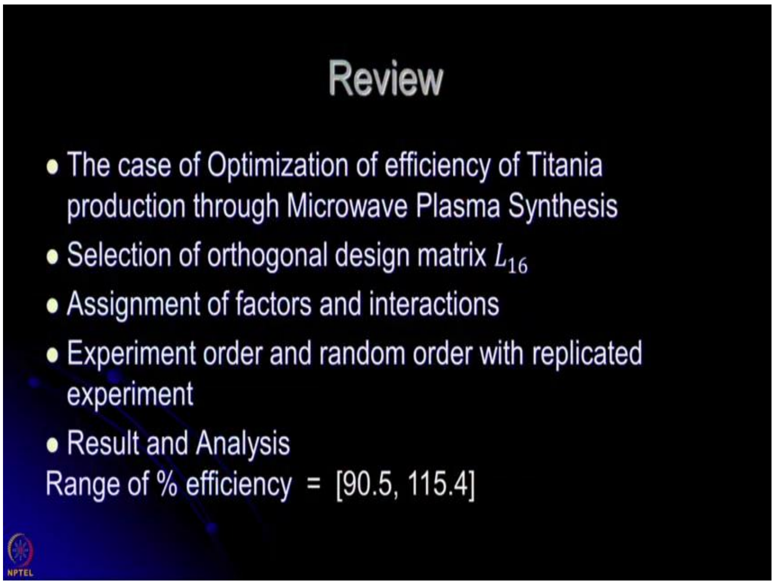
Dealing with Materials Data: Collection, Analysis and Interpretation
Professor M P Gururajan
Professor Hina A Gokhale
Department of Metallurgical Engineering and Materials Science
Indian Institute of Technology, Bombay
Lecture 89
Design of Experiment III

Hello and welcome to the course on Dealing with Materials Data. For past couple of sessions, we are going through a case study on design of experiment and learning the process, how design of experiment methodology is applied. In the past, the first session we talked about the historical aspect of design of experiment, when can we apply the design of experiments, what are the basic three principles of design of experiment.

(Refer Slide Time: 1:05)

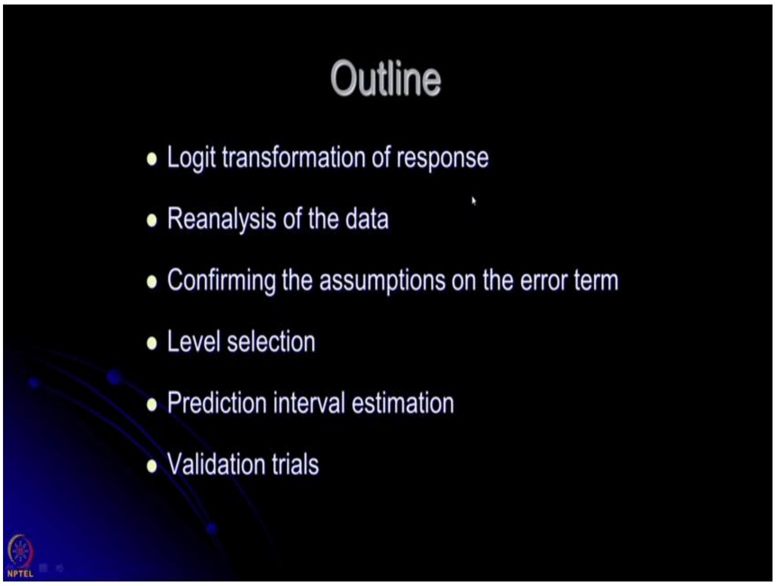


And we introduced the general guidelines, how to go step by step in implementing design of experiments and we introduced the case of optimization of efficiency of Titania production through microwave plasma synthesis. In the second session, we worked out the complete analysis of the design of experiment as to how we did the factor assignment, interaction assignment in a design matrix L16. We also showed how to make a matrix L16.

Then we said that what is the experimental order that is standard order and a random order of with a replicated experiment. We got the result. We did not discuss the analysis of the part of this results in great details, but we quickly went through it to show that the range of efficiency, by that I mean the interval estimation of the predicted efficiency under the selected random selected factors falls in the range of 90 percent to about 115 percent.

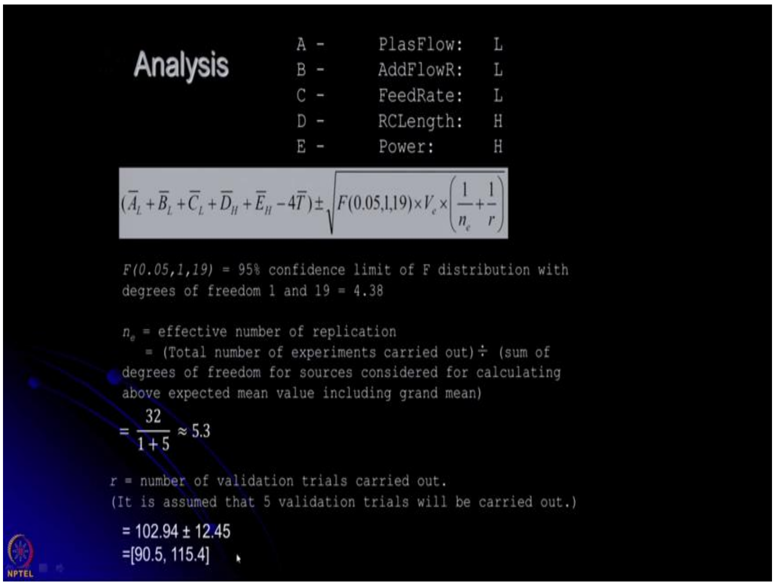
The question we later on asked was that, is everything okay? And one can very easily say that everything is not okay, because efficiency of any system cannot be 115 percent and therefore, something has gone wrong.

(Refer Slide Time: 2:38)



What we wish to do in this session is we would like to introduce in such situation logit transformation of response. Then we will reanalyse the data and this time we will go through both the tables in great details. We will do the confirmation of assumption on error terms. We will show the selection of levels in order to come to the optimized values which would give you the best results. And then we will talk about the prediction interval estimation and the validation trials.

