

**INDIAN INSTITUTE  
OF  
TECHNOLOGY  
KHARAGPUR**

**NPTEL  
National Programme  
on  
Technology Enhanced Learning**

**Applied Multivariate Statistical Modeling**

**Prof. J. Maiti  
Department of Industrial Engineering and Management  
IIT Kharagpur**

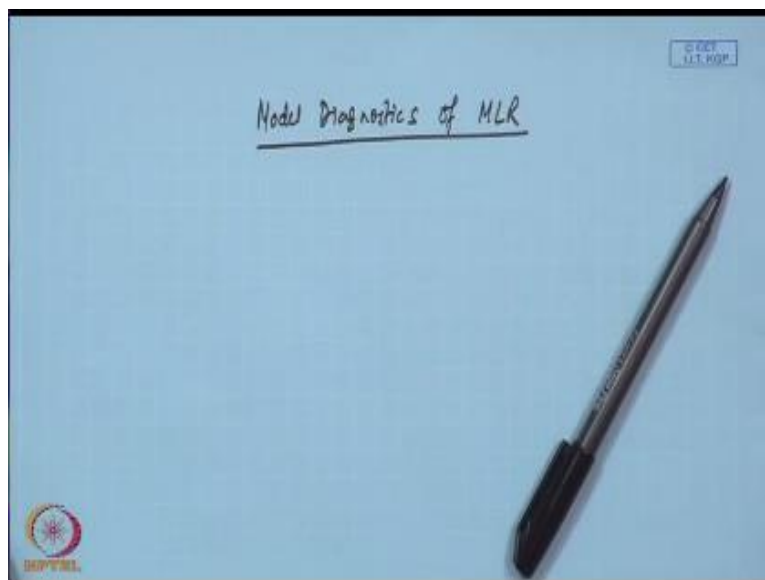
**Lecture – 25**

**Topic**

**MLR -- Model diagnostics**

So, we will start now, model diagnostics of multiple linear regressions.

(Refer Slide Time: 00:23)




So under model diagnostics what are the issues we will be covering?

(Refer Slide Time: 00:39)

**Contents**

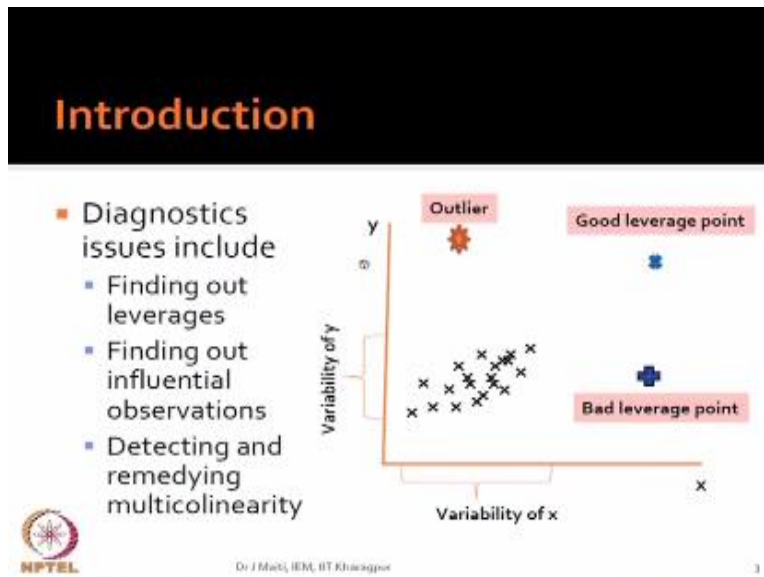
- Introduction
- Leverage
- Influential points
- Multicollinearity

 NPTEL

Dr J Mahi, IIT Kharagpur 2

Leverage points, influential points and multicollinearity.

(Refer Slide Time: 00:48)



Now, you see this figure so it is a scatter plot between  $y$  and  $x$  and if you see the majority of the data points they are scattered around this ellipse, and if we consider the major majority of this point or the mass of the points. Then you see that this is the variability of  $y$ , this is a range where  $y$  varies and this side is the variability of  $x$  with respect to the mass of the data points. Now, you consider this point, suppose your observations, one of the observations is observations is like this, this one lies that much distance away from this centre of this mass of the points. But, it lies in the, in the direction of  $y$  you see the variability of  $y$  is this one and it is basically much away from that that  $y$  portion.

But, if you see this portion for the  $x$  it is within this variability, so outlier is a point which is necessarily related to the variable  $y$ . So, outlier is an observation which lies much away from the general mass related to  $y$ . Now, you come to the other two points this point bishop is this point if you see this point which is if I say the variability of  $y$  it, basically belongs to this variability that range within this range along  $y$ . But, along  $x$  if we see this is away from the general range of the  $x$  and similarly the other one also, this one also.

Now leverage point is a point which lies beyond the, that general mass of  $x$ . So, that means outlier is necessarily related to the  $y$  related observations and  $x$  leverage points related to  $x$  variability that range point of view. Now, all these observations can have influence on the regression estimates, if any observations which influence the regression estimate is known as influential observations okay generally what will happen. You will found out that outlier will not affect the regression estimate much, but the leverage point will affect which.

For example there is good leverage point this good leverage point is one which is not which is basically almost lining on the straight line you see if I draw a straight line here, the regression line it is very close to the regression line although it is out of the, I mean far away from the general mass of the data points. But, from the regression point of view what is happening, it is basically lying almost on the regression line. So, it may be representing something different which is which will help in understanding behavior of the system that is why it is good it is not distorting the regression line regression line.

