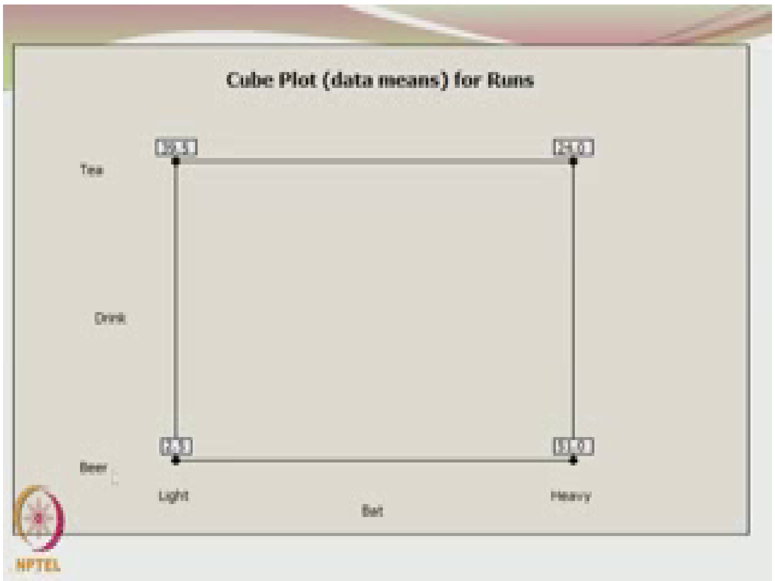


↓	C1-T	C2-T	C3
	Bat	Drink	Runs
1	Light	Beer	5
2	Light	Beer	0
3	Light	Tea	40
4	Light	Tea	39
5	Heavy	Beer	50
6	Heavy	Beer	52
7	Heavy	Tea	25
8	Heavy	Tea	23

So, I am presenting data in a table which contains the runs scored by a batsman in the game of cricket, I hope all of you know this game and know its rules. There are two innings and the batsman gets two chances in a match, first innings and second innings. So, the scores are listed in the last column. And in this experiment, what we are trying to do is find out whether the type of drink a person has or the batsman has before coming out to play influences his performance. Another factor we are considering is the weight of the bat he uses. Normally, we will think that the type of drink he has had or the weight of the bat he is using will be independent, but it may so happen that they may affect one another or there may be an interaction between the two factors that is what we are going to demonstrate in this example.

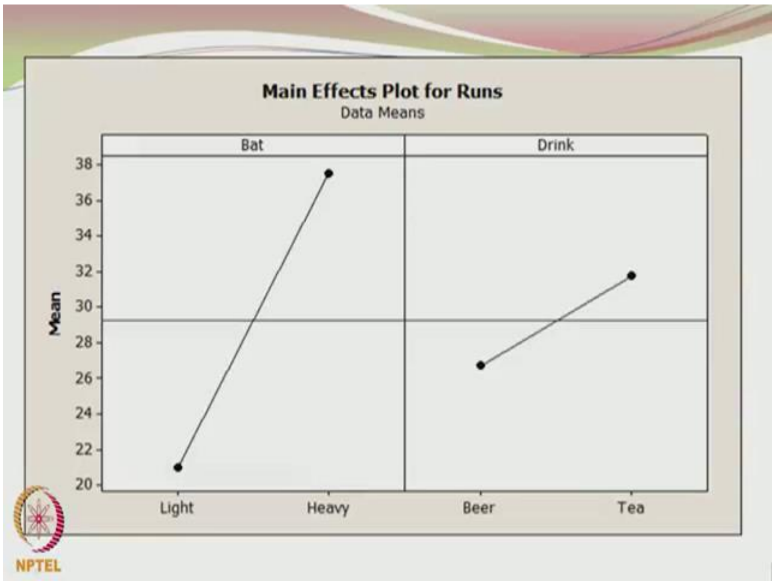
This example has been analyzed using mini tab software; this is a statistical analysis software. It also includes the design of experiments. You can download a trial version and check this for yourself. So, we have the type of bat or the weight of the bat in the first or the second column, you can have light bat and heavy bat, and then the batsman drinks beer, I don't know whether it is allowed in India by I think in foreign countries these things are pretty common. So, the batsman might have had beer before coming out to play, just to put him in good mood or he might have had tea, if he is of the serious type. So, these are the runs scored and you can see that there is a considerable variance in the run scored by the batsman.

(Refer Slide Time: 02:34)



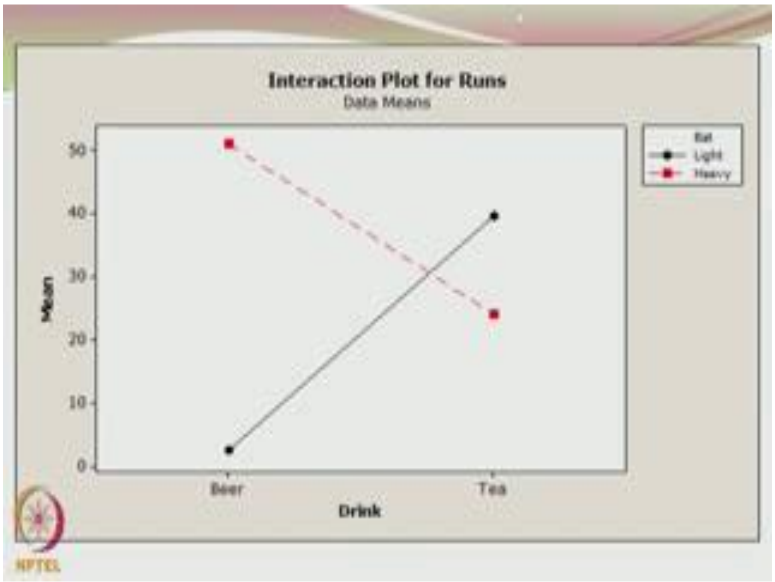
And you can see that the type of the drinkers represented on the vertical scale, and the type of the bat or the weight of the bat is shown on the horizontal scale. So, this is the average performance of the batsman, when he is having a light bat and beer and this is the average performance of the batsman, when he is having a light bat and tea. You can see 24 is the average performance of the batsman when he is having tea and heavy bat; and then beer and heavy bat the average performance is the highest at 51.

(Refer Slide Time: 03:19)



So, let’s analyze this further. You can see that this is the main effects plot; this is the effect of the bat alone and the effect of the drink alone. So, it might appear when a batsman is going from a light bat to heavy bat, the average performance improves, and when he is going from beer to tea the average performance improves, but this just not tell us the complete picture.