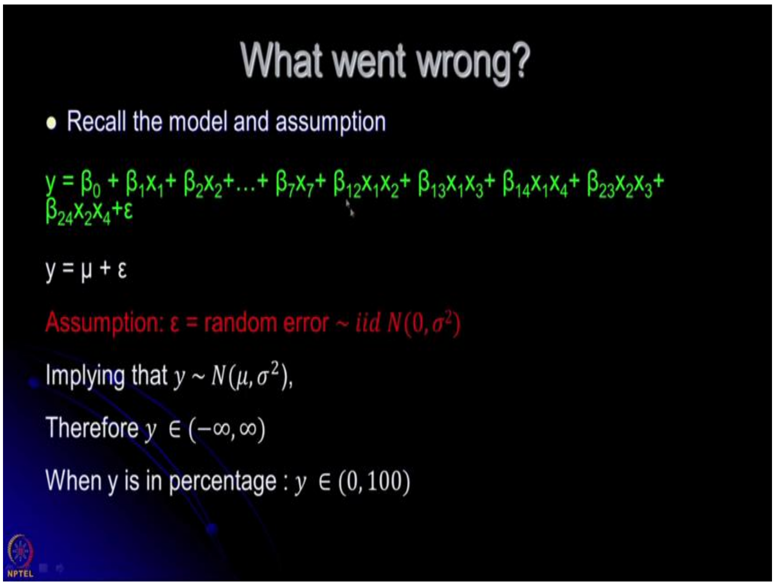
(Refer Slide Time: 3:19)



So, if you recall this was the analysis which I did not explain in great details, which we will do now. But we came up with this interval as a prediction interval, 95 percent prediction interval and for 5 validation trials and we find that there is something wrong in it. So, we asked ourselves a question, is everything okay? And the answer is no, because there cannot be 115 percent efficient system. So now, I would like to emphasize yet one more thing, it is very important that when you do this design of experiment, keep asking your question, yourself a question now and then, is everything okay?

Because what really happens when you conduct the whole experiment, you get so much involved in the experimentation and then finally in the analysis which is all variety of numbers that you deal with that sometimes you lose the focus on main issue of the problem. And therefore, it can very easily happen that we said that I have done everything and this is my prediction interval for 5 validation trials, but does it make sense? This question you have to ask time and again in order to see that you are going in the right path, in the right direction for your analysis. So, here the question is, is everything okay? And the answer is no.

(Refer Slide Time: 4:53)



What went wrong? Well we assumed this model, remember we assumed the model that response is a regression model with the parameters which are the 7 main effects and 5 interaction effects, so there are 13 parameters. I can simplify this model by saying that it is

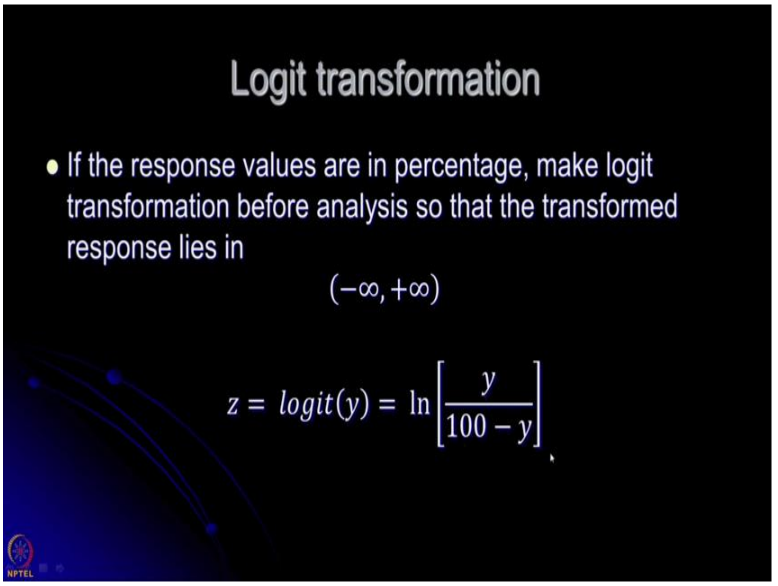
$$y = \mu + \epsilon$$

and our assumption is that epsilon is a random error and it is distributed as normal with mean 0 and unknown variance sigma square.

Which in turn implies, it means that this assumption actually has made an assumption on the response variable y which says that $y \sim N(\mu, \sigma^2)$ and which means that we are assuming that y, the response value is going to lie between minus infinity and infinity, but our response is percentage and percentage lies between 0 and 100.

Therefore, we have violated this assumption itself in taking this model and therefore, something needs to be done about it.

(Refer Slide Time: 6:13)

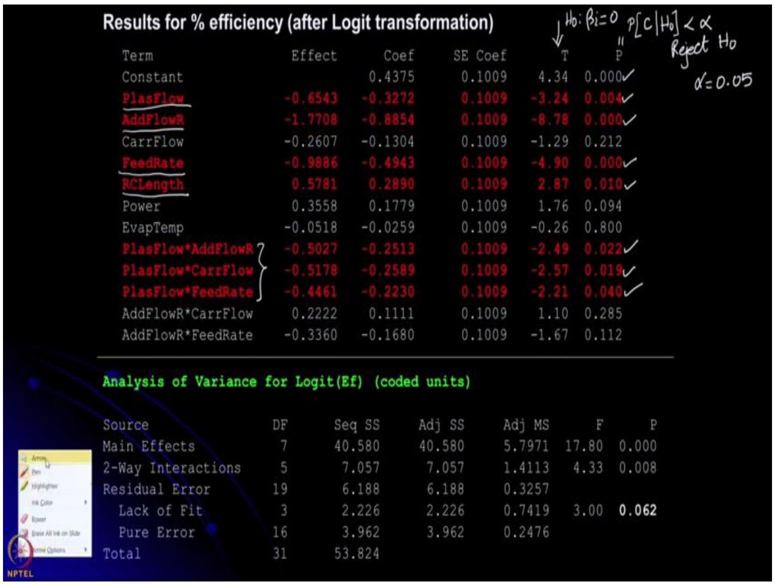


And the answer to such situation in this case is logit transformation. So, if your response values are in percentage, it means that it lies between 0 and 100. It makes sense to make a logic transformation before analysing the data and we should analyse the data with a logit transform response.

$$z = logit(y) = \ln\left[\frac{y}{100 - y}\right]$$

This logit transformed y which is z lies between minus infinity and infinity and therefore, in our model assumption, we replace the response y by the logit transformation of y, z and then we analyse the data.