But, the bad leverage point is this one what will because of this your regression line will shift okay so by regression diagnostic in case of multiple regression we try to find out all those influential observations including outliers leverage points, good leverage as well as bad leverage points many time what will happen suppose one point is, here somewhere here okay apparently it looks that there is no problem with this data with this particular observation but, if you carefully observe the errors you will find out that this has influence also in the regression line.

So, today's discussion we want to find out the leverages, we want to find out other influential observations and another issue which is also very important that our one of the assumption is that independent variables all the explanatory variables are independent in nature, so if they are there is some amount of dependency between or amongst the independent variables what will happen, it leads to again distortion in the estimates and that is termed as multicollinearity. So, we will find out how to identify these observations what are the remedies to high leverages or high influential points, and what is multicollinearity and how to detect and remedy multicollinearity.

(Refer Slide Time: 06:56)

## An example

Sl. No.	Months	Profit in Rs million	Sales volume in 1000	Absenteeism in %	Machine breakdown in hours	M-Ratio
1	April	10	100	9	62	1
2	May	12	110	8	58	1.3
3	June	11	105	7	64	1.2
4	July	9	94	14	60	0.8
5	Aug	9	95	12	63	0.8
6	Sep	10	99	10	57	0.9
7	Oct	11	104	7	55	1
8	Nov	12	108	4	56	1.2
9	Dec	11	105	6	59	1.1
10	Jan	10	98	5	61	1.0
11	Feb	11	103	7	57	1.2
	March	12	110	6	60	1.2



This is our example and these are the.


(Refer Slide Time: 06:59)

## Regression parameter estimates

$Y = 130.22 - 1.24X_1 - 0.30X_2 + e$

Observed	Fitted	Residuals
100	100.44	-0.44
110	102.88	7.12
105	102.32	2.68
94	94.82	-0.82
95	96.41	-1.41
99	100.69	-1.69
104	105.02	-1.02
108	108.45	-0.45
105	105.07	-0.07
98	105.71	-7.71
103	104.42	-1.42
110	104.77	5.23

$C = (XTX)^{-1}$		
40.9	0.114	-0.703
0.11	0.012	-0.004
-0.7	-0.004	0.012
SSE	$Y'(I-H)Y$	155
$se^2$	$SSE/(n-p-1)$	17.22



Dr. J. Mallick, IITM, IITKharagpur 5

Fitted values.


(Refer Slide Time: 07:00)

## Goodness of fit test

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	171.23	85.62	5.01	0.034
Residual Error	9	153.68	17.08		
Total	11	324.92			

**R-Sq = 52.7% R-Sq(adj) = 42.2%**



Dr. J. Malvi, IEM, IIT Kharagpur


6

And regression lines.

(Refer Slide Time: 07:02)

### Parameter test

Predictor	Coef	SE	T	P	VIF
Constant	130.22	26.43	4.93	0.001	
Absenteeism%	-1.2432	0.4480	-2.78	0.022	1.092
Machine BH	-0.2999	0.4586	-0.65	0.529	1.092



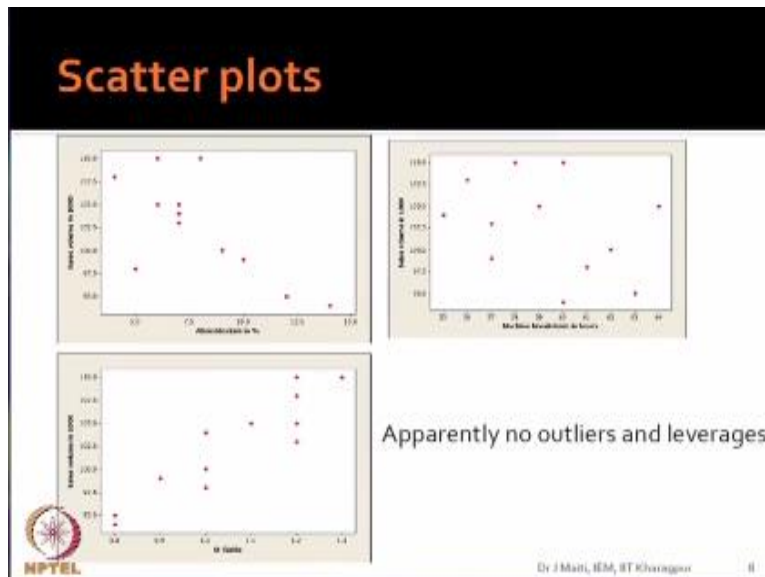
Dr J Malvi, IEM, BT Kharagpur

7

Parameter tests.

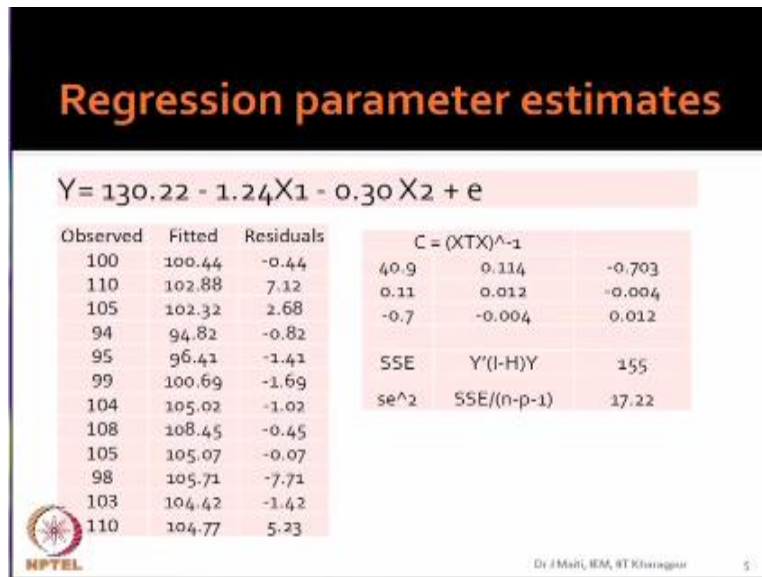


(Refer Slide Time: 07:03)



Now, see the outliers or leverages based on scatter plot it is visible that is there any outliers difficult I think it is even. This first figure it is, it is not clear that outlier is there or not or residual what I can say influencing our observations, second one you see that sales volume versus bishop is machine break down in hours here, what happen it is almost random no relationship and third one M ratio which we have not taken into consideration in this regression equation.

