

Applied Multivariate Statistical Modeling
Prof. J. Maiti
Department of Industrial Engineering and Management
Indian Institute of Technology, Kharagpur

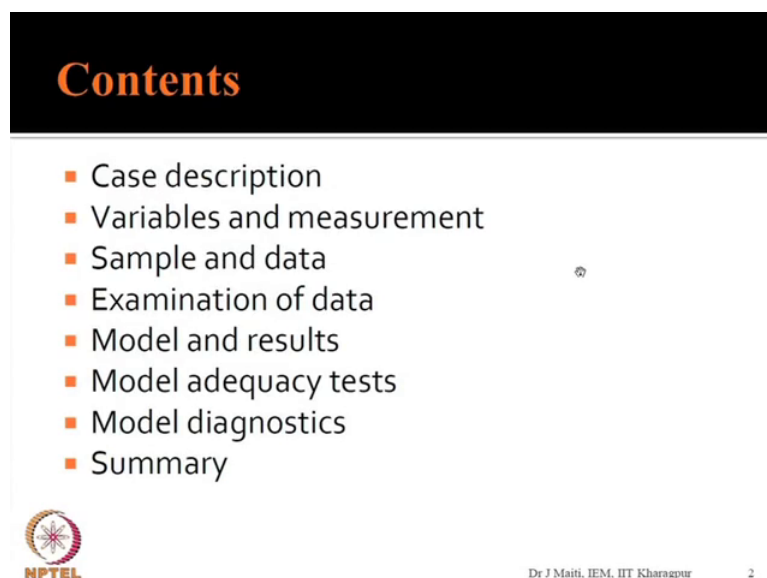
Lecture - 26
MLR - Case Study

(Refer Slide Time: 00:24)



Good morning, we will discuss today one case study on multiple linear regression. Let us see this case study first.

(Refer Slide Time: 00:29)




Today discussion will be on the case description, what are the variables and how we measure the variables, the sample collected, examination of the sample data, model and results, model adequacy tests, model diagnostics, fine finally followed by summary.

(Refer Slide Time: 00:53)

Case description

- The study is conducted in a worm gear manufacturing plant of India.
- Manufacturing process comprises
 - heating of ingots in crucible furnace
 - casting of molten metal in centrifugal casting machine
 - gear cutting in a hobbing machine
- Purpose: to model relationship of worm wheel quality with hobbing process variables

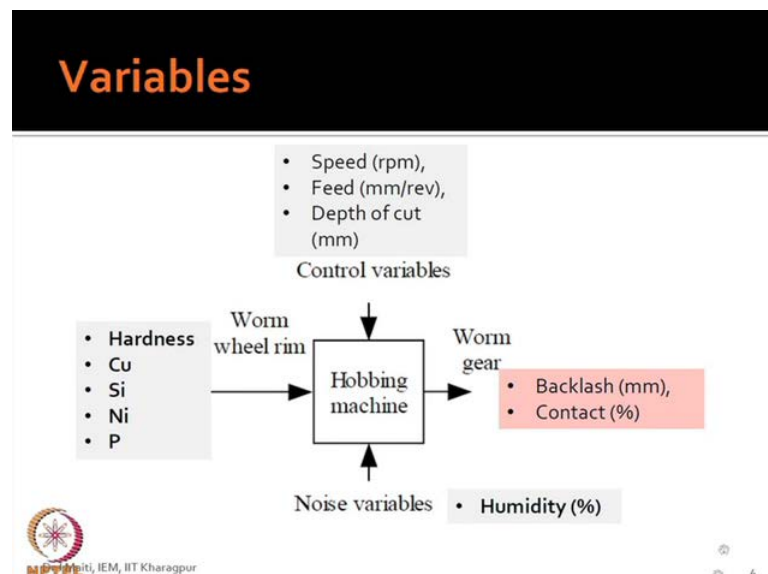


Dr J Maiti, IEM, IIT Kharagpur

3

This study was conducted in a worm gear manufacturing plant of India. Manufacturing process comprises, heating of ingots in crucible furnace, casting of molten metal in centrifugal casting machine, gear cutting in a hobbing machine. The purpose of this case study is to model the relationships of worm wheel quality with hobbing process variables.

(Refer Slide Time: 01:30)

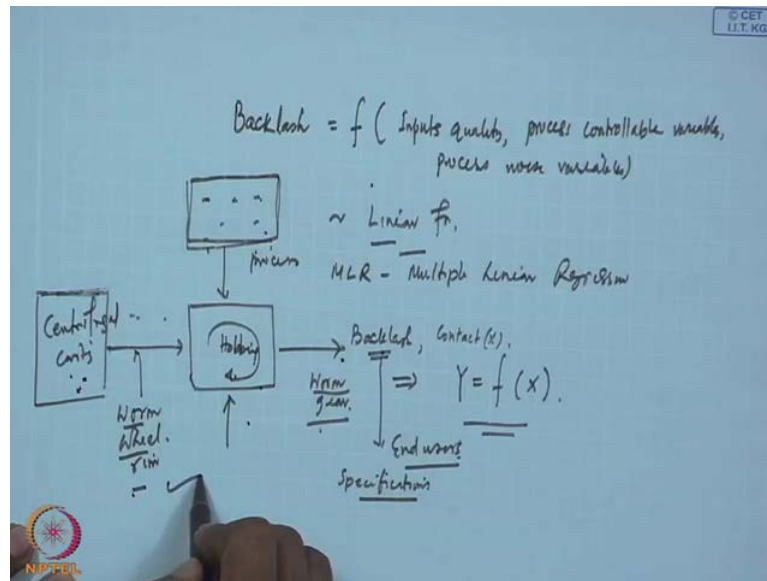


Now, see the schematic diagram of the entire process. Here, what we are saying this is the hobbing machine, that worm wheel rims are the input to this machine. This machine basically makes the or cuts the heat gear, heat and the worm gear from the point of view backlash and contact are the two important quantity variables and the worm wheel which is input to the hobbing machine. Here, the input worm wheel quality is measured through hardness copper percentage silicon percentage nickel percentage and phosphorous percentage.

So, what I mean to say that worm wheel as an input to this hobbing machine has certain quality features. These are known as input quality and this input quality variables are measured through hardness copper, silicon, nickel, phosphorous with appropriate units of measurement. Now, this hobbing machine is basically as I told you that this will cut the gear teeth. So, it has certain process parameters which are given here, speed in r p m, feed millimetre per revolution, depth of cut in millimetre. These are the variables which can be controlled by the operators.

Now, there are noise variables for example, humidity percentage and also ambient temperature that could be noise variable. So, here our sole interest is to use multiple linear regression to show that multiple linear regression is applicable, and it will give us some benefit in analyzing the relationship between the worm gear quality variable, like backlash, like contact percentage with the controllable process variables like speed feed depth of cut and noise variable, like humidity and the worm wheel rim quality variables which worm wheel is the input to this machine.

(Refer Slide Time: 04:12)



So, essentially what do you want to do? We want to say that backlash this can be function of that input quality can be process controllable variables and process noise variables process. Noise variables and this function we are saying it is a linear function linear function. So, we definitely test the linearity as well as other assumptions related to MLR, that is multiple linear regression.

And we want to say that whether we will be able to transform these hobbing process, hobbing process where the quality variable we are at present considering only backlash. Although, the contact is another quality variables which output from the hobbing process controllable and uncontrollable variables are there. So, can this process be represented in terms of an equation? Y is function of X , so that is our sole purpose.

So, in order to do, so let us see that the characteristics of the variables which is governing the total hobbing process and their specifications what is given by the customers. So, hobbing process is customer to the preceding process. Here, it is basically centrifugal casting, centrifugal casting is one process which basically produces that worm wheel, worm wheel rim. And depending on the process conditions and input to this centrifugal casting process, the quality of worm wheel will be determined. So, here hobbing is customer to casting process, ok?


Now, the worm gear is formed, this worm gear is sold to different customers end users. Now, the end users they give certain specification related to backlash, related to contact

percentage. These two are the quality characteristics of the worm gear from end users point of view. Similarly, when the hobbing is accepting the worm wheel rim from the centrifugal casting, so there is also hobbing is also given certain specification to the quality variables related to worm wheel rim which is produced by the casting process in addition. What is required, the operator here you will control the process variable within a certain range.