(Refer Slide Time: 7:10)



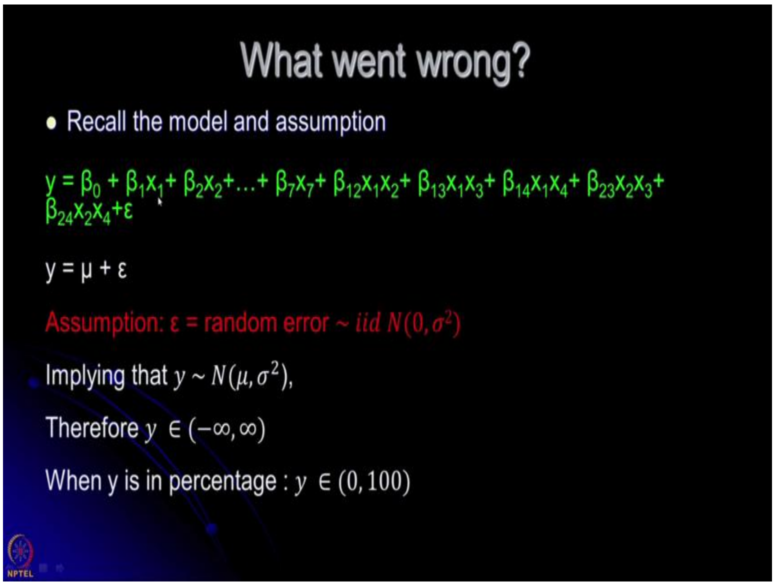
So, now after the logit transformation, these are the 2 tables of analysis, this is a table of all the coefficients and this is the analysis of variance table after logit transformation of efficiency. You see, all the results have changed. The tables have been created in the same way, your response y which you have been working on is now logit transform and therefore, your effects values are different and it has the P values are different.

Now, let us see, I already told you what is the t-distribution. This is the t statistic, this is a t statistic for testing the hypothesis that beta i is equal to 0, this is a t statistic which refers to the null hypothesis that beta i is equal to 0. Beta i refers to the each beta in your equation in your regression model.

And this is a P value which is actually the probability that of critical value when H0 is true and we say that this probability has to be less than alpha. You remember, this is what we had learned. So, if this is less than alpha, we reject H0, that means that probability of critical region is less than alpha, we reject H0.

And therefore, you can see that this constant is only a coefficient, so that is of course accepted, but these are the values which are shown in red which are less than our alpha is pre decided a 0.05, so these are the values which are less than 0.05, so it means that plasma flow rate, additional gas glow rate, feed rate, reaction chamber length and these 3 interactions are significant.

(Refer Slide Time: 10:07)



Because we reject the null hypothesis, it means that for these values betas are not 0, which in turn means that if you, sorry, which in turn means that, which in turn means that these sum of these parameters, the parameters referring to plasma gas flow rate, additional gas flow rate or plasma gas flow rate interaction with additional gas flow rate, these are the betas which are not 0, other all other beta we accept the null hypothesis that these betas are 0.

(Refer Slide Time: 10:30)

And therefore, we find that these are the betas which play the important role, these are your significant factors, all others are not significant. In other words, your model actually, your logit transformed efficiency depends upon the plasma gas flow rate, additional gas flow rate, feed rate, reaction chamber length and carrier gas flow rate only through the interaction. So, it is plasma gas flow rate interaction with the additional gas flow rate, plasma gas flow rate interaction with carrier gas flow rate and plasma gas flow rate interaction with the feed rate, these are the only important factors, all other white colour factors are not important.

Let us look at the analysis of variance. In analysis of variance, these effects are confirmed by the fact that again you take alpha is equal to 0.05. Then these are the F statistic which also says that the variation due to main effect is more is significant than the variation due to error. Here, what we are comparing? We are comparing two variances.

This is also an estimate of a variance sigma square and this is also a estimated value of a sigma square, this is estimated value of sigma square that is the main effect adjusted the mean square is estimate of sigma square provided your beta 0 which says that, main effects are important is 0, if that is the case then these 2 are exactly the same.

So, if we have two estimate of variances, the estimate due to the main factors and the estimate due to the error. Error is always an estimator of sigma square. So, if these two are same, it means that the ratio F is not large enough then you can say that main effects are not important, but here we find that main effects are important.

The column effect and the row effect, if you recall in analysis of variance, this effect is important because the P value is less than alpha, so the same criteria implies and similarly these effects are also important. So, here we are, though it looks small but these are significant enough. Now, what is lack of feed?