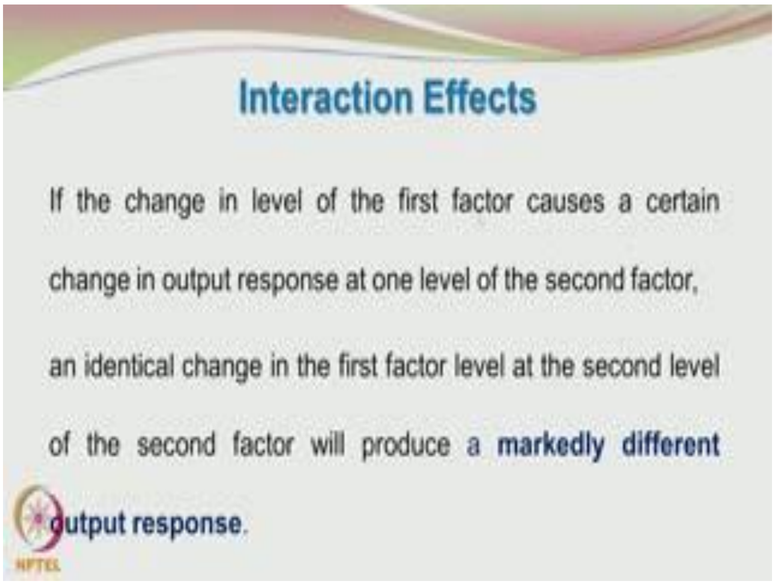
(Refer Slide Time: 03:50)



Okay so, we have to see the interaction. Now this is very interesting. When a batsman is going from beer to tea the performance is increasing, when he is using a light bat; but when he is using a heavy bat, the combination of beer and heavy bat leads to better performance than the combination of heavy bat and tea okay. So obviously, under the influence of the alcohol, the batsman is able to swing more wildly and connects more often than not and then scores more runs who knows, but this is an example where there is a strong interaction between the type of the drink and the weight of the bat.

If there had been no interaction between the two variables then the two lines would have been parallel. So, when you do your design of experiments, don't look at only the single effect or the single factor plots; look at the interaction plots. And if the interaction plots are parallel then the variables are not interacting with one another, but if they intersect or they approach one another or they diverge from each other, then it shows that there is an interaction. So, it's a very interesting and important aspect to consider in the design of experiments.

(Refer Slide Time: 05:21)

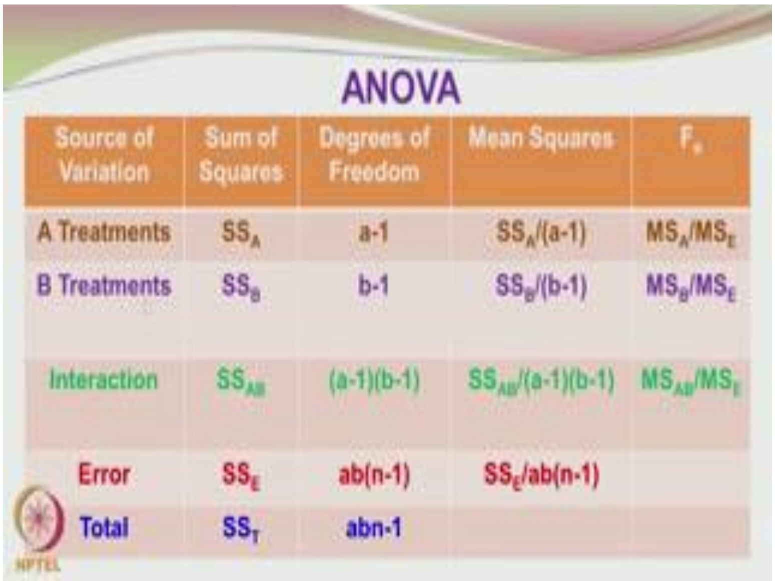


Interaction Effects

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, let us define the 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 second factor will produce a markedly different output response. So, you may want to read this couple of times to understand it fully. All it says is when you go from one level to another level for a particular factor, the response depends upon what level you have fixed for the second factor. If the level you have fixed for the second factor makes a difference then the two variables are set to interact.

(Refer Slide Time: 06:16)



ANOVA

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F_0
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		

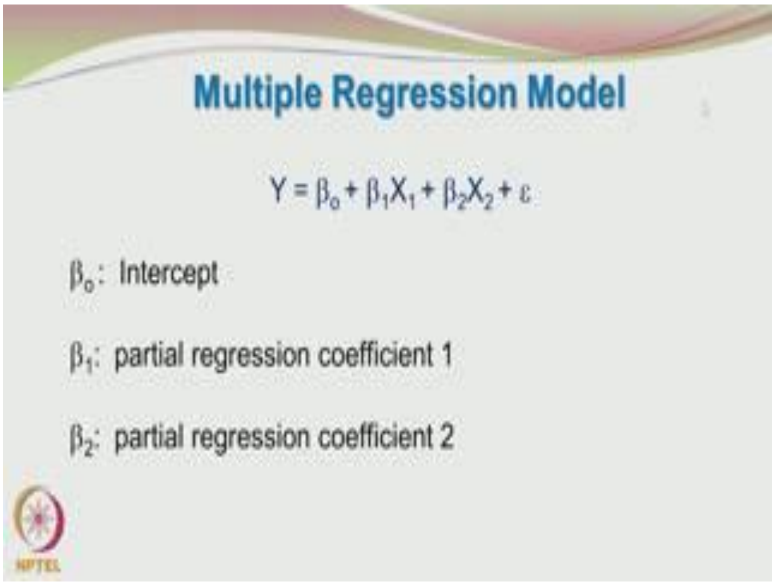
By now we know what is meant by ANOVA table. So, we have listed the source of variation; A treatments, B treatments and then the interaction between the two treatments. The difference between the earlier ANOVA table and the current one is that now we are dealing with two variables, and we also have to consider the interactions between the two variables. So, we find the sum of squares of the deviations for the A treatment, for the B treatment and also for the interaction. And we scale each of the sum of squares by the appropriate degrees of freedom; A is the number of levels for factor-a, B is the number of levels for factor-b. So, we subtract 1 from a and 1 from b and then we divide the sum of squares of a by a minus 1 sum of squares of b by b minus 1 to get the mean squares that is what is presented here.