So, what do I mean? I mean that the output of worm wheel, these are known as quality variables. Here, these variables are certain specifications which are determined by the basically end users, which are given by the end users not in the precise that numerical form. But end users customer requirement converted to specification here. Now similarly, in order to achieve or produce the worm gear within the specifications, you are also specifying certain range for the this variables process, variables and hobbing as a customer from the centrifugal casting process point of view. Hobbing is also specifying certain ranges for variables related to worm wheel quality, ok?

(Refer Slide Time: 09:17)

| Variables: specifications | |
|---------------------------|--|
| Input: Worm wheel rim | • Hardness (>90); Cu (87.22-88.72), Si (9.75-10.75), Ni (1.25-1.75), P (<0.03) |
| Process | • Speed (5-10 rpm), feed (1.1-1.2 mm/rev), depth of cut (0.25-2mm) |
| Noise | • Humidity (65-95 %) |
| Output quality | • Backlash (1.17-1.68 mm), Contact (30-40 %) |

 Dr. J Maiti, IEM, IIT Kharagpur 5

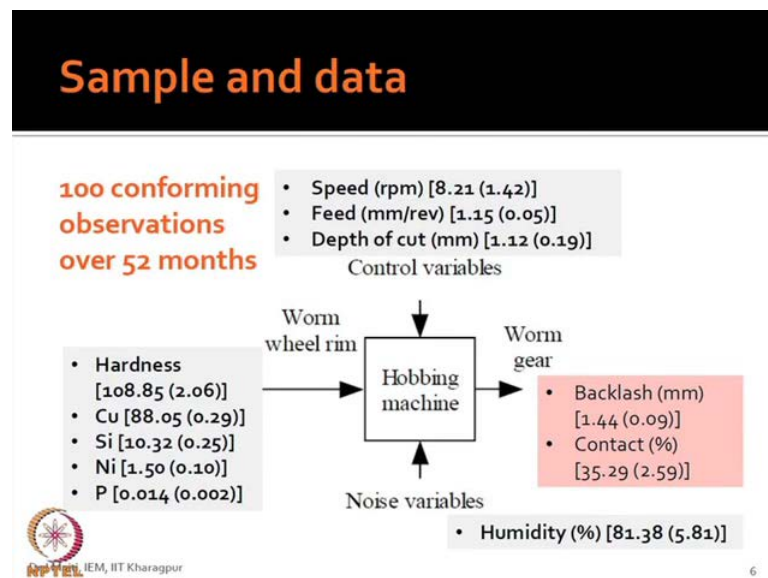
So, this is what is given here, that hardness. Suppose, input is worm wheel rim hardness measured hardness number, so that should be greater than 90. Copper is a chemical composition that percentage should be 87.22 to 88.72, in this range must be there silicon. Similarly, 9.7 to 10.75 nickel phosphorous, so would be less than 0.03 percent, all are percent figure. So, this is what hobbing section requires from casting centrifugal casting

section in terms of worm wheel rim quality and output quality. If you see here that backlash 1.17 to 1.68 millimetre and contact 30 to 40 percent contact, this is given by the end users, that mean the who are purchasing the worm gear.

So, in between what happen the process controllable variable by speed feed depth of cut, these are the specification 5 to 10 r p m, 1.1 to 1.2 millimetre per revolution. 0.25 to 2 millimetre, this is what is the process variable range determined by the engineers there and the plant is operating under humidity conditions 65 to 95 percentage that is throughout the year. Now, this humidity is such a thing it cannot be controlled, whether it is weather condition. So, this is that is why this is termed as noise.

So, you have input quality, you have process variables, you have noise variables and you require to have certain output quality. So, can there be relationship between output quality to input quality process variables or process conditions and noise present while producing the worm gear. So, this is our sole purpose for modelling the total system and there are many techniques available to model, but here we will use multiple linear regression.

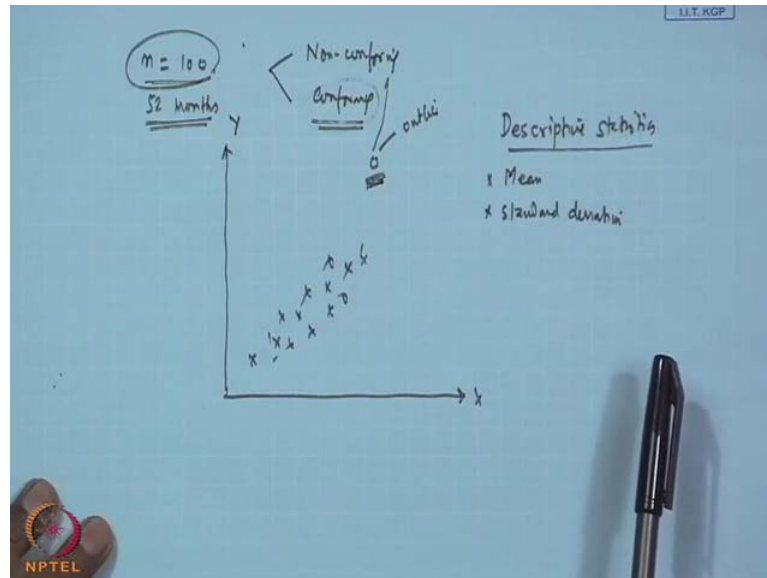
(Refer Slide Time: 11:45)



So, now in order to what will be the next step? Your next step will be data collection means, you must know that what is the manufacturing system here. That is the hobbing process, we have given here and you have to collect data. The hobbing process is totally

characterised by their, with their specification. Here in this figure clearly it is given, ok? So, 100 conforming observations over 52 months were collected from this plant.

(Refer Slide Time: 12:26)



So, that means what is our n sample size, n equal to 100. How many what is the period of collection, 52 months period of collections. So, during 50 months the production unit produces certain worm gears, certain number of worm gears out of which some are rejected or that goes to reworks and some are accepted. Though the rejected that is known as nonconforming and those accepted, those are known as conforming. So, this conforming items are considered conforming observations are considered. So, 100 conforming items were considered, ok?

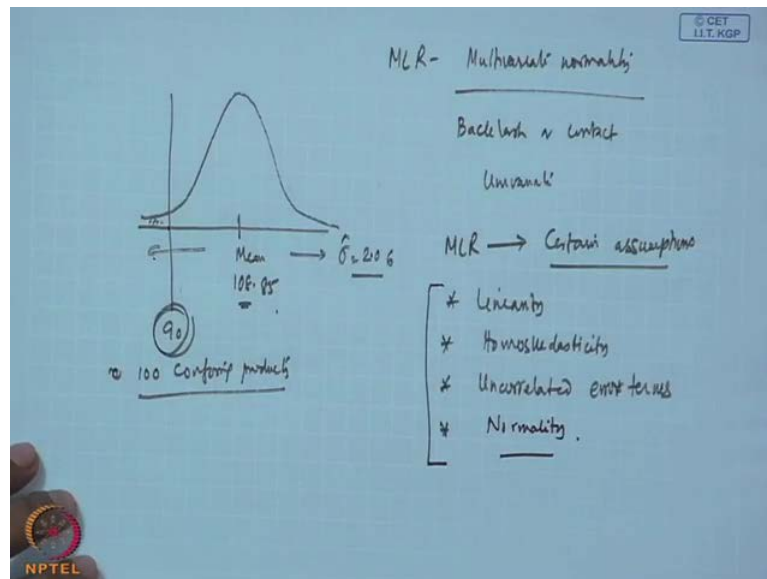
So, why 100 conforming? One is considered, because we want to model a regression or frame a regression model in such a manner that it will basically talk about the general behaviour of the process, you are getting me? Suppose, in case we have two variables. Suppose, Y versus X and this is my data set collected and the scatter plot is showing like this some data point is here. So, it is outlier and you will find out it will be most likely nonconforming. So, we do not want this type of observations in the model fitting. So, we will be considering only the general mass here. From customers point of view all conforming products are considered correct.

So, this is our data when you collect data from case study point to immediately, we have to find out the descriptive statistics. By descriptive statistics when you have continuous

data, we talk about mean and standard deviation. This will give you much idea about the behaviour of the process, getting me? Now, see if I see the 100 that means wheel rim, worm wheel rim that is considered. So, here the hardness the mean value is 108.85, standard deviation is 2.06, copper mean value is 88.05, deviation is 0.29, silicon 10.32 0.25, nickel 1.50 and standard deviation is 0.01, phosphorous 0.014 and that standard deviation is very less.