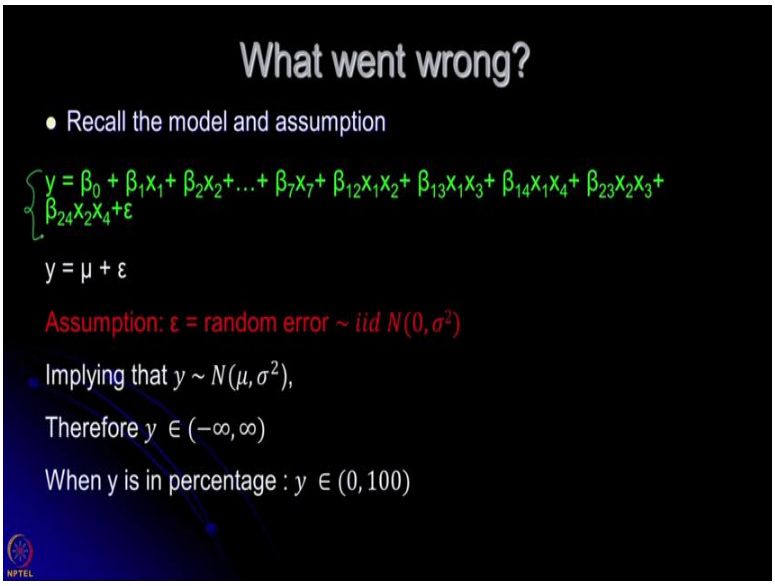
Lack of feed is calculated by using the three left out columns. So, this is calculated that error due to the three columns which we have not accounted and the pure error, this is we are dealing with only this, this is the ratio of 2, this is also an estimate, this, now remember, maybe I should use a different colour, let us use a colour green.

So, this particular factors, these together, these are your lack of feed and this these two are your lack of fit, what does this say? This part actually refers to the case of replica, this refers to the replicated replica. We have done 2 sets of 16, 16 experiments to find out if there is a error due to this experimentation itself.

And that error, here it shows that this error is not significant because this is greater than alpha which is 0.05, so we find that, this is, it is in the it is not in the critical region, and therefore, we cannot, not rejected the hypothesis. The hypothesis is that there is lack of, there is no lack of fit, both of them are exactly the same, our hypothesis is there is a lack of fit.

So, hypothesis is that there is a lack of fit and it is rejected, sorry, there is no lack of fit. And, therefore the hypothesis, we cannot reject it. So, there is no lack of fit. It means that your model adequately defines your complete data.

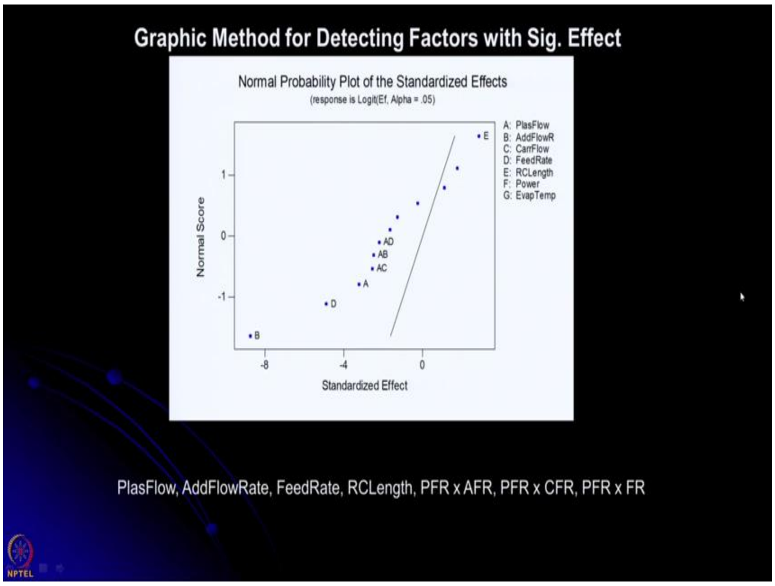
(Refer Slide Time: 15:56)



It means that this model this model this model is completely able to explain the logit transformed response in terms of the selected parameters which are given here.

So, this says that there is no lack of fit is your null hypothesis which is not rejected, so there is no lack of fit. The hypothesis, the result is that it completely, the model of logit transformed efficiency completely is explained by these selected variables.

(Refer Slide Time: 16:42)



We go to the next. These are the P, I mean this is coming up. There are graphical methods also available to detect the factors with significant effect.

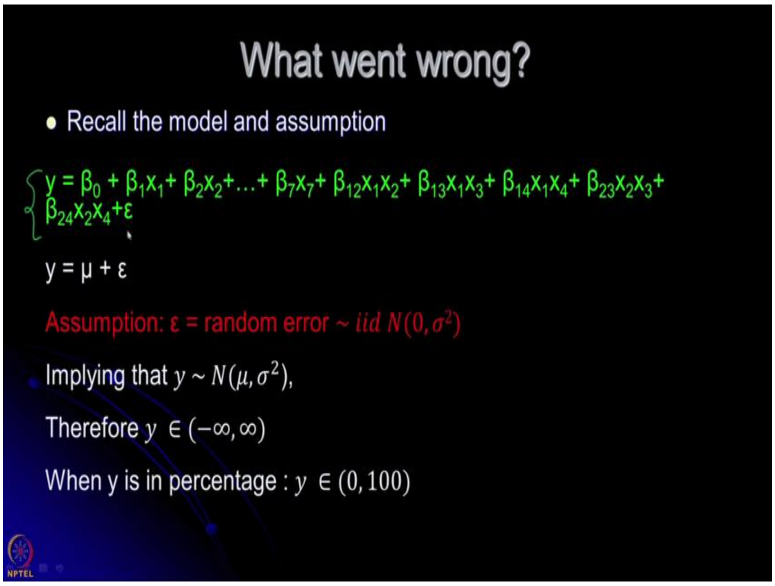
In this, the standardised effect, you remember this is your effect and you can divide it by the appropriate value of standard error of coefficient then you get a standardised effect and standardised effect is plotted against normal score on a normal probability plot, it is a normal probability plot. If the values follow the correct normal path, it should follow on this line, but when it deviates, it actually tells you that these are the values which are, which have the beta value not equal to 0.

These are significantly effecting the your result, if your result was only a random error. See, this also shows that this line would come if the effect is only due to epsilon. Let us try to understand this, when you say that beta 1, beta 2, beta n, etcetera is 0, it means that your variation that you see in logit transformed efficiency is purely due to error, there is no effect of mu.