Similarly, we also get to know that the degrees of freedom for the interaction between a and b is a minus 1 times b minus 1 and so you have the mean squares as sum of squares of ab by a minus 1 into b minus 1. So, these are representing the variations caused by the actual experiments. Whereas, we also have error which is based on the variation shown when you repeat these experiments, either you repeat all the experiments or you repeat the experiments at the center point. What is really meant by center point, right now do not worry, we will come to it a bit later. We can look at the individual mean squares for example, compare the treatment a, with the variability shown by the error and get the appropriate f value. Obviously, if a treatment is very important then it will completely dominate, what will completely dominate? The mean squares of a will completely dominate over that of the error and so you will have a very high F value.

Similarly for b, if the interaction effect is not important then what might happen is the mean square for ab would be comparable to that of the error, and you will get a value in the order of magnitude of 1 and that would indicate from the F probability chart that this variable is not important okay. There is something called as the p value, but I would leave that for you to understand yourself. It is quite straight forward; it has something to do with the type one error and okay no need to keep it in suspense; p value refers to the probability of committing the type one error. And obviously, if the F value is very, very high the p value would be very small, because you are going on to the tail end of the f distribution and the probability would be very small and so the probability of committing a type one error would be very small. So, I hope you know what is meant by type one error, if you don’t remember please go back to the previous lectures and then understand.

Now that we have done some experiments and you analyzed them, carried out the experiments that repeats, put them in statistical software like mini tab, got the response plots, and you also know which variables are significant which variables are not significant and so on. There is another interesting thing you can do with your experimental data. Many researchers and students get pride if they develop their own mathematical model. So, we are not going to discuss about very fundamental first principles model, we are going to discuss about very simple models, these are called empirical polynomial regression models. And even though, they may not have much of a theoretical background they provide useful formulae, if you want to call it like that which can be used to predict the response at different factor settings.

(Refer Slide Time: 10:48)



Multiple Regression Model

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

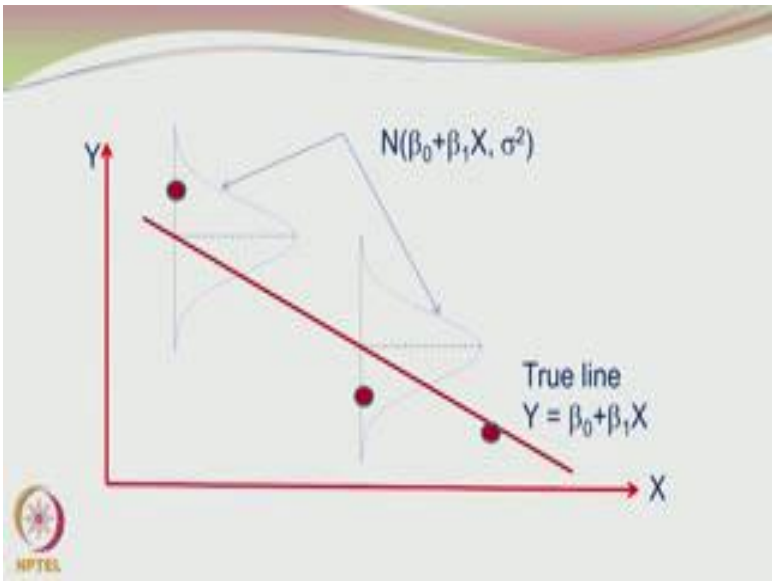
β_0: Intercept

β_1: partial regression coefficient 1

β_2: partial regression coefficient 2

So, now, we are getting into model building, which is based on experimental data. So, you have the process response, y is equal to beta naught plus beta 1 X 1 plus beta 2 X 2 plus epsilon, here X 1 and X 2 are the variables or the factors. These are values taken by the variable 1 and the value were taken by variable 2. So, X 1 represents variable 1, and X 2 represents variable 2 and beta 1 and beta 2 are called as the partial regression coefficients associated with these variables 1 and 2. Beta naught is not associated with the any variable it has standing on its own it is the so called intercept, for example, if you write y is equal to m X plus b then the b is the intercept. So, beta naught is referred to commonly as the intercept, whereas the beta 1 and beta 2 are referred to here as the partial regression coefficients.

(Refer Slide Time: 11:52)



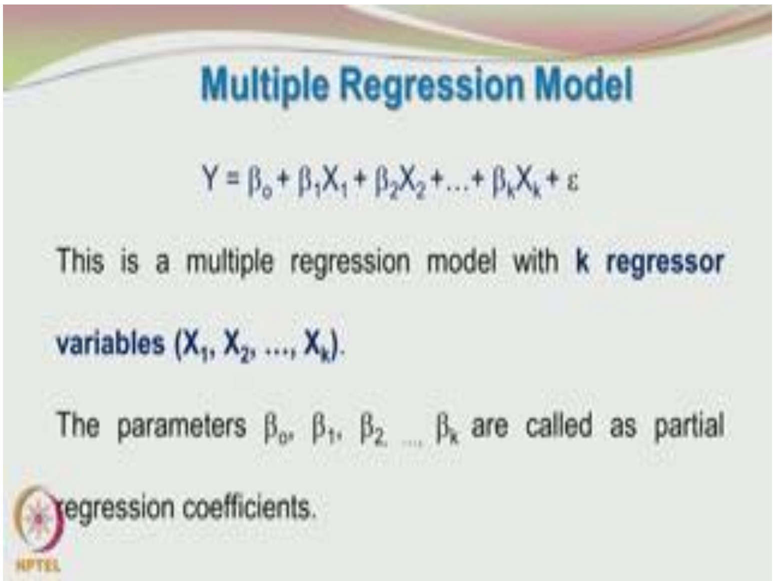
This is a very interesting diagram. You can see that the experimental data points are sort of scattered around a mean straight line. This is a true line okay, but the data points do not lie on the straight line that would have been an ideal occurrence or a intentional occurrence, but under usual circumstances, we can always find scatter of the experimental data around the straight line. So, beta naught and beta 1 are the true values of the parameters which unfortunately we do not know; and since we do not know beta naught and beta 1, we also do not know the value of the response y.

Here, for the purpose of illustration, I am just taking a case where there is only one factor or variable influencing the response. So, we have these experimental responses as being distributed around a mean value given by beta naught plus beta 1 X. So, if you take the level for X as X 1 at this point then the actual true, but unknown value would be given by beta naught plus beta 1 X a, calling this point or calling this variable setting as X a. So, this is the true response, but the experimental data because of the randomly varying error component would be scattered and so this is where the point lies.