So similarly, if you see the backlash that is 1.44 is the mean and 0.09 is the standard deviation, contact 35.29 degree that we are talking about. This is in terms of percentage and 2.59 is the standard deviation. Now similarly, humidity all over the year it is 80.38 that is the mean value and 4.87 is the standard deviation. And here the standard deviations are 8.21 1.42, standard deviation 8.21 is the mean value for speed. Similarly, feed and similarly depth of cut if you assume that the process is multi multivariate normal, then what will happen? The individual variables will be univariate normal.

(Refer Slide Time: 16:31)



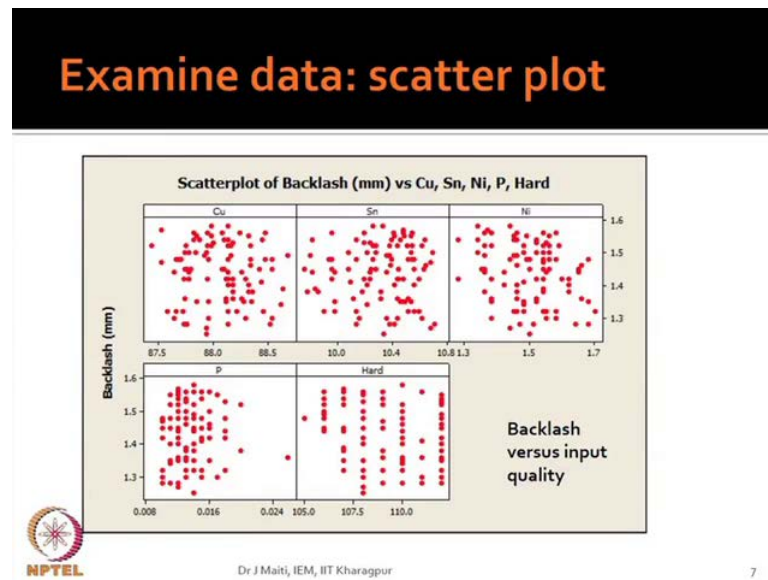
So, under such conditions for everywhere you can find out what is the mean value and then as you know the standard deviations, you will have a distribution like this. For example, if I consider hardness, so hardness that data mean is 108.85 and this side the sigma cap that is 2.06. Now, what is the specification given for hardness, that it should be greater than 90. If it is 100, it maybe 90 will be somewhere here, 90 will be somewhere here. Now, we have to see that what is this percentage coming here.

So, these are the first level of work you have to do before fitting into the this, but as I told you that n 100 nonconforming products are considered conforming products are considered. So, it seem it indicates that it does not indicate that this one the hardness is not less than 90, because this conforming or nonconforming these are basically related to the backlash. The worm wheel quality related to worm wheel quality.

So, you will be having a idea with the mean and standard deviation and the your this one, what is this? I said the normal distribution, this is the first level. So, M L R talks about that one of the assumption is multivariate normality multivariate normality. So, we will assume that that the backlash or contact they will definitely be univariate normality, in addition what we have assumed that even in the process variables they also behave like a coming from a normal distribution or normal population. Then that is also we are considering, although in multiplication that we are not interested in the X distribution. We are rather interested in the Y distribution, because error will capture the Y spot and X will be treated as a fixed values, different fixed values, ok?

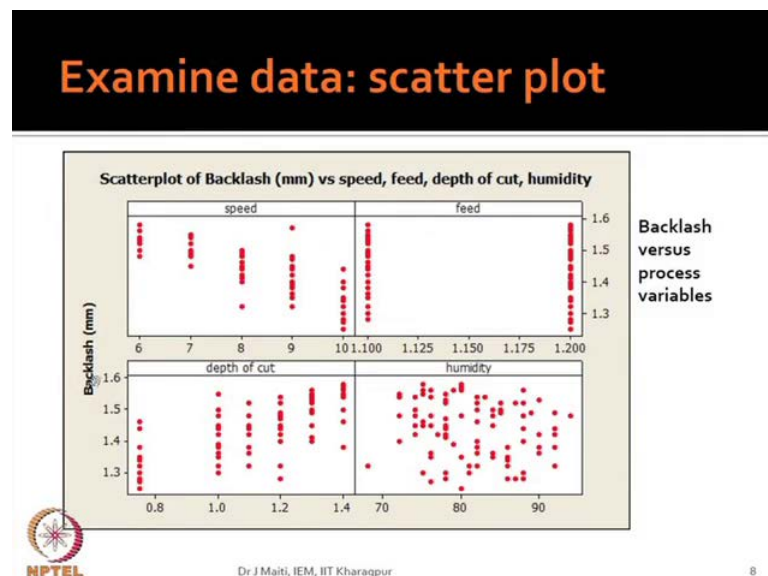
So, next step is we will all of you know that M L R that certain assumptions, what are those assumptions that must be followed? For assumptions will be, one will be linearity, second one will be homoskedasticity, third one will be uncorrelated error terms and fourth one will be normality. What we have seen that errors are independent, identically distributed and errors follows, basically normal errors will be normally distributed also that is why normality, ok? So, we should not wait for model feet and then get the your error value and then you test the normality. That we will do definitely only fit the model, but here before that let us see from the data raw data collected. We will get any kind of information which will help us to judge about the assumptions of the model.

(Refer Slide Time: 20:44)



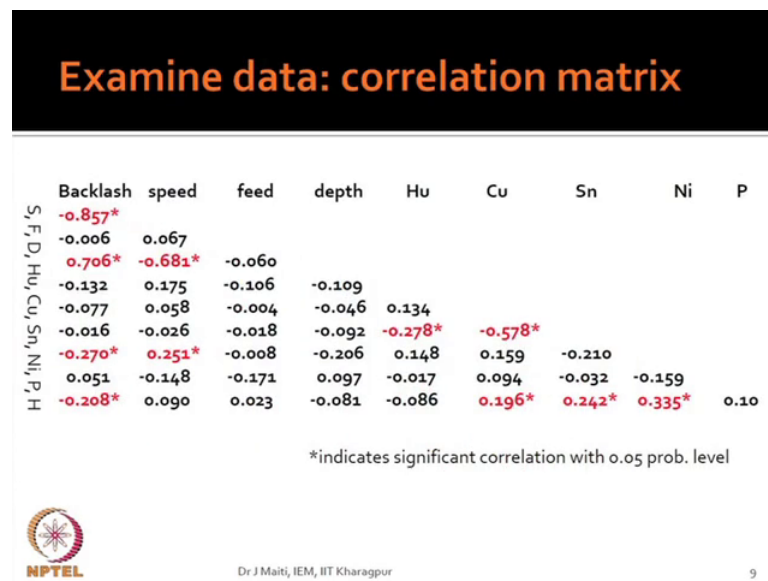
For example, if you see one of the things is the linearity with respect to the Y variable, if we consider Y is backlash here, although Y can be contact. But first we will consider that Y is backlash, then backlash versus copper you see it almost happens at random relationship. Perhaps, that copper percentage is not contributing perhaps similarly, if you see silicon, if you see nickel, phosphorous and hardness it is very difficult to say there is relationship, but it is also not true that there is non-linear relationship, ok?

(Refer Slide Time: 21:30)



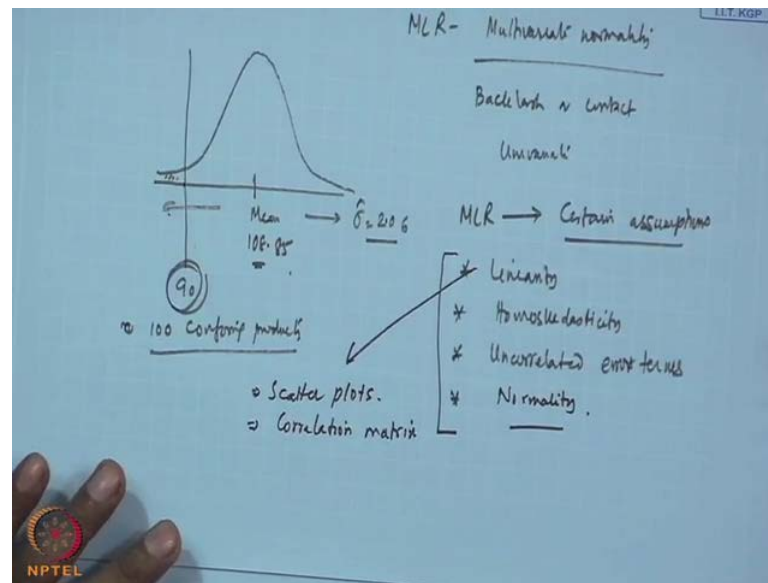
So, if you see the second plot where we are talking about backlash versus speed. It is continuously decreased backlash with respect to increase in speed, you see this way it is coming. Similarly, if you see feed your, it is difficult to tell, but the variability here is more here is less. But I am not clear that whether there is any decrease or increase in relationships with increase in speed, backlash will decrease or increase? That type of relation is not there, but in case of depth of cut there is clear cut increase in relationship, but in case of humidity again it is random, very difficult.