(Refer Slide Time: 07:55)




This regression equation we have not considered this M ratio we have consider  $X_1$  is the absenteeism  $X_2$  is the breakdown hours.

(Refer Slide Time: 08:06)

### Parameter test

Predictor	Coef	SE	T	P	VIF
Constant	130.22	26.43	4.93	0.001	
Absenteeism%	-1.2432	0.4480	-2.78	0.022	1.092
Machine BH	-0.2999	0.4586	-0.65	0.529	1.092

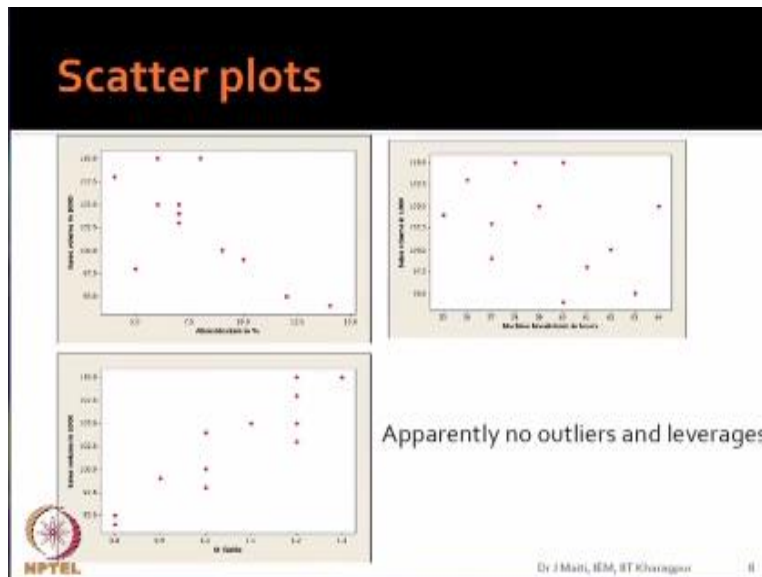


Dr J Malvi, IEM, IT Kharagpur

7

And if you see the regression coefficients, now that absenteeism has affect because P value is 0.022, but absenteeism case it is 0.53, so it is that has no effect.

(Refer Slide Time: 08:17)



And which is also rebuilt in this picture there is no effect. So, if we include M ratio what will happen, ultimately your regression fit will be better  $r^2$  will go to the higher side because here is perfect almost perfect correlation in this particular case. So, by seeing scatter plot it is not always possible to find out that whether there are outliers or leverages.

(Refer Slide Time: 08:48)

## Identification of leverages

$$H = X(X^T X)^{-1} X^T$$
$$H = \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1n} \\ h_{21} & h_{22} & \cdots & h_{2n} \\ h_{i1} & h_{i2} & \cdots & h_{in} \\ h_{n1} & h_{n2} & \cdots & h_{nn} \end{bmatrix}_{n \times n}$$


$h_{ii}, i=1, 2, \dots, n$   
measures the leverage values  
of observations  $i = 1, 2, \dots, n$

$$\sum_{i=1}^n h_{ii} = p+1$$
$$h_{ii} = (p+1)/n,$$

if each obs contributes equally

$$\frac{\left(h_{ii} - \frac{1}{n}\right)/p}{(1-h_{ii})/n-p-1} \text{ follows } F_{p, n-p-1}$$
$$F_{p, n-p-1}^{\alpha=0.05} < 2$$

**So, cut off for leverage point  $> 2(p+1)/n$**



NPTI, IEM, ST Kharagpur

So, in order to identify leverage points you have to understand the hat matrix okay so, I think we have described hat last class.

(Refer Slide Time: 09:01)

The slide is titled "Model Diagnostics of MLR". It shows the formula  $H = X(X^T X)^{-1} X^T$  with an arrow pointing to "X-space". Below the formula, there is a matrix representation of H. The matrix is square with diagonal elements  $h_{11}, h_{22}, \dots, h_{nn}$  and off-diagonal elements. A circle highlights one of the diagonal elements  $h_{ii}$ , with a note "Cut-off value" pointing to it. To the right of the matrix, there is a vertical list of indices  $1, 2, \dots, n$  corresponding to the rows of the matrix.

Not last, but one this is the hat matrix see ultimately this one all related to x space, so when you are talking about leverages it is related to the space created by the x matrix and this one you have already seen that this is basically  $h_{11}$  to  $h_{11}, 12$  to  $1n, h_{21}, 22, 2n$  like this I think  $h_{n1}, h_{n2}$  to  $h_{nn}$  and there will be somewhere  $h_{ii}$  okay so, leverage values are the diagonal elements, so in the head matrix for we have  $i$  of  $i = 1$  to  $n$  observations and you find out the  $h_{ii}$  values these are known as leverage values.