We assumed that the errors are varying randomly with mean 0, and constant variance, so when you super impose the error effect on the response, the response is also normally distributed, but the mean value given by the true line equation and the variance given by sigma squared. Sigma squared is the error variance, which is assumed to be constant. Similarly, at the second location or second level for X, let us call it as b you find of response line below the true line. And by random fluctuations, the third value lies pretty close to the true response. So, in reality, we may never get to know, what is the true value of the parameter beta 1 and the true value of the parameter beta naught. So, through a linear regression analysis and your experimental data, we can only estimate the parameters beta naught and beta 1.

So, each time you do an experiment, you will get different responses and each time you fit the data to the model, you may get different values of the parameters. So, this is because the random error component is influencing a response in several unknown ways which you cannot predict a priori and so that’s the reason, why your model parameters are coming slightly differently each time you do the experiment. So, you do not have to worry, if you do not get the same model parameters if you do the experiments several times; it is expected that this is going to happen.

(Refer Slide Time: 15:22)



Multiple Regression Model

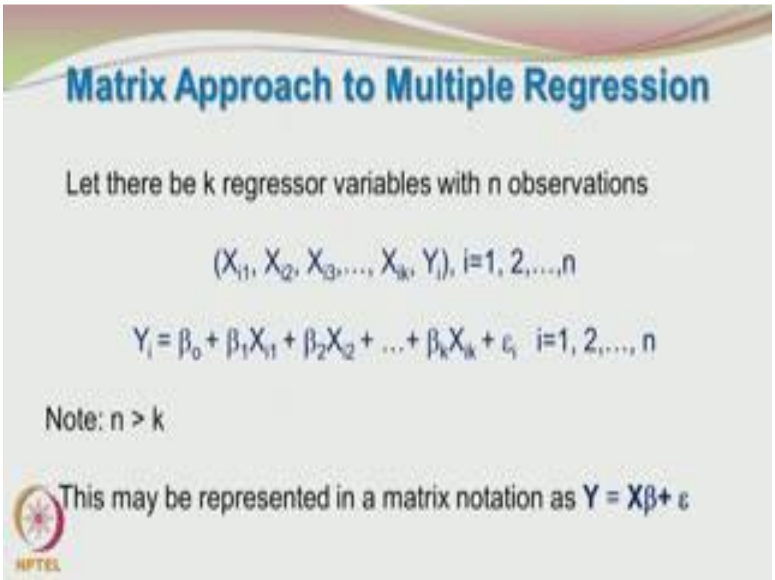
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

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

The parameters β_0, β_1, β_2 ... β_k are called as partial regression coefficients.

We can have any number of variables okay; you do not get restrict yourself to one variable or two variables at the most. You can analyze several variables together and if you use the linear algebra concepts, it becomes very straightforward. So, when you have multiple variables, you have the multiple regression model but k regressor variables X 1, X 2, so on to X k. And again to retreat beta naught beta 1, beta 2, so on to beta k are called as partial regression coefficients. So, how many partial regression coefficients you have? It is k plus 1; we have k regressor variables and we have hence k partial regression coefficients associated with these k regressor variables. In addition to these, we also have the intercept which is beta naught, so we have k plus 1, unknown parameters we want to estimate; in some places we may refer to this k plus 1 as p.

(Refer Slide Time: 16:30)



Matrix Approach to Multiple Regression

Let there be k regressor variables with n observations

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

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

Note: $n > k$

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

Now, you have as I said earlier k regressor variables, but you are not going to do a single experiment involving different settings of these k regressor variables. You might to want to fix some variables and vary some variables; next time you may fix those variables and then vary some other set of variables so on. So, let us assume that you do this in a very systematic manner, and you have n such experiments or n such runs. So, you have now two subscripts i and j; one subscript referring to the run order okay; i is the run order, you can go from 1 to n and the j subscript referring to the index of the regressor variable okay. So, if I say X i 2 then I am referring to the ith run and the second regressor variable, I think now that should be clear to you. So, the ith run's response Y i is given in this way, beta naught plus beta 1 X i 1 plus beta 2 X i 2 plus so on to beta k X i k plus epsilon i, i running from 1 to n.

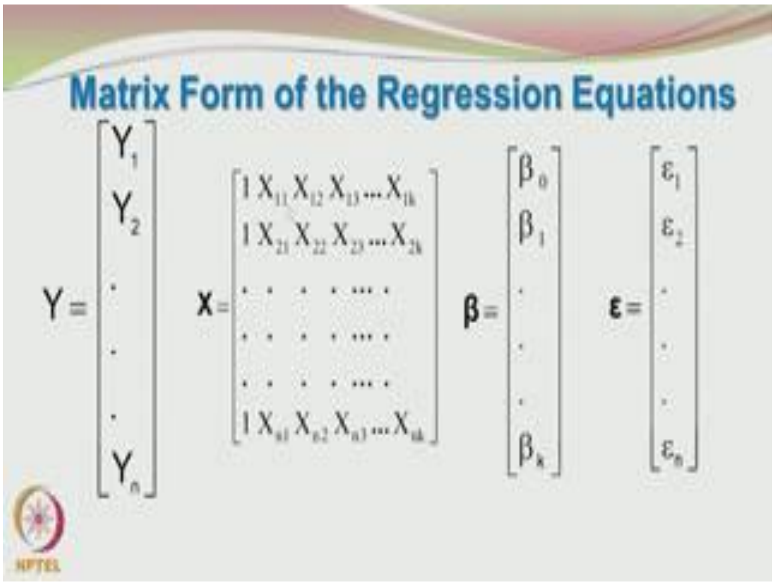
Obviously, n should be greater than k, if n is less than k you don’t have enough data to analyze problems satisfactorily. If you have n is equal to k even that is not sufficient because the total number of unknown parameters you want to estimate is in fact k plus 1, the partial regression coefficients for each of the regressor variable and then you also have the intercept. So, you have k plus 1 parameters to estimate and you need at least k plus 1 runs. So, if you have k plus 1 runs, and if you have k plus 1 parameters to estimate then you have an exact problem in the sense that your model will fit the experimental data perfectly because you are essentially solving an equation involving k plus 1 variables and k plus 1 unknowns.

But this is not really good because it shows that you have scaled up your regression model, you have introduced many terms in the regression model, and it is fitting your data properly. But this is not really good because you are taking up too many of the parameters, and the regression model may become also very unwieldy and complex to use.

The beauty of a regression model is appreciated when it is very simple and even with the few parameters you are able to explain the process reasonably well. Anyway, we know that this is an empirical model and it is only applicable in the very narrow region occupied by the ranges of these variables; you cannot extrapolate this model all the time beyond the ranges of these variables. So, it’s a very simple model. So, better leave it as a simple model involving two or three parameters, and then see how this explains your experimental trend.