(Refer Slide Time: 22:12)



So, first is that means what is you require?

(Refer Slide Time: 22:19)



You require to see the scatter plot, then you will find out certain relationship, but these are subjective. Then you go for correlation matrix, you go for correlation matrix. So, correlation matrix we have computed here and the correlation matrix is shown like this. So, backlash vis a vis speed feed depth of cut humidity, these are process related variables. Copper, silicon, nickel, phosphorous and hardness, these are input quality variable. Now, if you see backlash vis a vis all those input that independent variables. What you are considering in this case? You see that backlash versus speed there is negative correlation and which is significant.

Similarly, backlash versus depth of cut there is positive correlation which is also significant. Backlash versus nickel, there is negative correlation and significant correlation and backlash versus hardness, this negative, but significant correlation. So, this correlation coefficient will give you certain idea, that whether these variables, the independent variables so called independent variables in the model, they are contributing to the dependent variable backlash or not. In regression, dependent and independent side will be there.

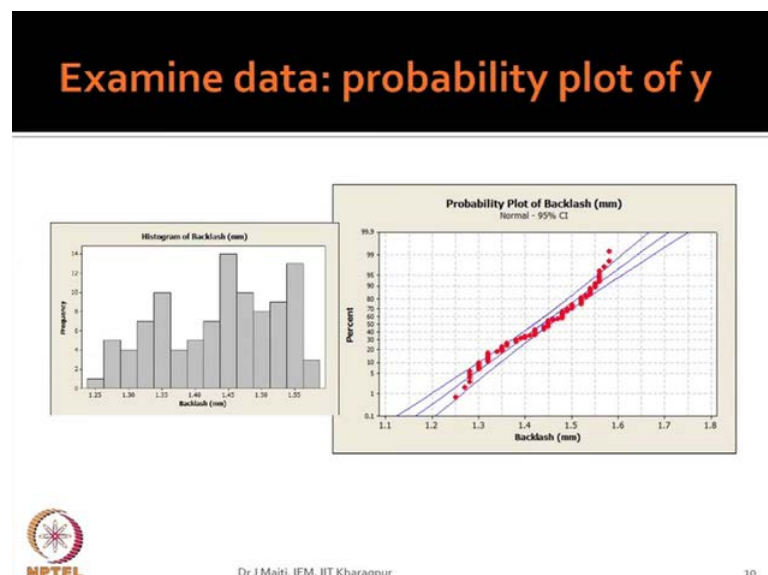
So, in this case we are talking about one dependent variable that is backlash and several independent variables 1, 2, 3, 4, 5, 6, 7, 8, 9 independent variables from correlation matrix. One another important aspects you are finding out that the independent variables are also not perfectly independent, because there are some correlations between the

independent variables which are also significant. For example, speed versus your depth of cut that there is negative significant correlation. Similarly, speed versus nickel, positive significant correlation humidity vis a vis silicon, there is negative correlation. And copper vis a vis silicon, there is negative significant correlation. Similarly, copper vis a vis hardness, this correlation is there.

So, we can say that there are certain correlations amongst the independent variable, but there are many other coefficients correlation coefficients which are not significant. So, we cannot say 100 percent that these variables are independent. In that sense, from correlation structure point of view, but what we can say that we can proceed for regression, because these type of correlated structure.

In the independent variables if it is large enough, then it will distort and it will be also known from the diagnostics from variance, that is implacent factor variance, implacent factor another things are there we will check. At this point in time what I mean to say yes that there are certain independent variables which are basically contributing towards backlash, that is the dependent variable. So, our aim is we can go for multiple regression, correct?

(Refer Slide Time: 26:09)



So, what we have tested? Then we have tested that the linear relationship is there and we can go from multi multiple linear co regression, but we also want to test from the raw data. We want to examine that the normality of the Y variable. So, if you know how to

go for normality, there are different approaches. If you plot that is probability plot, you can see the histogram, you can go for quantile plot.

So, there are many plots. Now, you see this P P plot, here if you see the histogram of backlash, it is not perfectly normal. You cannot say this is the 100 percent normal, but little departure from the normality and if we see the probability plot, also there are certain points out of the confidence band. So, it is not 100 percent normal, but the departure also not that much that we should be worried.

(Refer Slide Time: 27:22)


Model and results

Backlash = 1.35 + 0.0086 Cu + 0.0133 Sn - 0.0088 Ni - 2.39 P
- 0.00583 Hard - 0.0433 speed + 0.0839 feed
+ 0.105 depth of cut + 0.000239 humidity

R-Sq = 78.7% R-Sq(adj) = 76.5%

Analysis of Variance

| Source | DF | SS | MS | F | P |
|------------|----|------|-------|-------|-------|
| Regression | 9 | 0.60 | 0.067 | 36.87 | 0.000 |
| Residual | 90 | 0.16 | 0.002 | | |
| Total | 99 | 0.76 | | | |



Dr J Maiti, IEM, IIT Kharagpur 11

At this point in time we can we can think for those departures after model diagnostics, then you have to feed the model. So, this is our regression fit or regression equation.

(Refer Slide Time: 27:36)

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

$$X = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix}$$

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

$p = 9$
 $p+1 = 10$

$$y = X\beta + \epsilon$$

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_9 \end{bmatrix}$$

Labels for $\hat{\beta}$: constant, C_u , S_n , Ni, P, H, S, F, D, H_0 .

I am sure that you can recollect that what is the formula? We have used to calculate beta. Beta cap is X transpose X inverse X transpose y, this is the regression equation. We have used and using the square X is your desired matrix 1 1 1 to 1 and like this your X 1 1, X 2 1, then X n 1. Similarly, X 1 p, X 2 p like this X n p and y is here one variable only. So, y 1, y 1 dot dot dot y n and your equation regression equation is y is X beta plus epsilon, where y is n cross 1 X is basically n into p plus 1. So, beta definitely will be p plus 1 cross 1 and epsilon will be n cross 1.

So, beta our beta is beta 0, beta 1 to beta p. In this case we have p equal to 9 so your total p plus 1 will be 10 and your beta estimated is that is why first one is the constant, second one is our we have taken the that copper related. Then silicon related, nickel related, phosphorous related, hardness related, then also your speed feed depth of cut and humidity speed feed depth of cut and humidity. So, this many things will be there. So, I can say this is my beta 0 cap, this is beta 1 cap. So, like this 1, 2, 3, 4, 5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10. So, this one will be your beta 9 cap, ok? Now, what are the values, we found out from our regression model.

(Refer Slide Time: 30:05)

$\hat{\beta} = \begin{bmatrix} 1.35 \\ 0.0086 \\ 0.0133 \\ -0.0088 \\ -2.39 \\ -0.0058 \\ -0.0433 \\ +0.084 \\ +0.105 \\ +0.00024 \end{bmatrix}$
 $R^2 = 90\%$

$\hat{y} = X\hat{\beta}$
 $\hat{e} = y - \hat{y}$

$SSE, SSR = SST - SSE$
 $R^2 = \frac{SSR}{SST} = \frac{0.60}{0.76} = 78.7\%$
 $R_a^2 = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)} = 76.5\%$
 $SSE = 0.76 - 0.60 = 0.16$
 $n=100, p=10, n-p-1=90$

We found out like this, that is 1.35 then 0.0086, 0.0133, minus 0.0088, minus 2.39, minus 0.0058, minus 0.0433, plus 0.084, plus 0.105, plus 0.00024. So, this is your beta value for the regression equation. We found out this is related constant and all these things, fine? So, once you have estimated equation that the fitted equation, that mean we are now having \hat{y} equal to $X\hat{\beta}$ having this is my beta cap.