So,  $h_{11}, h_{22}$  like  $h_{nn}$  they are all leverage values okay now what will be the value of  $h_{ii}$  that  $h_{ii}$  value what will be the cut off value for  $h_{ii}$  that means when we say that the observation is influential or it is basically leverage points. So, there must be a cut off value, now if you see this.

(Refer Slide Time: 11:03)

## Identification of leverages

$$H = X(X^T X)^{-1} X^T$$

$$H = \begin{bmatrix} h_{11} & h_{12} & \dots & h_{1n} \\ h_{21} & h_{22} & \dots & h_{2n} \\ h_{i1} & h_{i2} & \dots & h_{in} \\ h_{n1} & h_{n2} & \dots & h_{nn} \end{bmatrix}_{n \times n}$$

$$\frac{\left(h_{ii} - \frac{1}{n}\right)/p}{(1-h_{ii})/n-p-1}$$


$h_{ii}, i=1, 2, \dots, n$   
measures the leverage values of observations  $i = 1, 2, \dots, n$

$\sum_{i=1}^n h_{ii} = p+1$

$h_{ii} = (p+1)/n$ ,  
if each obsv contributes equally

follows  $F_{p, n-p-1}$   $F_{p, n-p-1}^{\alpha=0.05} < 2$

**So, cut off for leverage point  $> 2(p+1)/n$**

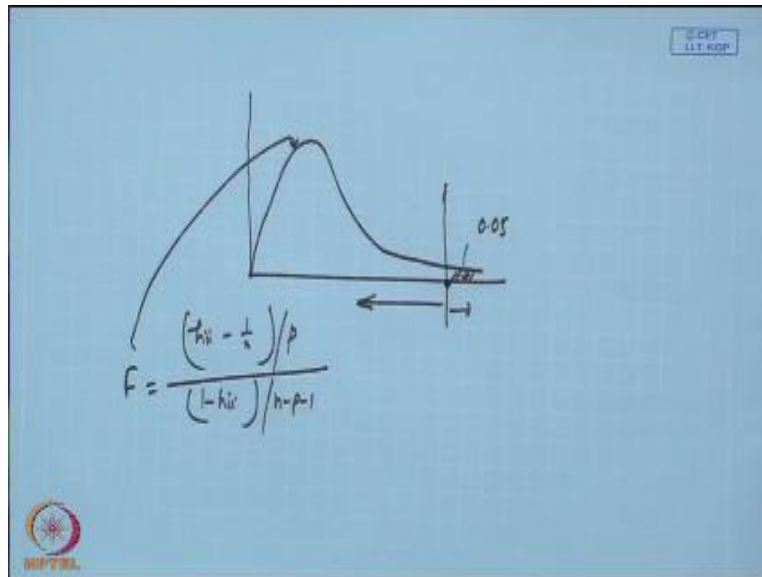


ANNA UNIVERSITY, CHENNAI

Distribution of the  $h_{ii}$  you will find out that  $h_{ii} - 1/p$  and  $1 - h_{ii}$  by  $n - p - 1$  this quantity follows  $F$  distribution with  $p$  and  $n - p - 1$  degrees of freedom. Now, if your  $p > 10$  and  $n - p - 1 > 50$  means what we are saying if you take large observations as well as your  $p$  is little more and for  $\alpha = 0.05$ . This  $F$  value is always less than equal to 2 this is we are talking about that when we talk about the multiple regression large number of variables large number observation this is the practical case. So, if this is the case then all values will be irrespective of the  $p$  and  $n - p - 1$  when this condition satisfies.

So it will be less than equal to 2. So that what we mean by this when say that whether it is influential or not.

(Refer Slide Time: 12:19)



That mean you are considering chi square distribution and you will be considering this region that is 0.05. So, you are saying it is within this side then it is not influential when it goes to this side this is influential observation, so that mean we want to find out  $h_{ii}$  value for this point okay and it is it is given that that  $h_{ii} - 1/n$  divided by degree of freedom and  $1 - h_{ii}$  divided by this is, this is suppose this is F this one you are finding out here. So, we want to know this and there are there are several cut offs given.



(Refer Slide Time: 13:18)

## Identification of leverages

$$H = X(X^T X)^{-1} X^T$$

$$H = \begin{bmatrix} h_{11} & h_{12} & \dots & h_{1n} \\ h_{21} & h_{22} & \dots & h_{2n} \\ h_{i1} & h_{i2} & \dots & h_{in} \\ h_{n1} & h_{n2} & \dots & h_{nn} \end{bmatrix}_{n \times n}$$

$$\frac{\left( h_{ii} - \frac{1}{n} \right) / p}{(1 - h_{ii}) / (n - p - 1)}$$


$h_{ii}, i = 1, 2, \dots, n$   
measures the leverage values of observations  $i = 1, 2, \dots, n$