Now, if you are familiar with linear algebra, you will easily appreciate what I am going to say it. Now, but even if you don’t know linear algebra or you have forgotten, it is very good time and opportunity to go through this concepts quickly and then understand what I am going to say next. In fact, linear algebra you don’t have to do research or going to very great depth, all you need to know is how to add matrices, subtract matrices, multiply matrices, and represent matrices first and then also some brief ideas about how to find Eigen values and Eigen vectors.

(Refer Slide Time: 20:47)



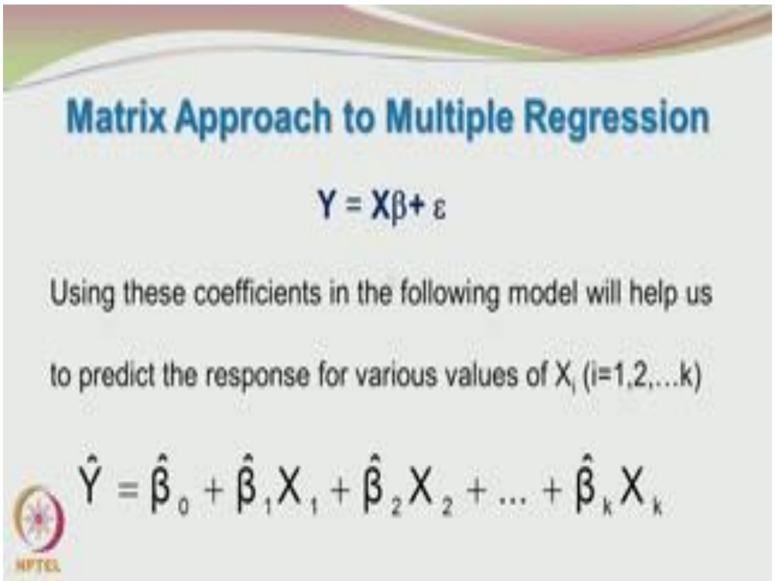
Matrix Form of the Regression Equations

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

So, when you represent this in the matrix notation, you get nice compact forms. You have the Y column vector, which is comprising of the n experimental observations running from Y 1, Y 2 so on to Y n. Then you also have the X matrix, it’s a very important matrix which is having a column of one’s this corresponds to the intercept. I will come to this in a moment then you have X 11, X 11 represents the first run and the setting of the first factor or the first variable X 1. X 12 again refer’s to the first run, but it refers to the setting of the second factor or the second experimental variable X 2 and so on. Since we have k factors, this one runs all the way from X 11 to X 1 k; similarly you have the other entries in this matrix.

For example, X n 2 will refer to the nth run and the setting of the second experimental factor or the experimental variable X 2. Now, let us look at the beta column vector. By now you should know that this comprises of beta naught which is the intercept and then beta 1 the partial regression coefficient to X 1 so on to beta k, which is a partial regression coefficient to X k. And this is the error term - the mysterious error term which is running from epsilon 1 epsilon 2 so on to epsilon n right.

(Refer Slide Time: 22:27)



Matrix Approach to Multiple Regression

$$Y = X\beta + \varepsilon$$

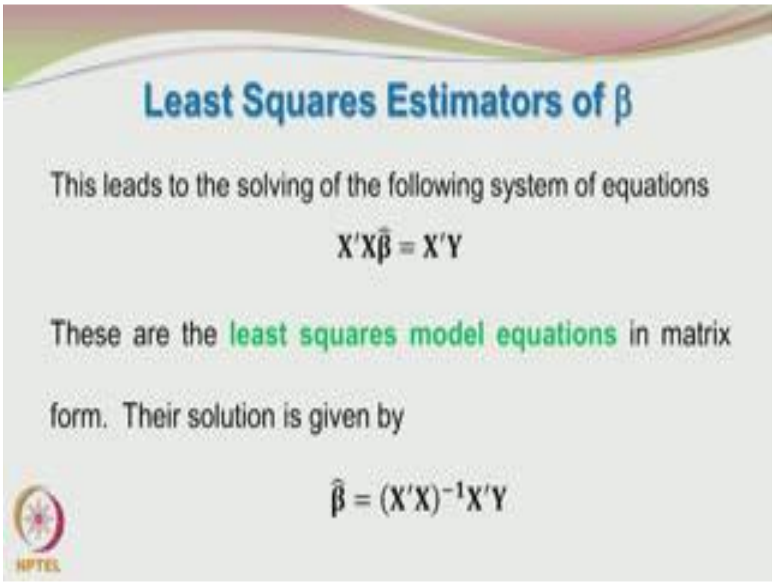
Using these coefficients in the following model will help us to predict the response for various values of X_i $(i=1,2,\ldots k)$

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

So, you have y is equal to X beta plus epsilon. Now, we have to find the beta column vector. We have to estimate the values of beta naught, beta 1, so on to beta k. It looks like a very tough problem, because you are familiar with solving two equations and two unknowns and things like that even three equations in three unknowns becomes quite difficult to solve. But here it looks like, we are going to solve at least 10 variables, it may even be 20, I have seen very elaborate models involving a large number of variables to estimate. So, it looks like a very complex problem, but if you are familiar with linear algebra you will be able to easily appreciate how quickly we are able to solve even large systems of equations right.

So, we do not know the true response because we do not know the true value of beta. So, what we instead do is try to estimate the response or try to estimate the parameters of this model, so that we can predict the response. So, we define our regression model as Y hat is equal to beta naught hat plus beta 1 hat X 1 plus beta 2 hat X 2 plus so on to beta hat k X k. So, this is very interesting. Now, we have put this hat on each of the parameters and also the response; this hat indicates that it’s a predicted response and not the experimental response. Again beta naught hat and beta 1 hat, etcetera refers to estimated parameters and not the true parameters. So, please be clear about the notation.