So, when that happens we say that in all the results should fall in this line, all the results should fall on this line. When they go away from the line, it means that your assumption that percentage efficiency is completely a random process, it is not true, there is a systematic error in it and those systematic changes are because of these factors, these are because of these factors which are farthest away from the line which shows that it is only a random effect.

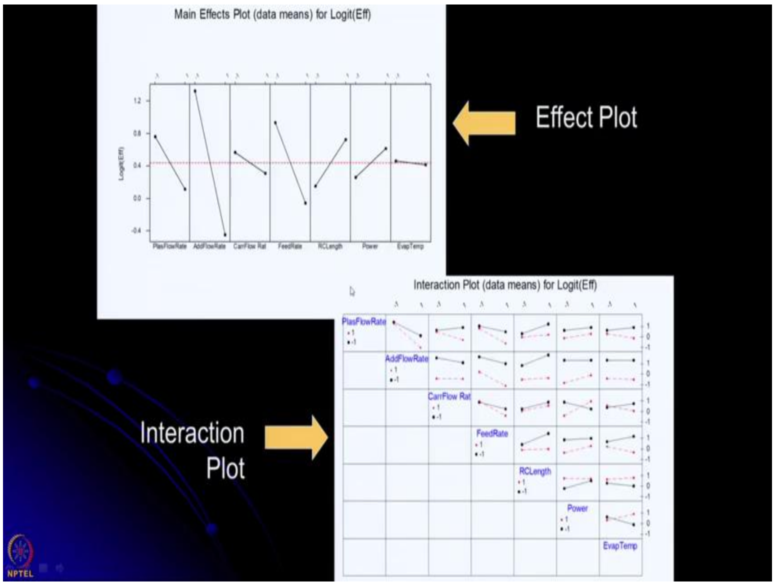
So, please understand in this whole process we are trying to separate out the effect due to the systematic change in the factors versus the random errors which we say that it is may have caused due to human error, machine error, etc.

(Refer Slide Time: 19:19)



So that random error and systematic errors are separated in this particular model. This is the systematic error or systematic changes and this is a random error, so what it really shows is that if the error was only random, it would fall all the data points would fall on this line, but as they are going farther away it shows that there is these factors which have been very systematically changed from level minimum to level maximum has an effect on your percentage or logit transferred percentage efficiency.

(Refer Slide Time: 19:55)



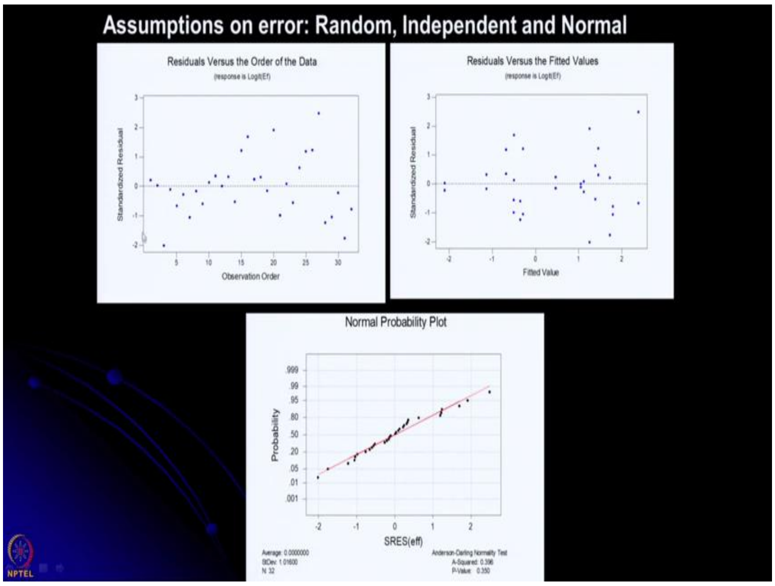
This can also be found by plotting an effect plot and a interaction plot, these are very simple plots compared to the previous one. The plasma flow rate is varied from minus1 and 1 and it is the logit of efficiency is plotted on y axis and the slope of the line shows that how severe is the effect, how large is the effect. So for example, carrier gas flow rate or you look at evaporating temperature, it hardly has an effect, it is almost horizontal.

While these are some effect, but you can see that these are the ones which have a larger effect. Carrier gas flow rate is also almost flat like a power, so that is not shown in the, when you look at the interaction plot, it is plotting the same thing for a the two effects together. So, on one axis is additional gas flow rate, on in the other axis for exam is plasma gas flow rate and the effects are plotted.

And whenever these lines cross each other, it means that there is an interaction effect, if the lines are almost parallel then we say that there is no interaction effect. So, you can see that these are almost parallel, but these has an interaction effect, this has a clear interaction effect, as you can see.

So, you can also see graphically. Generally, most of the software package and I believe even in the hour you can have these plots very easily and understand it. Now, another thing we have to make sure that we are not deviating from our assumption, the assumption on error. So, error is the logit transferred that is Z value and estimated Z value, the difference between the two is called an error. So, Zi minus Z hat i is called an error.

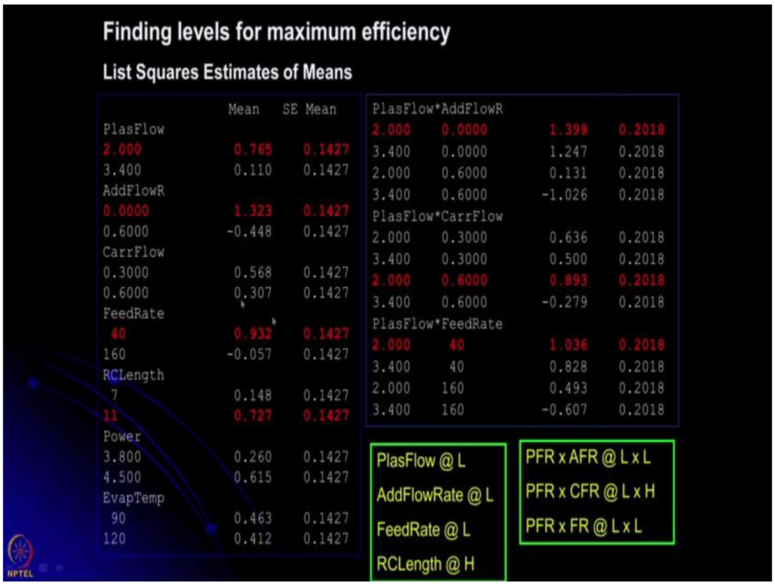
(Refer Slide Time: 22:01)



And there in this error if this is plotted, the observation order vs standardised residual, if you remember we did this exercise during regression analysis. If this is completely randomly plotted, then we are very happy that it is indeedly a random error. If we plot the fitted value against the standardised residual order, residual value, standardised residuals and if it does not show any pattern, it also means that there is no pattern in it, it is a randomised order.