So, you are able to find out the error, error is $y - \hat{y}$ as you are able to find out the error. Now, you are also able to find out the $SS E$ sum square errors, you are able to find out $SS R$ which is $SS T$ minus $SS E$. You know the degrees of freedom, so you can go for finding out R square. R square is $SS R$ by $SS T$ and in this case our $SS R$ is 0.60 and $SS T$ is 0.76 which is giving you 78.7 percentage R square value.

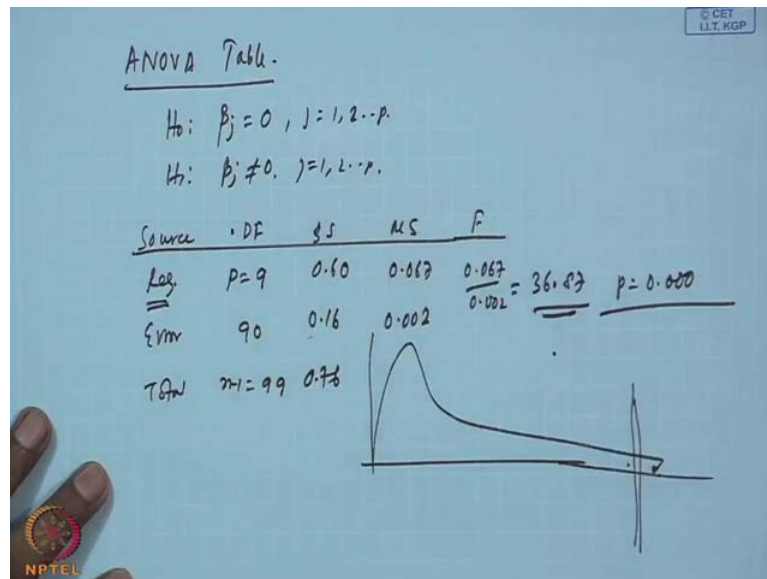
So, that means what I can say in total 79 percent variability of backlash can be explained by the independent variable. Considered now whether 78 percent or 79 percent is to be considered significant or not. Considering the case whether what we have taken, we have taken data from the day to day operation, it is not that experimental data. So, there are lot of variability involved. So, as a result, so we cannot expect that 95 percent R square we will be getting. So, I think it is almost 80 percent.

So, it is considering the variability involved form material, from operation, from production, from service, many points of view. So, I think this this is a reasonable model to consider. Once you consider that R square, then you also may be interested to know

whether there is the effect of parameters to the number of observations. So, n and n and p , that relationships that you want to test.

So, R^2 a square what you will do, R^2 a square will be $1 - \frac{SSE}{SST}$ by degrees of freedom, that is $n - p - 1$ by SST by degrees of freedom. That $n - 1$ in this case, this one is 75.65 percent, because our SSE is I think 0.60 0.16 and your n is 100, $p + 1$ is 10. So, $n - p - 1$ is 90. So, $1 - \frac{0.60}{0.76}$ by your $n - 1$ is 99. This is giving you 76.5 percent variability of backlash explained, getting me? Okay? So, this is 1, so R^2 is not bad. Now, we will go for anova table, your test case study must show this.

(Refer Slide Time: 34:39)



Anova anova table we test to hypothesis β_j equal to 0, for j equal to 1 to p . And all case and H_1 at least one β_j not equal to 0, that is the case. Then you calculate the anova table, anova table will be your source then degree of freedom SS MS , then you test the F source is one is regression, second one is error, third one is your that total is divided like this. So, your degrees of freedom is how much? There are $p + 1$ 10, so it will be p , it will be 9 and total will be $n - 1$ equal to 99.

So, error will be that the difference between the two, this is 90. Our SS regression is 0.60, error 0.16 and 0.76, then MS is SS by degrees of freedom 0.60 by 9. This will be 0.067 then 0.16 by 90, this will be 0.002. So, your F value will be 0.067 by 0.002 this is

a very high value 36.87, very high value. So, very high value means you are, if this is the case then it may be falling somewhere here.

So, probability if you calculate, probability it will show almost show 0 0 0. So, that means the independent variables are actually contributing in explaining the variability which is also obvious from the R square table, ok? So, we are accepting the model from overall adequacy point of view. So, then we require to test, what we require to test? We require to test the individual parameters test of parameters.

(Refer Slide Time: 37:00)

Test of parameters.

$$\frac{\hat{\beta}_j - E(\hat{\beta}_j)}{SE(\hat{\beta}_j)} \sim t_{n-p-1} \text{ under } H_0$$

$$E(\hat{\beta}_j) = \beta_j = 0 \quad t_{n-p-1} \sim \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$$

$$SE(\hat{\beta}_j) = s_e \sqrt{(X^T X)^{-1}} \quad s_e^2 = \frac{SSE}{n-p-1}$$


Test of parameters and you all know what we say that beta j cap minus expected value of beta j cap by standard error of beta j cap, this follows t n minus p minus one under H 0 null hypothesis. So, under null hypothesis expected value of beta j cap equal to beta j equal to 0. So, ultimately your t n minus p minus 1, what you are basically computing getting from this equation, that will be that to be compared, that to be compared with beta j cap by S E beta j cap, and how do you know S E beta j cap? This will be s e square X transpose X inverse, you have seen earlier. Now, what is s e square? s e square is S S E by n minus p minus 1 degrees of freedom. So, you know everything, now see this slide.

(Refer Slide Time: 38:16)

Model and results

Backlash = 1.35 + 0.0086 Cu + 0.0133 Sn - 0.0088 Ni - 2.39 P
 - 0.00583 Hard - 0.0433 speed + 0.0839 feed
 + 0.105 depth of cut + 0.000239 humidity

| Predictor | Coef | SE Coef | T | P |
|--------------|-----------|-----------|--------|-------|
| Constant | 1.346 | 1.912 | 0.70 | 0.483 |
| Cu | 0.00860 | 0.02051 | 0.42 | 0.676 |
| Sn | 0.01332 | 0.02666 | 0.50 | 0.619 |
| Ni | -0.00884 | 0.05295 | -0.17 | 0.868 |
| P | -2.394 | 1.773 | -1.35 | 0.180 |
| Hard | -0.005833 | 0.002648 | -2.20 | 0.030 |
| speed | -0.043340 | 0.004227 | -10.25 | 0.000 |
| feed | 0.08395 | 0.08762 | 0.96 | 0.341 |
| depth of cut | 0.10467 | 0.03105 | 3.37 | 0.001 |
| humidity | 0.0002394 | 0.0007886 | 0.30 | 0.762 |



Dr. J. Maiti, IEM, IIT Kharagpur

13

This is the slide the constant copper silicon to all the predictors, then these are the values, beta values. This is the standard error s_e square into $X^T X^{-1}$, that particular value that corresponding that with that matrix, the corresponding value we have to consider. So, you see here then T value, T value is nothing but coefficient by standard error. If you see hardness, this coefficient is minus 0.005833 and your standard error of the coefficient is solo that the T value is becoming large minus 2.20, which says that that these influence is significant at 3 percent probability level of significance.

So, we consider this is a significant contribution in terms of explaining the variability of backlash. Similarly, speed you see speed also this divided by this is minus 10.25, this is very large. So, the probability value is almost 0 0 0, we have used mini tap to find out this solution feed is not significant, because the probability value is 34 percent. And you see although coefficient value is more than speed coefficient in terms of magnitude, but the standard error is high, much higher compared to speed standard error. And as a result what happens speed values, becomes quite low and it makes it insignificant. Similarly, depth of cut is also very, very significant, ok?

So, then from this result we can say that that, yes multiple regression is able to explain all around 80 percent of variability of backlash. And the variables which are contributing the maximum or significantly contributing, we can say these are hardness speed and


depth of cut, correct? Now, you may be interested to know that what are the individual contribution to the overall sum square.