(Refer Slide Time: 24:17)



Least Squares Estimators of β

This leads to the solving of the following system of equations

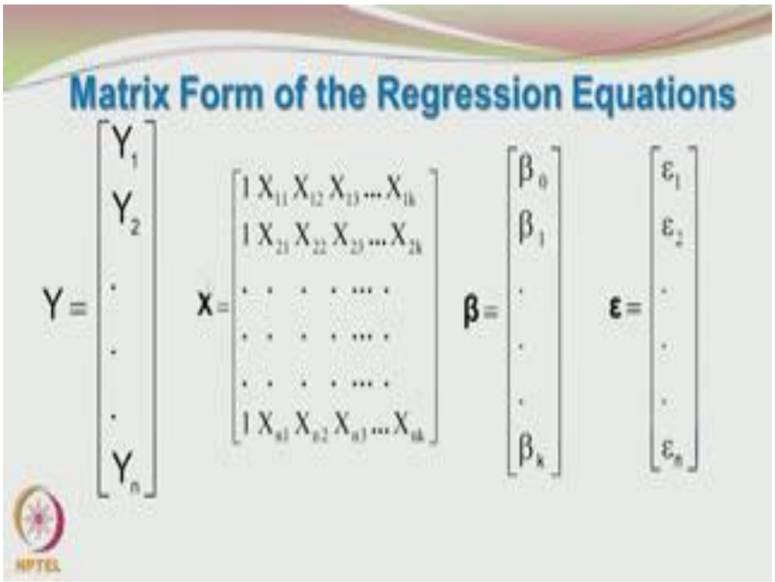
$$X'X\hat{\beta} = X'Y$$

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

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

So, what do we actually do? We first construct the X prime X matrix, where X prime is the transpose of the X matrix. We saw the X matrix to be defined as shown in the slide, and the transpose of the X matrix is represented by X prime, and the X prime or the transpose of X simply involves interchanging the rows and columns of the X matrix, so to get X prime, interchange the rows and columns of the X matrix.

(Refer Slide Time: 24:58)

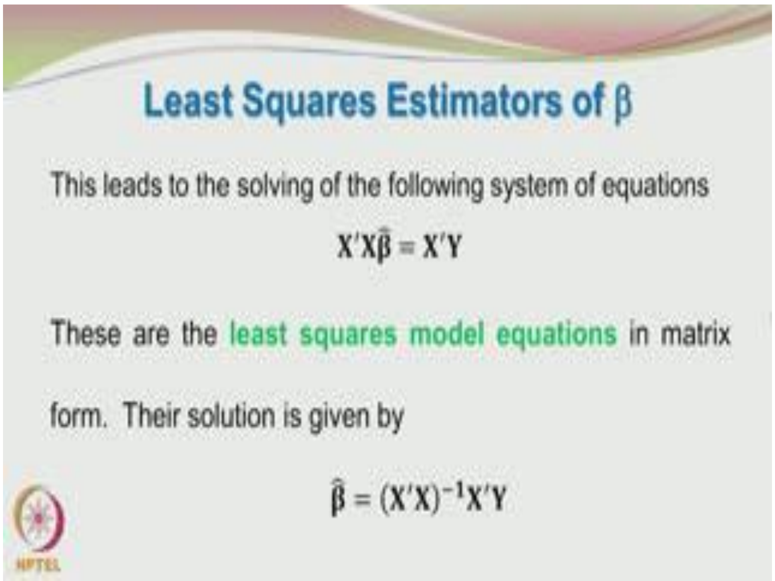


Matrix Form of the Regression Equations

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

So, here if I put X prime, you will have 1 X 11 X 12 so on to X 1 k as the first column okay. So, I think you know transposes pretty well. So, you multiply X prime on X first.

(Refer Slide Time: 25:15)



Least Squares Estimators of β

This leads to the solving of the following system of equations

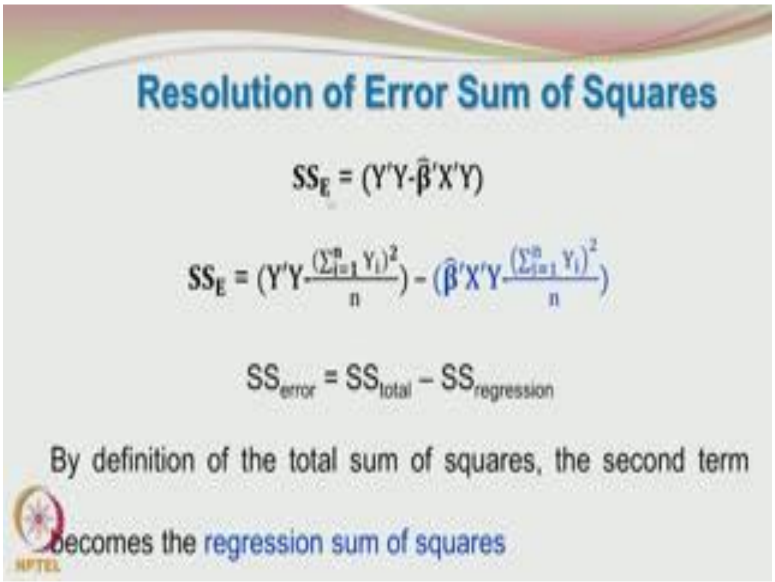
$$X'X\hat{\beta} = X'Y$$

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

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

And then this is pre-multiplying the column vector of the unknown parameters which required to be estimated. Then on the right hand side we have X prime Y, where X prime is again the transpose of the X matrix and Y is the matrix of responses. So, when you do that you are estimating beta hat to be X prime X inverse X prime Y okay. So, very quickly you should be able to get the estimated parameters. Doing the inversion and multiplying the matrices is a very simple task in many software; MATLAB you might be familiar or Scilab you may be using, where all these matrix manipulations are done very quickly and in a very simple manner.

(Refer Slide Time: 26:22)



Resolution of Error Sum of Squares

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

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

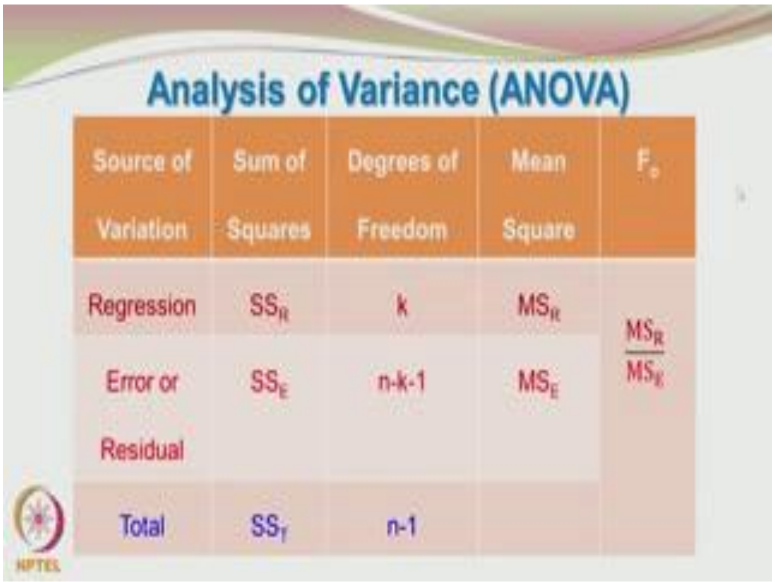
$$SS_{error} = SS_{total} - SS_{regression}$$

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

Okay now, we are looking at something, which is apparently very complicated; it's in fact quite straight forward, you do not want have to worry too much about it. We have the sum of squares of the error. The error component is very essential to see if your regression parameters are significant or not whether the model parameters you have estimated, how many of them are indeed having an impact on the process response and how many of them are dud parameters which are of little value. So, to do that we have to find the sum of squares of the errors; you might have already come across the sum of squares of the errors in earlier analysis of variance tables.