$\sum_{i=1}^n h_{ii} = p + 1$

$h_{ii} = (p + 1) / n$ ,  
if each obsv contributes equally

follows  $F_{p, n-p-1}$   $F_{p, n-p-1}^{0.05} < 2$

**So, cut off for leverage point  $> 2(p+1)/n$**

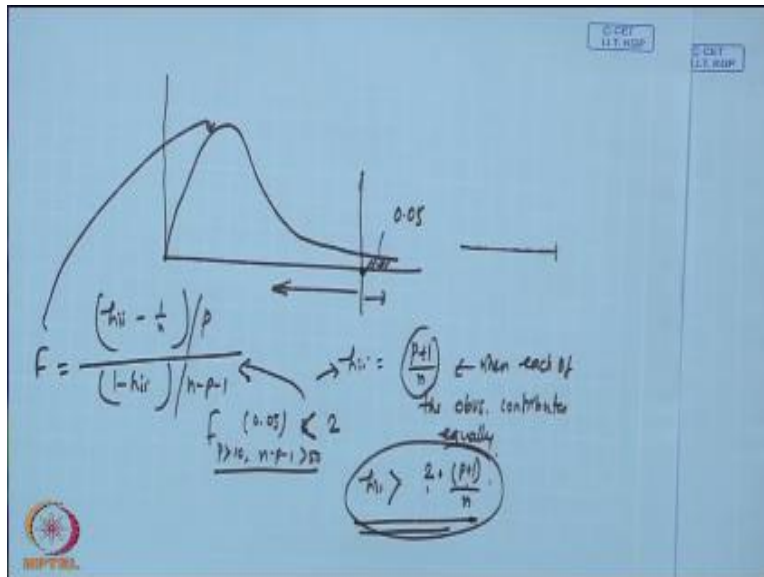


ANNA UNIVERSITY, CHENNAI

But the most widely used is that cut off for leverage point that will be greater than  $2(p + 1 / n)$ .  $(p + 1 / n)$   $p + 1$  is the number of parameter to be estimated.  $N$  is the number of parameter estimated  $n$  is the number of that is the sample size and these two is coming because we have seen that it will be less than 2 for most of the situations. So, when this condition satisfy we say this the point is a leverage point in the sense it has influence on the regression estimates, and what  $h_{ii}$  measures it measures the leverage values and the sum total of  $h_{ii}$  this will be equal to the number of parameters.

So, then if all our all points are equally influencing then what will happen  $h_{ii}$  value be equal for all the points and that value will be  $p + 1 / n$ . So, that mean what we mean to say.

(Refer Slide Time: 14:33)



If they are equally  $h_{ii} = p + 1$  by  $n$  when each of the observations contributes equally which we want also, but it is not possible. So, as a result this distribution and this from this distribution what we are seeing that this one for  $F_p > 10$   $n - p - 1 > 50$ . If we take  $\alpha$  equal to this point this will always less than 2 irrespective of any other  $p$  when this condition satisfy. So, that is why they are saying that if you multiply this by 2, so what will happen  $h_{ii}$  this greater than you are multiplying by 2 into  $p + 1 / n$ , this is the average value this one.

So, from average how much you are going this side depending on this value that is why 2 is multiplied here, so if your value  $h_i$  value, any  $h_i$  value which is more than  $2(p + 1) / n$  that is leverage value okay now, see this for our case, our case.

(Refer Slide Time: 16:01)


## Identification of leverages

$h_{ii}$	$\frac{\left(h_{ii} - \frac{1}{n}\right)/p}{(1-h_{ii})/n-p-1}$	follows $F_{p, n-p-1}$	$F_{p, n-p-1}^{\alpha=0.05} < 2$
0.164531			
0.106083			
0.391255			
0.495383			
0.340328			
0.235316			
0.296724			
0.309304			
0.123412			
0.251441			
0.145325			
0.140896			

**So, cut off for leverage point  $> 2(p+1)/n$**

Cut off =  $2*(2+1)/12 = 0.50$

**Conclusions: No leverage points**



ANNA UNIVERSITY, CHENNAI

You see that  $h_{ii}$  values are observation 1 to observation 12 and  $h_{ii}$  values are given these are all the diagonal values of the head matrix okay now, what will be the cut off value, cut off value will be we have 2, how many parameters we are estimating 3, what is your sample size n. So, p is  $2 + 1(2/n)$  this is 0.50 is there any value which is greater than 0.50, you see we have not got any value, here which is greater than 0.50, so we can conclude that, here is no leverage points for the problem we have undertaken.

(Refer Slide Time: 17:02)

## Identification of leverages: Cook's distance

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})^T X^T X (\hat{\beta}_{(i)} - \hat{\beta})}{ps_e^2}, i = 1, 2, \dots, n$$

$$D_i = \frac{r_i^2 h_{ii}}{p(1-h_{ii})}, i = 1, 2, \dots, n$$

$$r_i = \frac{e_i}{\sqrt{s_e^2(1-h_{ii})}}, i = 1, 2, \dots, n$$

$$D_i \sim F_{p, n-p-1}$$

COOK'S D

0.000877

0.131387

0.147671

0.025554

0.030221

0.022519

0.012249

0.002592

0.000014

**0.520644**


0.007862

0.102077

Cut off:  $D_i > 1$

$D_{10} = 0.52 < F_{2,3}(0.25) = 1.62$

Conclusions: No influential observations



ANNA UNIVERSITY, CHENNAI

Okay now, this leverage point is definitely very good it will give you the, you identify that if any observation is influential or not. But, Cook has given something different also that means you can go by  $h_{ii}$  values and the formulation what we have discussed so far that you can use cook distance also cook's distance.

(Refer Slide Time: 17:30)

Cook's distance

	C	V	X
1			
2			
⋮			
i			
⋮			
n			

$n = 100$   
 $100 - 1 = 99$

$$y = X\beta + \epsilon \quad (n)$$

$$g = X\hat{\beta} \quad (n)$$

$$\hat{y}_i = X_i \hat{\beta}_i \quad (n_i)$$

$$D_i = \frac{(\hat{\beta} - \hat{\beta}_i)^T X^T X (\hat{\beta} - \hat{\beta}_i)}{p \cdot h_i}$$