(Refer Slide Time: 40:44)

Model and results

Backlash = 1.35 + 0.0086 Cu + 0.0133 Sn - 0.0088 Ni - 2.39 P
 - 0.00583 Hard - 0.0433 speed + 0.0839 feed
 + 0.105 depth of cut + 0.000239 humidity

| Source | DF | Seq SS |
|---------------------|----------|-----------------|
| Cu | 1 | 0.004578 |
| Sn | 1 | 0.004292 |
| Ni | 1 | 0.057740 |
| P | 1 | 0.000122 |
| Hard | 1 | 0.004071 |
| speed | 1 | 0.509094 |
| feed | 1 | 0.001161 |
| depth of cut | 1 | 0.020508 |
| humidity | 1 | 0.000167 |



Dr J Maiti, IEM, IIT Kharagpur
 12

You see this is basically table, where which is also obtained through mini tap. Here, what happened? You see that if your total sum square is 0.76 and you will find out the regression sum square is 0.60. So, if you add all those things this will be 0.60 and if I see from the speed is contributing, the maximum 0.509 so divided by 0.60, it will be 90, more than 90 percent, more than 80 percent. Then followed by nickel, followed by depth of cut, but although S S contribution, nickel is more than hardness, but nickel is not significant. Hardness is significant, because of the variability of estimates for hardness is much lower than the variability of estimate for nickel, getting me?

(Refer Slide Time: 42:08)

Test of parameters.

$$\frac{\hat{\beta}_j - E(\hat{\beta}_j)}{SE(\hat{\beta}_j)} \sim t_{n-p-1} \text{ under } H_0.$$

$$E(\hat{\beta}_j) = \beta_j = 0 \quad t_{n-p-1} \sim \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}.$$

$$SE(\hat{\beta}_j) = s_e \sqrt{(X^T X)^{-1}}$$

$$s_e = \frac{SSE}{n-p-1}$$

© CET IIT KGP

NPTTEL

What do we say that this is your T distribution. So, your this is the mean value of beta under H_0 beta j equal to 0. So, suppose your values comes here, the computed value and this side is 0.025 and this side also 0.025. So, if your value falls here then it is insignificant. So, usually we take 0.05 so the p value, what you are getting? If you find out any p value which is less than this, so p less than equal to this, then you blindly accept this, because this is significant with 5 percent error, ok?

(Refer Slide Time: 42:56)

Model adequacy tests

R-Sq = 78.7% R-Sq(adj) = 76.5%

Analysis of Variance

| Source | DF | SS | MS | F | P |
|------------|----|------|-------|-------|-------|
| Regression | 9 | 0.60 | 0.067 | 36.87 | 0.000 |
| Residual | 90 | 0.16 | 0.002 | | |
| Total | 99 | 0.76 | | | |

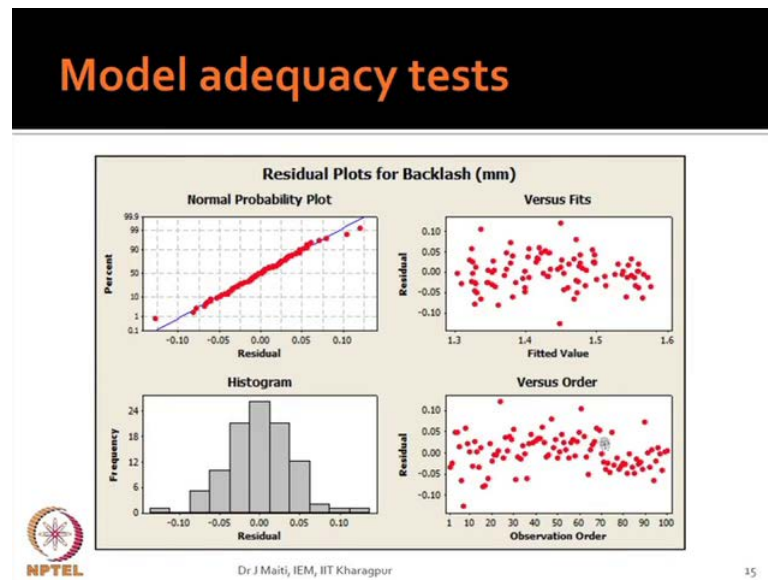
NPTTEL

Dr J Maiti, IEM, IIT Kharagpur

34

Then what we require to do?

(Refer Slide Time: 43:05)



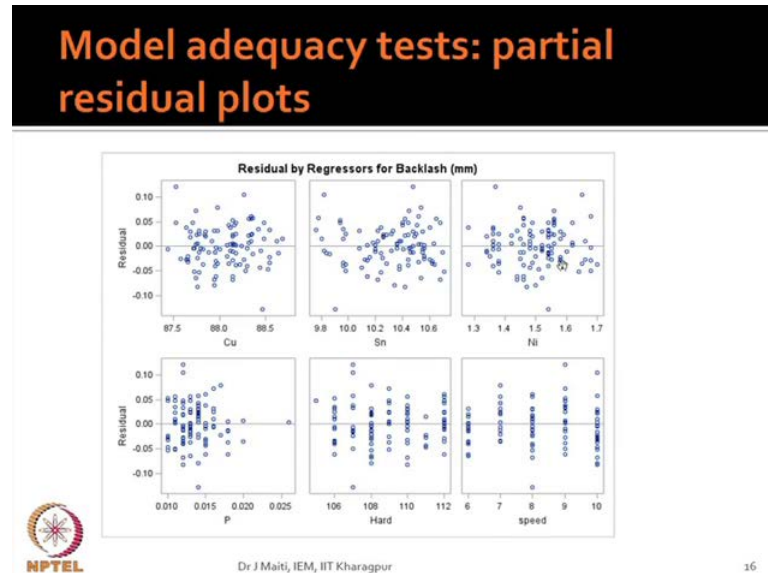
You require to once you estimate then you accept the model, you estimated the parameters effect in terms of T test and then you require to do the model adequacy test, getting me? So, what are the model adequacy tests? Under model adequacy tests definitely what I have told earlier that R square is definitely coming under model adequacy test, that basically talks about the model value does not apart from R square or the analysis of variance. This anova, that one, that F test we require to test the residuals whether residuals are basically following the assumptions of the regression model or not.

So, you see this the diagram. Here, you see although we have seen in y plot probability probability plot for y we find found out that some of the points are out of the straight line within the from the confidence band out of that band. But in the residual case it is very much close to the straight line, and I think all maybe following within the band. If I draw a 95 percent confidence band, it will be coming like this.

So, it simply indicates that that little bit departure has not affected much here. So, that errors are showing normal and if you go for the fitted value versus residual, that mean independent observations it should be that there, should not be any trend. It is there is no trend and if I see the histogram also it is almost normal and if you see thus auto correlation point of residual versus observational order. I do not find out any correlation because there is no trend available, so then what does it signify? It signifies that yes the

model is adequate from residual analysis point of view also, but if you see clearly here is one point, here is one point, some points are little away from the general mass.

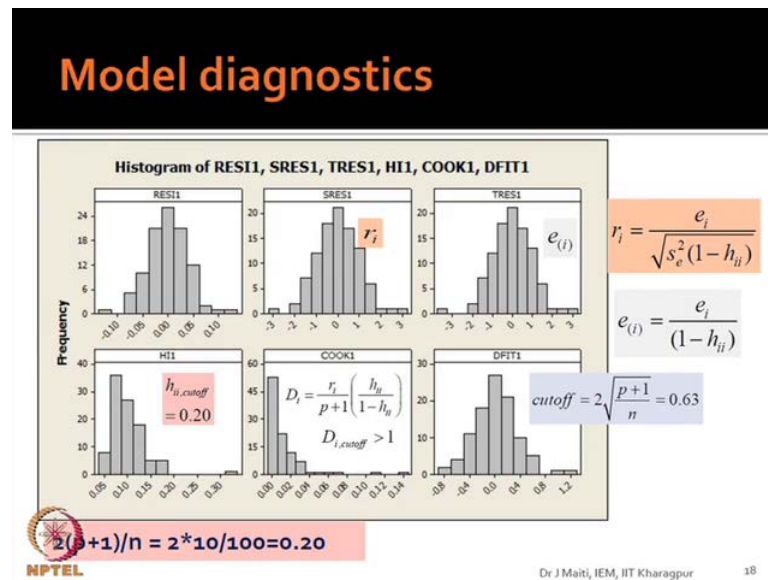
(Refer Slide Time: 45:28)



So, these are other plot, these is other plot like residual versus individual variable plot. Can you remember, I told you that partial residual plot this is in order to see that whether any of the if there is nonlinearity. If any of the independent variables contributing to the nonlinearity or not. So, this one is S output and here you see that residual versus copper, nothing silicon, nothing, we are not getting any where any departure from that.