So, how do we find that we subtract beta hat prime X prime Y from Y prime Y. This can also be written as shown here. And the first term in the bracket refers to total sum of squares, and the second term in the brackets is the regression sum of squares. Ideally you would like to have the total sum of squares to be equal to the regression sum of squares, but it never happens the total sum of squares often exceeds our regression sum of squares and more the difference between the two; more is the contribution from the random error component. So, we have the total sum of squares minus the regression sum of squares giving the error sum of squares.

(Refer Slide Time: 27:38)



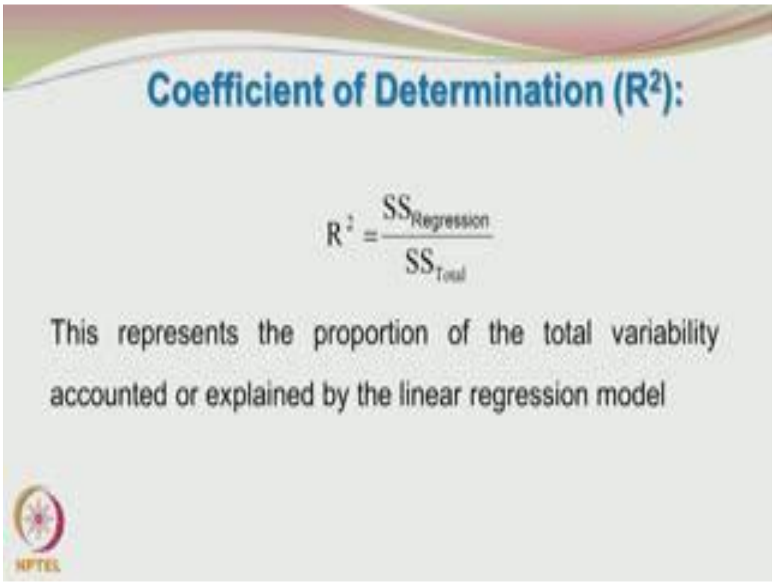
Analysis of Variance (ANOVA)

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

So, again we come to the very familiar by now analysis of variance table, where we have the source of variation in the first column, sum of squares in the second column, degrees of freedom in the third column, mean square in the fourth, and the F value in the fifth. So, we compare the regression sum of squares with the error sum of squares and not immediately, we first have to find the mean squares. So, we scaled the regression sum of squares with the degrees of freedom with the regression sum of squares and so we have SS R by k, which will give you the mean square regression. Similarly, sum of squares of error divided by n minus k minus 1, which is the degrees of freedom associated with the error component and that would give you the mean square error. The ratio of MS R to MS E will give you the F value.

I am not really getting into how you calculate the degrees of freedom for each of these contributions, regression and error. It is very interesting I hope you will be sufficiently interested and motivated to find out how, for example, the regression sum of squares has k degrees of freedom, while the error sum of squares has n minus k minus 1 degrees of freedom, the total number of degrees of freedom is simply the addition of these two and it becomes n minus 1; and the addition of sum of squares of regression and sum of squares of error is the total sum of squares.

(Refer Slide Time: 29:09)



Coefficient of Determination (R^2):

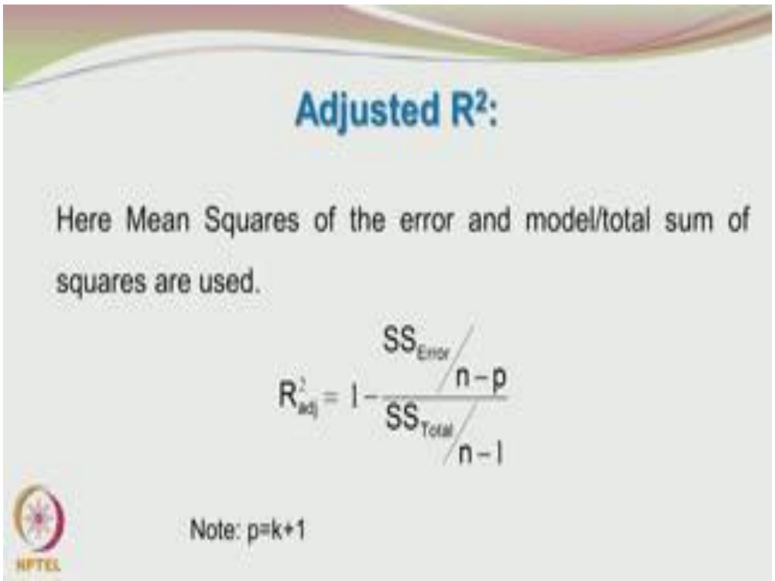
$$R^2 = \frac{SS_{Regression}}{SS_{Total}}$$

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

Many researchers in their publications proudly present their experimental results and save they got an R squared value of 0.99 or 0.991, sometimes they may even say 1, what is meant by the coefficient of determination? R squared is the short form or the nomenclature for the coefficient of determination. So, this represents the proportion of the total variability accounted or explained by the linear regression model. Experiments means variations or variability and you are developing the mathematical model. If all the variability or variations are completely described by the proposed model, it looks like you have solved the problem or cracked the problem, and you are getting an R squared value of 1. You will get an R squared value of 1, the number of variables you want to estimate is equal to the number of experimental runs then you are solving the linear algebraic problem and not an actual statistical regression analysis problem. So, you will always get R squared is equal to 1.