$\propto F_{p, n-p-1}$

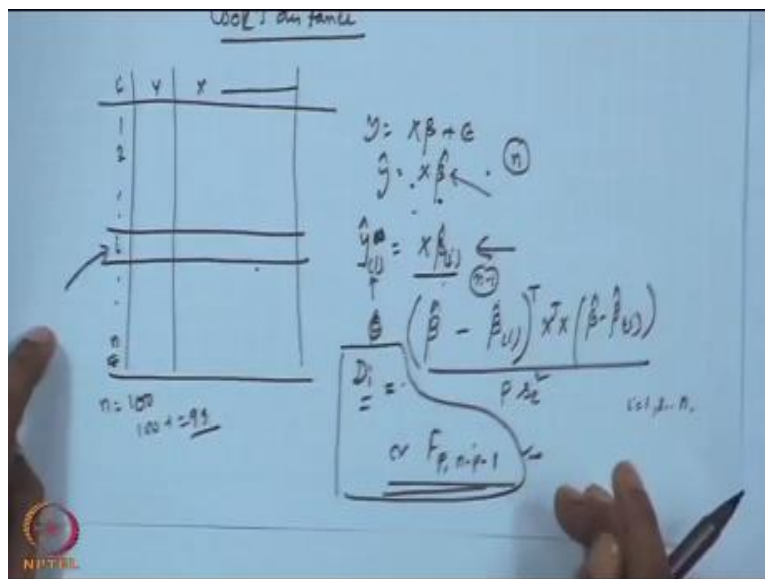
What is the procedure in cook's distance the procedure is like this, so  $i = 1, 2, i \dots, n$ , so  $n$  observation are there  $y$  values are there and  $x$  values are there fine. You have used the regression equation like this  $x\beta + \epsilon$  and using all that observations you have computed  $y^\wedge = x\beta^\wedge$ . Now, what is our interest, our interest is we want to know is the  $i$ -<sup>th</sup> observation is influencing the regression estimate or not.

Now if  $i$  is might belong the general mass then what will happen as  $n$  is quite large. If you eliminate one observation there will not be almost no difference in the  $\beta$  estimate because if I take  $n = 100$  or  $100 - 1$  that is  $99$ . Then this estimate should not be distorted to the general mass, general mass if that point does not belong to the general mass that means it is a leverage point with respect to  $x$  definitely we are talking about  $x$  space then what will happen it will affect the  $\beta$  estimate so now what you will do you go for another regression without the  $i$ -th observation getting me so if I say this one suppose if I write this one the  $i$ -th observation is not there  $y_i$  cap this is  $x\beta$  I can write here  $i$  let it be here only  $\beta_i$  within bracket let us give like this that is better parity will be there so what is the second equation.

Second equation is, here you have taken n data points here you have taken n - 1 data point the  $i^{\text{th}}$  the 1 this i the  $i^{\text{th}}$  1 is eliminated then what we are saying this should not the difference between this  $\beta_{\text{cap } n} - \beta_{\text{cap } i}$  this difference should not be much really it should that ineffectively there should not be any difference only rounding error some difference will be there then Cook has created one statistics the  $D_i$  which is  $\beta_{\text{cap } n} - \beta_{\text{cap } i}$   $X^T$  then  $\beta_{\text{cap } n} - \beta_{\text{cap } i}$  this divided by  $p$  into  $S^2$  and definitely  $i$  equal to 1 to  $n$  so he created one statistics this type of statistics so what happen you eliminate the  $i^{\text{th}}$  observation do the second round in regression modeling find out the  $\beta$  values.

You have several x values these are all matrix vector values or matrix of the order  $p \times p + 1$  cross 1 and then you create this type of statistics that  $D_i$  equal to this one follows f distribution with  $p$   $n - p - 1$  degrees of freedom sir 1 - when we will eliminate the  $i^{\text{th}}$  observation means x as well as y that is total observation so that time this x matrix 1 we consider from that matrix also we have to eliminate that total that xy total as you said the  $i^{\text{th}}$  observation including x and y you eliminate.

(Refer Slide Time: 22:02)



Now this quantity this quantity follows F distribution now when what will be your say that this  $D_i$  what we are trying to say this will become as close as possible so, we will be looking for this

Di value as small as possible then we will say that it is not away from the general mass and fine so as a result what happened using this now can you not find out that what will be the influential observation it all depends on that where you want to put the cut off value depending on the F distribution we will be able to do.

(Refer Slide Time: 22:53)

The image shows a whiteboard with handwritten mathematical notes. At the top left, it says 'n = 1000' with a checkmark. To the right, it says  $\beta(i)$  and  $(i=1, 2, \dots, n)$ . Below this, there is a diagram of a vertical line with a horizontal line intersecting it, and another horizontal line below. To the right of the diagram, the following equations are written:

$$D_i = \frac{r_i \cdot h_{ii}}{p \cdot (1 - h_{ii})}$$

$$r_i = \frac{e_i}{\sqrt{S^2 (1 - h_{ii})}}$$

On the right side, there is a vertical line with 'SSE' written next to it, and 'n-p-1' written below it, indicating the degrees of freedom for the error sum of squares.

Now question is there are suppose 1000 data points so should I go for that 1000 + 1 when every time you are eliminating 1 the first one second one like this so many observations so you will be having so many regression fittings regression equation you have to develop you have to calculate  $\beta$  that  $\beta_i$ ,  $i$  equal to 1 to  $n$  but it is not like this you do not require this several times there is the way out is that  $D_i$  is  $R_i$  by  $p \cdot h_{ii}$  by  $1 - h_{ii}$  so  $h_{ii}$  is this these are the basically the diagonal elements of the head matrix this is known  $p$  is known then what is  $r_i$  is basically  $e_i / \sqrt{S^2 (1 - h_{ii})}$   $e_i$  is the basically the error one error one.