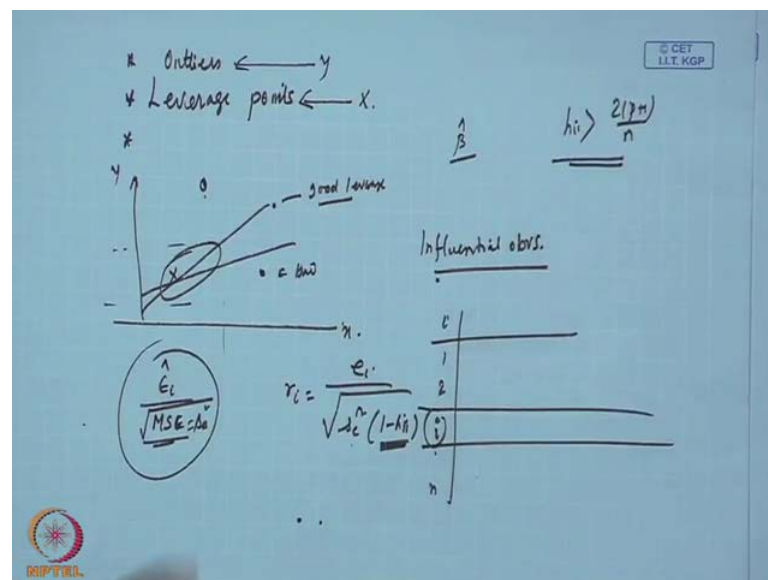
And interestingly what is happening here, this copper silicon nickel they are not contributing much in explaining the variability. And as a result here also you are not getting any trend, if there is any variable which contribute then there will be a trend. So, here I think the speed is contributing. So, this line is little inclined, this line is little inclined, but not apparently visible that inclination.

(Refer Slide Time: 46:44)



There are apart from these four tests, what I have shown you earlier that your probability plot fit versus residual fit versus order versus residual. And your partial residual plot, there are certain other statistics to find out the leverage or points.

(Refer Slide Time: 47:08)



Can you remember leverage points? There are leverage points, there are outliers. So, we talk about outliers with respect to y, we talk about leverage points with respect to X. I think you can remember this, we have given one explanation here. Suppose, this is the

general mass, here is one point and this side is y, this side is x, this is outlier, because this is does not belong with the variability zone of y.

But suppose, some point is here, some point is here, this point is within y variability, but much away from x variability. This is your good outlier good leverage, this is your bad leverage. We say good leverage because under good leverage there is no problem in the regression estimate beta cap bad average distort the line that mean the regression equations are distorted.

So, these are the things so what we want then we also require to find out the leverage points, other way in one words including outliers, you can say influential observations. So, influential observations is very, very important to find out, ok? So, there are many ways to find out influential observations, one is your standardised residuals, second one is your studentized residuals, third one is your phase residuals or deleted residuals, then there is your h_{ii} that is the you know the hat matrix and diagonal elements. These are basically leverage points, then you there is cook's distance, then there is defit, these are different way, different statisticians. They have identified different procedures to identify the your what I can say influential observations.

What is your standardised residual? Standardised residual is suppose, your residual is this one, this divided by M S E, this is your standardised residual. What is your studentized residual? Studentized residual says that this M S E may be affected by large observations, large residuals like outliers. So, then it should be normalised, that normalisation is done here like your studentized residual. If I say r_i which is e_i by some s_e square into $1 - h_{ii}$. So, M S E is here s_e square.

So, we will write then square root of M S E here correct and then this one is. So, this is the leverage point, these values are subtracted from 1 and this is your studentized residuals. This value should not be this all values, should not be large value then there are deleted residuals. Deleted residuals means suppose you have i equal to 1 to n observations. So, i th observations you delete then you model that run, the regression model, find out the residuals for the i th value, ok?

So, deleted residuals can be found out by this, then there is your h_{ii} that leverage point values like this. So, the cut off h_{ii} is 2 into $p + 1$ by n , getting me? 2 into $p + 1$ by

n. So, your h_{ii} value if it is $2p + 1$ by n , it is more than this it is more than this. This is influential, now you can find out deficit, you can find out cook distance.

(Refer Slide Time: 51:49)

$$D_i = \frac{\hat{r}_i}{p+1} \left(\frac{\hat{r}_i}{1-h_{ii}} \right)$$

$$D_i > 1$$

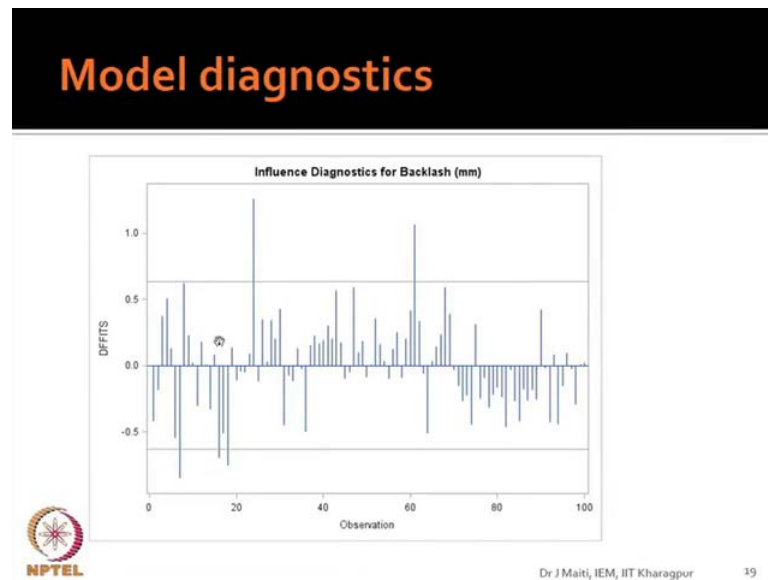
$$DFITS > 2 \sqrt{\frac{p+1}{n}}$$

$$h_{ii} > \frac{2(p+1)}{n}$$
 Montgomery et al (2003)

Now, if you go by cook distance, basically talk about D_i which is r_i by $p + 1$ into h_{ii} by $1 - h_{ii}$, this is your cook distance. It is cut off value if D_i value greater than 1, this is influential. So, what I say? I first say that h_{ii} greater than $2p + 1$ by n , that is influential D_i value, following this if greater than 1 that is influential. So, there is deficits value also $DFITS$ this deficits cut off is, if this one is greater than root over $p + 1$ by n that also denotes influential observations.

So, this how this difference residuals measures are coming, what is the logic behind it? The logic is ultimately you will find out that similar like that residual then normalised residual, then adjusted residuals and like this, but cook distance and deficits they talk about talk from the model prediction power point of view, ok? So, there is very good discussion of all those things in Montgomery et al under model diagnostics, I think chapter 5 probably under model diagnostics.

(Refer Slide Time: 53:31)



So, you see here what happened in this slide, for the particular case our D_i value defits value, I shared the defits value $2 \sqrt{\frac{1}{p+1} \frac{1}{n}}$. This defits value, this defits value this $2 \sqrt{\frac{1}{p+1} \frac{1}{n}}$ means $10 \sqrt{\frac{1}{100}}$, I think and this value $2 \sqrt{\frac{1}{p+1} \frac{1}{n}}$ $10 \sqrt{\frac{1}{100}}$ value. So, ultimately if you see this is coming under 0.63, I will go to the earlier slide. It will be clear this is 0.63 cut off, so if you put a absolute cut off value 0.63 and all the 100 observations residuals, if you plot what you are getting?

You are getting some observations, particularly this observation. This observations higher and here also they are crossing there are some observations, I can say 1, 2, 3, 4, 5 observations are influential observations. Although they are not distorting much about the regression estimate maybe, because the departure may not be that high, but if that this this value is very high large value, it will distort. So, what will be your action? Action will be it is better to remove those observations, you go for one more time regression model after removing the leverage bad leverage points.