Another way you will get R squared close to one would be keep on adding more and more useless terms to your regression model, you can put beta 11 X 1 squared beta 22 X 2 squared and so on. It may look as if you are introducing some non-linearity or curvature into your model, but as long as the coefficients are still remaining as beta 11 and beta 22, it remains as a regression model. So, you add under another or additional columns corresponding to X 1 squared and X 2 squared into your design matrix X and then manipulate it in the same way to get the regression parameters. So, we are still in the subject of R squared, so don’t aim for unrealistically high value of R squared, just because you want to make it as perfect as possible. In fact, a regression model with the R squared value of 0.84 may be probably more robust, and probably more informative and reliable than a regression model, which is having 15 parameters and then R squared value of 1. Such a model may be in fact, unstable and very sensitive. So, how to quantify which is an acceptable R squared?

(Refer Slide Time: 31:34)



Adjusted R²:

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

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

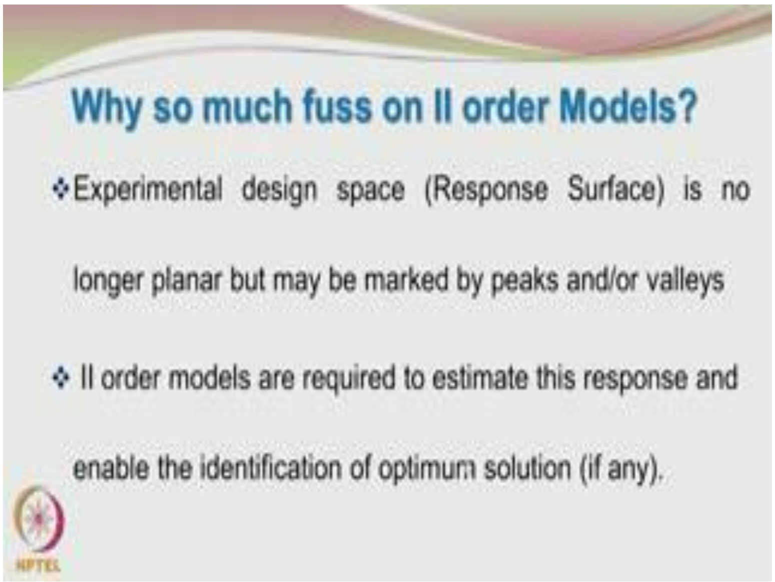
Note: p=k+1

We do something called as adjusted R squared which penalizes the R squared, if you are introducing too many parameters. So, now, we have adjusted R squared defined to be 1 minus sum of square of error by n minus p, by total sum of squares by n minus 1. So, we are getting 1 minus mean square error and divided by mean square total. So, essentially when the number of parameters increases, as I told you earlier p is equal to k plus 1. So, when p increases, this term n minus p will decrease, so the numerator term will start to increase. When the numerator term starts to increase then the R squared value or the adjusted R squared value will start to decrease. So, this tells us that even though R squared value is increasing; the adjusted R squared value may decrease because of this n minus p effect okay. Only when you are adjusted R squared also increases with increasing number of parameters then those parameters are required to improve the models predicting capability.

What I am trying to say here is if the mean square error term it has two components, the sum of squares of error and n minus p. When you add the additional parameter, if the sum of the squares of error comes down considerably then that would have a greater reduction than the increase caused by decrease in n minus p. So, I think if you look at it carefully, you will able to appreciate if SS error decreases at a faster rate than the decrease in n minus p, then there would be a reduction in the numerator overall numerator here, and you are adjusted R squared will in fact increase. So, any way don’t only look at R squared, but also please look at adjusted R squared to see whether the R squared adjusted is increasing or decreasing with the addition of more parameters. If adjusted R squared is decreasing with the addition of more parameters then you better not add those parameters, and leave with the simpler model.

Okay now, when you look at experimental design papers of the literature they keep harping on the second order models, why so much first on the second order models, when you look at the experimental design space it is no longer planar, but it may be marked by peaks and and/or valleys. Peak is mountain kind of thing and a valley is a depression kind of geography in the experimental response space.

(Refer Slide Time: 34:54)

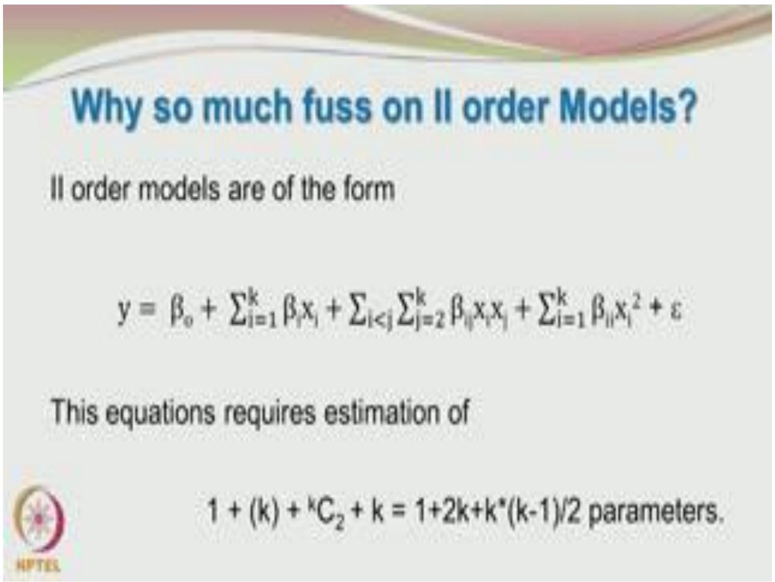


Why so much fuss on II order Models?

- 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).

So, in order to capture these peaks and valleys, second order models are required and the moment you talk about peaks and valleys you are also looking at the best possible optimum condition or the best possible minimum condition. Sometimes the optimum can be maximum, and in some other cases optimum may be the lowest value okay. So, we have to make sure that we find the correct optimum. There can be multiple peaks and multiple valleys, and we should not be satisfied with the sub optimal solution.

(Refer Slide Time: 35:30)



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

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

Okay so, now, how do we represent the second order models? So, this is the form of the second order model. It is not very complicated. Even though it looks to be like that y is equal to beta naught plus sigma is equal to 1 beta i X i and then you have sigma i less than j j equals 2 to k beta i j X i X j plus the sigma equals to 1 to k beta i i X i squared plus epsilon okay. This not as complicated as it looks the first one is the combination of the symbol factors; the second one is the combination of all possible interactions involved between the factors, and then you also have the quadratic term effect X 1 squared and X 2 squared. For example, if you are talking about two factors or two variables, you will have beta naught plus beta 1 X 1 plus beta 2 X 2 plus beta 12 X 1 X 2, and then you will have beta 11 X 1 squared plus beta 22 X 2 squared plus the error component. For a general case, you can find out how many parameters you want to estimate, based on this model structure okay. I think we will stop now and continue in the next class.

Thanks for your attention.