So  $i$  given that like this epsilon I cap you know  $S^2$  is SSE by  $n - p - 1$  so that mean when you are fitting 1 degrees in equation you are getting everything now put this value  $r_i$  value here and find out this value and then you say whether it is what I can say what is the distance you measure the

distance and using F distribution you find out now the cut off value what it says that the cut off value.

(Refer Slide Time: 24:44)

### Identification of leverages: Cook's distance

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})^T X^T X (\hat{\beta}_{(i)} - \hat{\beta})}{ps_e^2}, i = 1, 2, \dots, n$$

$$D_i = \frac{r_i}{p} \frac{h_{ii}}{1-h_{ii}}, i = 1, 2, \dots, n$$

$$r_i = \frac{e_i}{\sqrt{s_e^2(1-h_{ii})}}, i = 1, 2, \dots, n$$


$$D_i \sim F_{p,n-p-1}$$

$D_{10} = 0.52 < F_{2,9}(0.25) = 1.62$

Cut off:  $D_i > 1$

COOK'S D
0.000877
0.131387
0.147671
0.025554
0.030221
0.022519
0.012249
0.002592
0.000014
<b>0.520644</b>
0.007862
0.102077

**Conclusions: No influential observations**

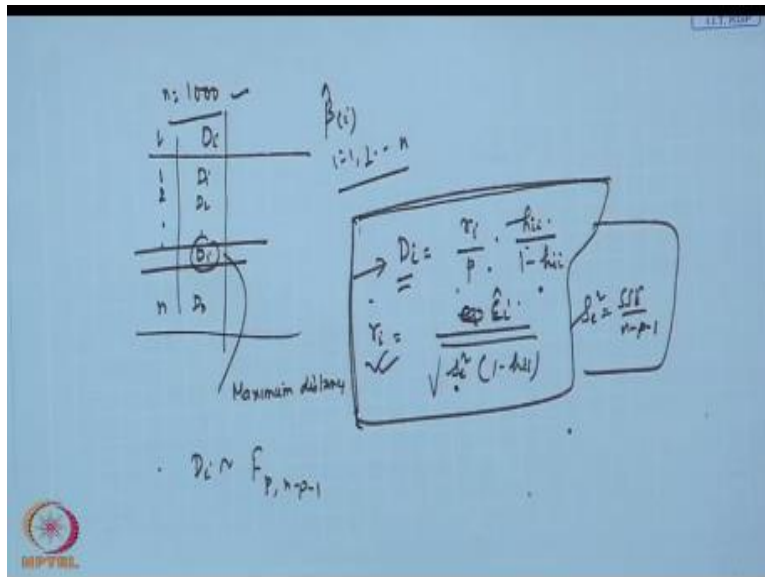


NPTEL | IIT Kharagpur

Is given that if  $D_i$  greater than 1 then it is basically significant this is Cook's, Cook has given this that for  $D_i$  greater than 1 the observation having this that will be significant now we will say what is the procedure is first you find out the Cook's distance using the formulation that formulation will be this one you will be using this use this find out this Cook's distance.

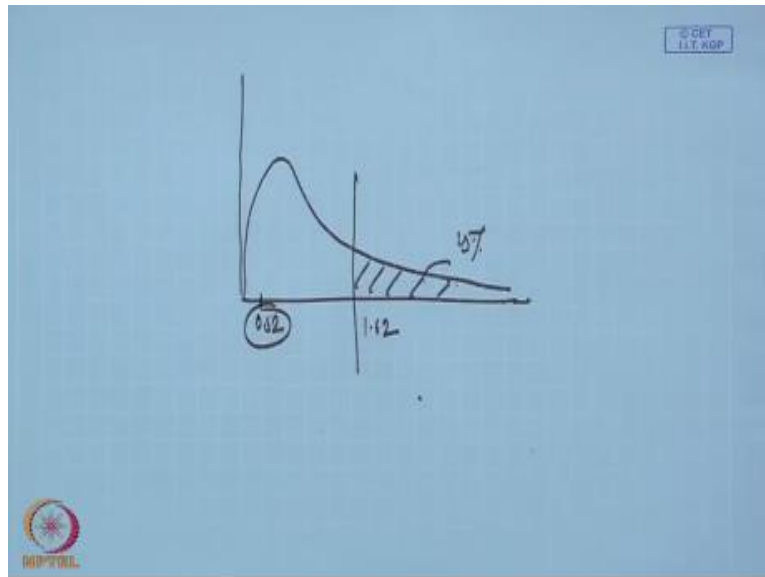


(Refer Slide Time: 25:32)



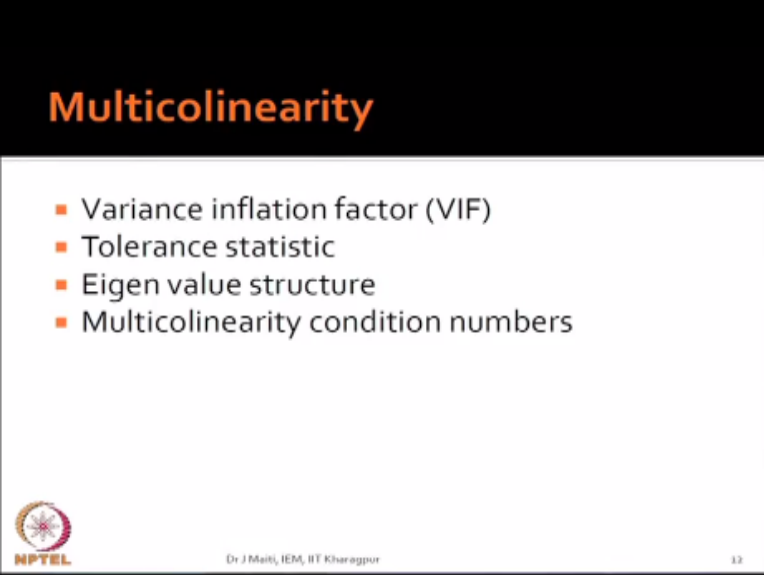
Yes not coming yeah so using this you find the Cook's distance so ultimately 1 to n then  $D_i$  value you are getting so if I say  $D_1, D_2$  like this  $D_n$  then you find out the maximum one which one is maximum let the  $D_i$  this one is the maximum distance correct so for this maximum distance what we say that  $D_i$  follow  $F_{p, n-p-1}$  so I will take the maximum distance and then what is  $p$  value in our case in our case  $p$  is 2-  $p-1$  is 9 and then I have taken 0.25 not 0.05 even when I have taken 0.25 this value is 1.62 that but our  $D_{10}$  value is 0.52 only so it is much.