This is a normal probability plot so we have plotted against normal probability the values of the logit, the error caused, the standardised error caused by logit transformed estimated values, efficiency values and it shows that this error, standard errors is also falling on the straight line. And therefore, it is our assumption of normality is also correct.

(Refer Slide Time: 23:09)



Finding the level for maximum efficiency, we find, you can do it this way, this is called the estimate tables where you take plasma flow rate at different values and you find the means value of its logit transformed efficiency and its standard error of mean. Remember, standard or error of mean is standard deviation divided by square root n, n is the number of experiment, you are conducting, you are looking at only 1 level, so you have conducted only 8 times 2, 16 experiments here, so accordingly this has been worked out.

So, this is a mean over the values when plasma flow rate first kept at minimum, this is the same when it was kept at maximum. You want to have a maximum efficiency, so you want to have the maximum mean. So, you know that plasma flow rate will attain maximum will, at when plasma flow rate is at minimum level it will give you maximum efficiency.

Similar, additional gas flow rate should also be at the minimum level to give you maximum efficiency. We can do the same thing with respect to the interactions. So, interactions will have four possibilities and we find that when plasma flow rate is minimum and additional gas flow rate is minimum, it gives you the maximum logit transformed efficiency, while plasma flow rate at minimum, but carrier gas flow rate at a maximum gives you the maximum logit transformed efficiency.

And similarly, minimum value, levels at plasma flow rate and feed rate gives you the maximum thing. So, our selected variable levels, selected factor levels are, we have selected 4 factors, plasma flow rate at low level, additional gas flow rate at low level, feed rate at low level and reaction chamber length at high level. The interactions are important, plasma flow rate, additional flow rate, both at low level which is good.

The carrier gas flow rate appears only as an interaction and it should be kept at a high level, while the feed rate and plasma flow rate both are at low level and their interactions are also important. So, this is what you derive looking at this.

(Refer Slide Time: 25:52)


Results for % efficiency (after Logit transformation)

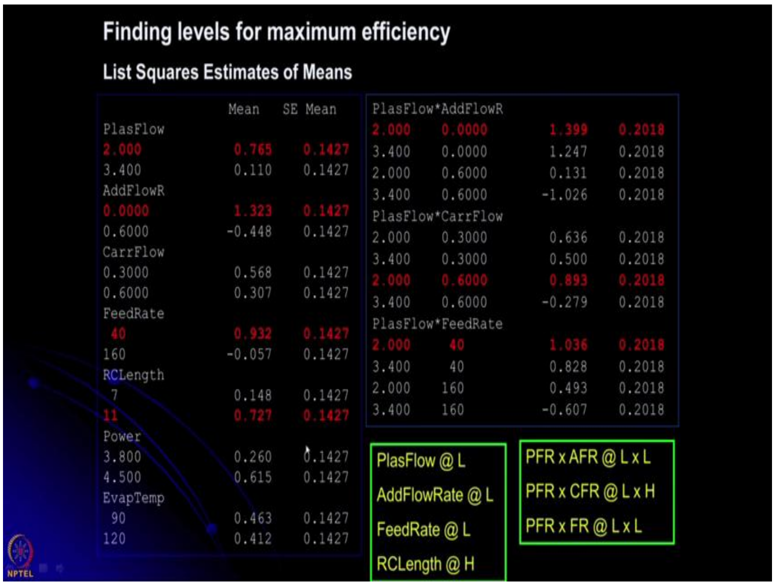
Term	Effect	Coef	SE Coef	T	P
Constant		0.4375	0.1009	4.34	0.000
PlasFlow	-0.6543	-0.3272	0.1009	-3.24	0.004
AddFlowR	-1.7708	-0.8854	0.1009	-8.78	0.000
CarrFlow	-0.2607	-0.1304	0.1009	-1.29	0.212
FeedRate	-0.9886	-0.4943	0.1009	-4.90	0.000
RCLength	0.5781	0.2890	0.1009	2.87	0.010
Power	0.3558	0.1779	0.1009	1.76	0.094
EvapTemp	-0.0518	-0.0259	0.1009	-0.26	0.800
PlasFlow*AddFlowR	-0.5027	-0.2513	0.1009	-2.49	0.022
PlasFlow*CarrFlow	-0.5178	-0.2589	0.1009	-2.57	0.019
PlasFlow*FeedRate	-0.4461	-0.2230	0.1009	-2.21	0.040
AddFlowR*CarrFlow	0.2222	0.1111	0.1009	1.10	0.285
AddFlowR*FeedRate	-0.3360	-0.1680	0.1009	-1.67	0.112

Analysis of Variance for Logit(Ef) (coded units)

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Main Effects	7	40.580	40.580	5.7971	17.80	0.000
2-Way Interactions	5	7.057	7.057	1.4113	4.33	0.008
Residual Error	19	6.188	6.188	0.3257		
Lack of Fit	3	2.226	2.226	0.7419	3.00	0.062
Pure Error	16	3.962	3.962	0.2476		
Total	31	53.824				

You remember, last time we had done it by looking at this tables that if this is minus it means that the values are minimum of plasma flow rate will give you maximum of logit efficiency.