(Refer Slide Time: 55:20)

Model diagnostics: unusual observations

| Obs | Cu | Backlash | Fit | SE-Fit | Residual | St Resid |
|-----|------|----------|---------|---------|----------|----------|
| 7 | 88.5 | 1.32 | 1.44747 | 0.01075 | -0.12747 | -3.09R |
| 16 | 87.8 | 1.28 | 1.36127 | 0.01374 | -0.08127 | -2.02R |
| 23 | 87.7 | 1.36 | 1.35531 | 0.02385 | 0.00469 | 0.13X |
| 24 | 87.5 | 1.57 | 1.44947 | 0.01561 | 0.12053 | 3.04R |
| 61 | 88.3 | 1.44 | 1.33598 | 0.01557 | 0.10402 | 2.62R |

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large leverage.

Durbin-Watson statistic = 1.50256 indicating little positive autocorrelation



Dr J Maiti, IEM, IIT Kharagpur 20

So, I told you there are five unusual observations, 1, 2, 3, 4, 5, correct? So, this also available in software, you ask mini tap it will give you what are the observations which are influential to be considered. So, then you see and what are the variables that is another important issue is that, see what are the variables. So, which are making the observations influential?

This copper is from the input X point of view, copper is making almost all the things influential and the first one is this is residual. This R denotes for large standardised residuals and X denotes whose X values is large leverage. So, it is that means this 23 has this observations has influence on in distorting the regression estimates, but residuals will not distort the estimates. So, that mean X that 23 observations should be removed, getting me?

So, whenever you find problem in fitting regression model, definitely you have to see the unusual observations and then if possible remove the unusual observations, and then again you rerun the model. And you require, what you require? More number of observations, that is why and when you find out that your regression coefficient R square value is very, very low. Then no need of going to all those things, because it will not give you any information.

So, R square is vital. So, if R square value is very low that means either you have not taken the right kind of independent and dependent variables or your data collection is


faulty or what will happen? Basically, the methodology you adopted in measuring the data that is also not correct. So, there are many issues, so that to be taken into consideration. So, what in nutshell? What I mean to say then that this particular case study shows you how you will you go for multiple linear regressions if applicable for the problem you are considering. So, apart from these things, another issue is that I have told you that auto correlation that Durbin-Watson statistics, here also we have used it is 1.5. So, Durbin-Watson or D W statistic should be 2 for no correlation, but it is 1.5. So, slight correlation is there, so 1 minus R is 1.5. So, it is negative positive autocorrelation is 3, ok?

(Refer Slide Time: 58:39)

Multicollinearity

| Predictor | VIF |
|--------------|-------|
| Cu | 1.940 |
| Sn | 2.332 |
| Ni | 1.426 |
| P | 1.092 |
| Hard | 1.620 |
| speed | 1.976 |
| feed | 1.058 |
| depth of cut | 1.978 |
| humidity | 1.146 |

No multicollinearity problem

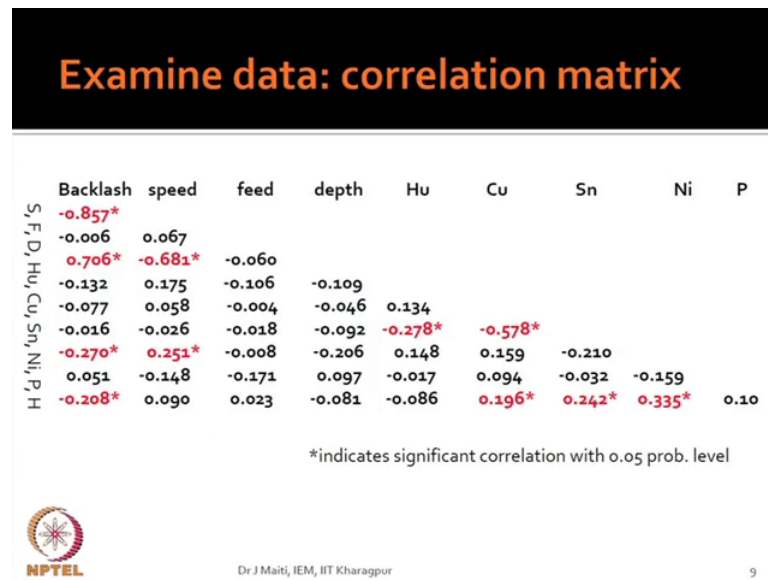


Dr J Maiti, IEM, IIT Kharagpur

21

Next issue is multicollinearity, you have to test it your things are fine. It is settled that you are happy with 80 percent variability, but whether there is multicollinearity because then if you go back to your regression, not regression.

(Refer Slide Time: 59:01)



That is correlation matrix you have seen that there are some of the variables in the independent side which are having high significant correlation coefficient, particularly speed versus depth. There is high correlation, this is high but not that high that it will distort your regression estimate that much. And that is also happened and you have seen that that much disturbance you have not faced apart from your R square value which is almost 0.8.

But R square cannot be improved unless you go for other kind of variables which are basically contributing here, because this type of process is process basically operates under many variables operators conditions, operators expertise, operators skill, those are not considered. So, how can we think that we will be getting that 90 percent or 95 percent of variability explained, it is not possible. So, now multicollinearity problem is there or not? So, 1.90 all those things I think can you remember the multicollinearity number? Can you remember that V I F variance impression factor?

(Refer Slide Time: 01:00:18)

$D_i = \frac{\sigma_i}{p+1} \left(\frac{h_{ii}}{1-h_{ii}} \right)$
 $D_i > 1$
 $h_{ii} > \frac{2(p+1)}{n}$
 $DFITS > 2 \sqrt{\frac{p+1}{n}}$
 $VIF = \frac{1}{1-R_j^2}$
 Montgomery et al (2003)

This is $1 - R_j^2$, remember? So, R_j^2 means what happened we have taken this X_j as a dependent variable and all other independent variables which contributing this, then you found out that R_j^2 . Now, if your R_j^2 is 0, then VIF will be 1, if your R_j^2 is 1 then VIF will be infinite, very large VIF.

(Refer Slide Time: 01:00:49)

Multicollinearity

| Predictor | VIF |
|--------------|-------|
| Cu | 1.940 |
| Sn | 2.332 |
| Ni | 1.426 |
| P | 1.092 |
| Hard | 1.620 |
| speed | 1.976 |
| feed | 1.058 |
| depth of cut | 1.978 |
| humidity | 1.146 |

No multicollinearity problem

Dr J Maiti, IEM, IIT Kharagpur 21

Now, in this case your VIF or none of the VIF values are large. So, they are not showing any multicollinearity problems. So, that means even though one or two things are like this, but no multicollinearity problem is there, getting me? So, then what is your


conclusion about the process that the process variables are able to explain including the input from input that input material, but is able to explain and we have found that two variables emerges significant.

(Refer Slide Time: 01:01:33)

Model and results

Backlash = 1.35 + 0.0086 Cu + 0.0133 Sn - 0.0088 Ni - 2.39 P
 - 0.00583 Hard - 0.0433 speed + 0.0839 feed
 + 0.105 depth of cut + 0.000239 humidity

| Predictor | Coef | SE Coef | T | P |
|--------------|-----------|-----------|--------|-------|
| Constant | 1.346 | 1.912 | 0.70 | 0.483 |
| Cu | 0.00860 | 0.02051 | 0.42 | 0.676 |
| Sn | 0.01332 | 0.02666 | 0.50 | 0.619 |
| Ni | -0.00884 | 0.05295 | -0.17 | 0.868 |
| P | -2.394 | 1.773 | -1.35 | 0.180 |
| Hard | -0.005833 | 0.002648 | -2.20 | 0.030 |
| speed | -0.043340 | 0.004227 | -10.25 | 0.000 |
| feed | 0.08395 | 0.08762 | 0.96 | 0.341 |
| depth of cut | 0.10467 | 0.03105 | 3.37 | 0.001 |
| humidity | 0.0002394 | 0.0007886 | 0.30 | 0.762 |


Dr J Maiti, IEM, IIT Kharagpur 33

Three variables, but that is speed and depth of cut. These are process variables where you have control, hardness cannot be controlled because unless it will be controlled by the centrifugal casting shop not here. So, we have control, you can improve the backlash. This is one step, but it requires further the optimisation that what will be the setting points process, setting points that statistical optimisation using response surface methodology can be found out, response surface methodology. So, we will not discuss this. In other word, we have considered only linear regression that maybe nonlinearity, there may be quadratic effects, there may be interaction effects we have not considered, but those things also required to be considered for a full model, different kind of model.

Thank you very much.