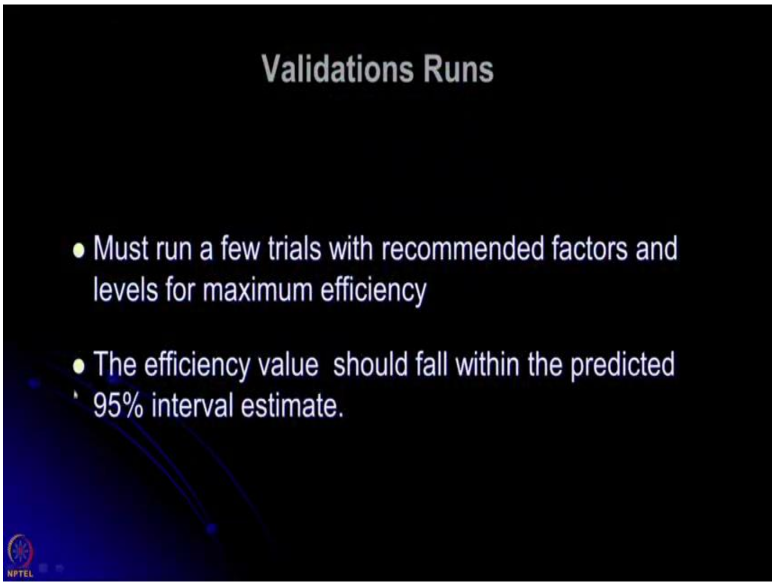
(Refer Slide Time: 26:07)



Here, instead we have done the actual calculations and found it out. So, this way also you can do it, this way it becomes very clear, particularly with respect to the interaction.

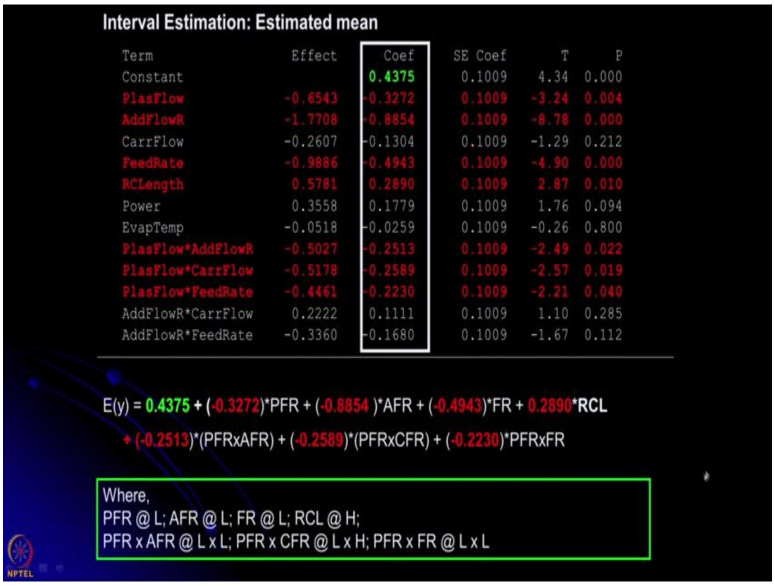
And then the comes that this results that we have got, we must validate, it means we must carry out a few experiments by keeping plasma flow rate at low level, additional gas flow rate and feed rate also at low levels, but RC length as the highest level and of course carrier gas flow rate at high level. Conduct a few experiment and see that it gives you the maximum possible efficiency, logit transformed efficiency. Logit transformation is also increasing function and therefore, if the logit transformation transformed efficiency is high, the efficiency is also high.

(Refer Slide Time: 27:13)



So, we must run a few trial at recommended factors and level for maximum efficiency and we should see that the efficiency value should fall in the predicted 95 percent interval estimate of the maximum efficiency. So, let us do the interval estimation.

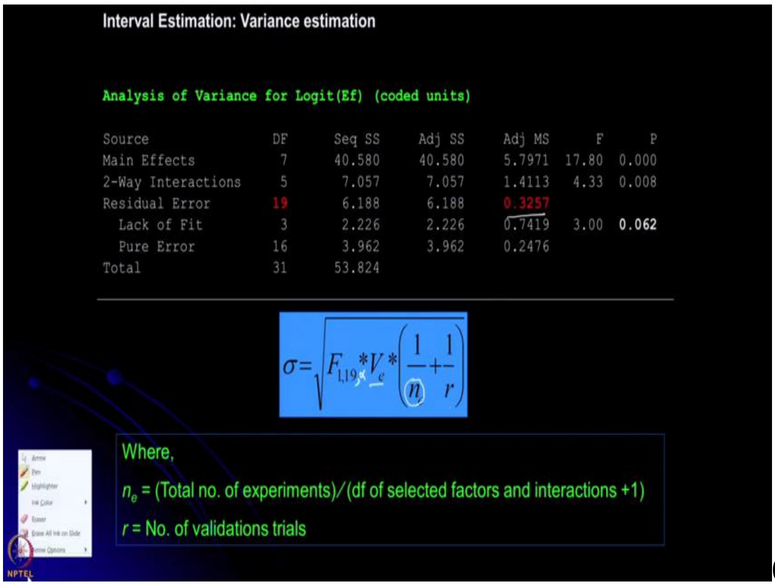
(Refer Slide Time: 27:25)



So, this is the model we have estimated. Expected value of y, it is actually logit transformed y is this, this is the coefficient and these are the red values which are shown as having the coefficient, they are written here and these are the actual parameters which we are putting here. And if you take, you remember that this is a coded one, so therefore, plasma flow rate at minimum is going to be minus1, this is going to be minus 1. Reaction chamber length is highest, so it is going to be plus 1. This is not the values that you have calculated here.

Please remember these are the coded values. I think we must write it down here, these are all four coded values, very important, this is for coded values. So, this is going to take minus 1 value, this will be minus 1 because we have to keep at minimum and this will be plus1 value, this will be minus 1 times minus 1, this will be also minus 1 times plus1 and this will be also minus 1 times minus 1. This is how you have to calculate. So, this is what I have shown here, this is what you have to keep these values.

(Refer Slide Time: 29:15)

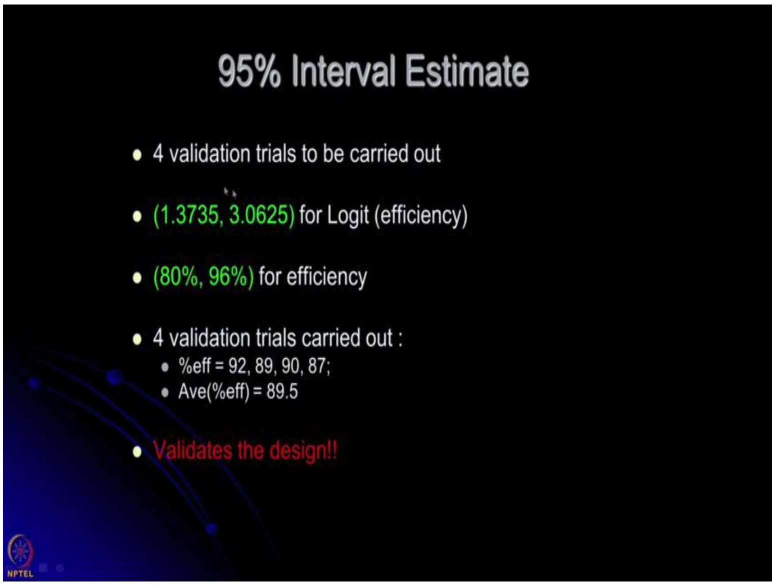


0

And then these are you have to do it in a coded values, so these are the values that you will take and then the, how do you find the range? The range is found in this manner. This is your F value and adjusted mean square is your F value is F value at 19, here something I have not written I must write it down here, this is F, it does not show well, let us change the ink colour, it shows F at alpha.

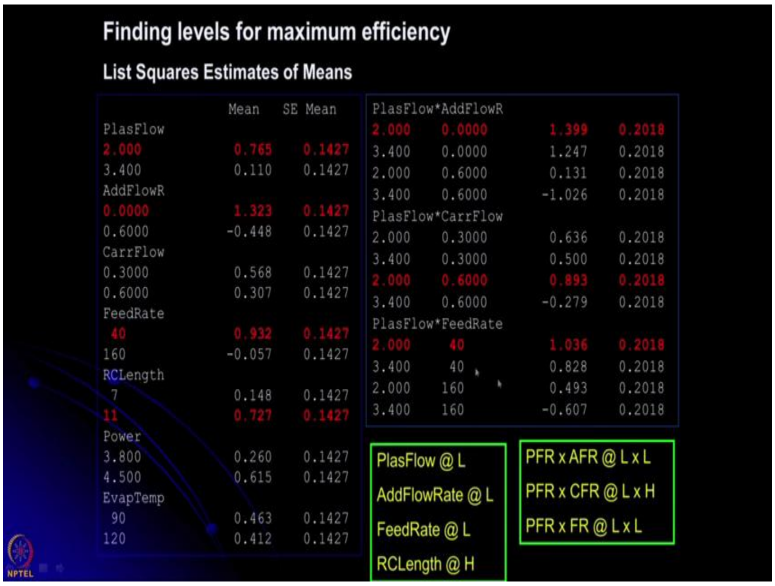
Please remember we have to take it at alpha level. And therefore, you take F at alpha 95 percent that is 0.05, this is the error variance which is shown here and then this is the, so this is going to be the effective number of replica which is total number of experiments divided by degrees of freedom of selected factors plus1, r is the number of validation trials.

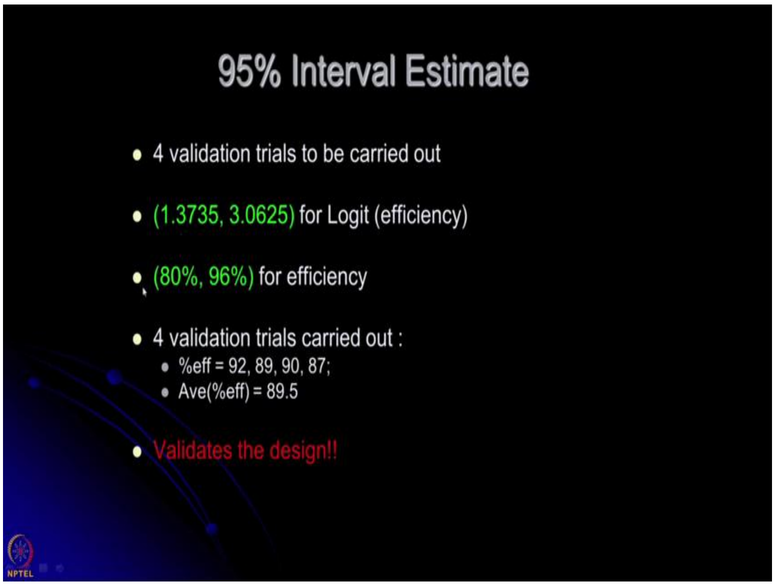
(Refer Slide Time: 30:44)



So, accordingly if you calculate, we were to calculate 4 validation trials. So, the logit efficiency interval is this, interval estimate for efficiency for 4 validation trial is 80 percent to 96 percent, these are the 4 validation trial results which is 92 percent, 89 percent, 90 percent and 87 percent and the average of that is 89.5 percent which is very much within the interval and therefore, we can say that our results are validated.

(Refer Slide Time: 31:35)





In other words, this model this model is the correct model in which you select only, here is the model, only logit efficiency described in the form of plasma flow rate at the lowest level, additional gas flow rate, this is the, plasma flow rate kept at low, additional gas flow rate kept at low level, feed rate kept at low level, reaction chamber length kept at high level with carrier gas flow rate at high level would give you the highest efficiency of the system and we validated it by going through the, finding out the 95 percent interval of predicted value of 4 validation trials and the design is validated. Thank you.