(Refer Slide Time: 26:42)




Close is if I see the F distribution table sorry graph like this we have taken 25% this is 25 % t and this value is 1.62 but your maximum value here it is 0.52 so it is not at all a influential point.

(Refer Slide Time: 27:17)



**Multicollinearity**

- Variance inflation factor (VIF)
- Tolerance statistic
- Eigen value structure
- Multicollinearity condition numbers

 Dr. J. Maiti, IEM, IIT Kharagpur 33

Then we will go for multicollinearity now multicollinearity as I told you multicollinearity is a is an issue where independent variables are not truly independent there is there is dependence structure amongst the independent variables under such condition what will happen if there is linear case linear dependence case the determinant of this  $X^T X$  will you will not get it will become 0 and ultimately inverse you cannot create and you will not get the estimate values so multicollinearity has to be tested and multicollinearity can be tested through four different procedures.

And these are known as variation inflation factor tolerance statistic Eigen value structure and multicollinearity condition number so what we will discuss we will first discuss the variance inflation factor what is variance inflation factor variance inflation factor.

(Refer Slide Time: 28:28)

## Multicollinearity: VIF & Tolerance

Diagram illustrating the regression model for variable  $X_j$ . Predictors  $X_1, X_2, \dots, X_p$  and the constant  $X_0 = 1$  influence  $X_j$  through coefficients  $\beta_1, \beta_2, \dots, \beta_p, \beta_0$ . The error term is  $\epsilon_j$ .

$$VIF = \frac{1}{1 - R_j^2}$$

Predictor	$\beta$	VIF
Absenteeism %	-1.2432	1.092
Machine BH	-0.2999	1.092

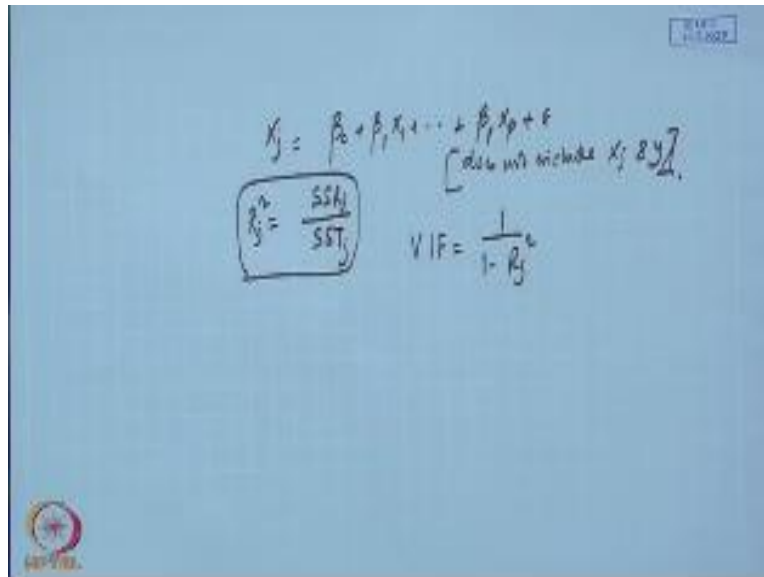
$R_j^2$	0	0.2	0.4	0.5	0.6	0.8	0.9	1.0
VIF	1	1.25	1.67	2	2.5	5	10	$\infty$
1/VIF	1	0.8	0.6	0.5	0.4	0.2	0.1	0.0

Dr J Mehta, IEM, BT Kharagpur

33

Is something suppose we are talking about that out of this  $p$  independent variables there is correlated structure in the sense dependence relationship so arbitrarily we are taking one independent variable as dependent variable we are not considering here we are considering only the independent variables then we are taking one of the independent variable as dependent variable and all other independent variables as independent as influencing that independent variable so  $X_j$  is now affected by  $X_1, X_2, \dots, X_p$  then you are making a regression equation so your regression equation is now.

(Refer Slide Time: 29:26)



$X_j$  is  $\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \epsilon$  so like this  $\beta_p X_p + \epsilon$  this does not include  $X_j$  and  $y$  then you find out the  $R_j^2$  so that will be  $\frac{SSR_j}{SST_j}$  that is for the  $j^{\text{th}}$  variable so then you create variance inflation factor equal to  $\frac{1}{1 - R_j^2}$  for example in our case absenteeism and machine breakdown.

(Refer Slide Time: 30:30)

### Multicollinearity: VIF & Tolerance

$$VIF = \frac{1}{1 - R_j^2}$$

Predictor	$\beta$	VIF
Absenteeism %	-1.2432	1.092
Machine BH	-0.2999	1.092

$R_j^2$	0	0.2	0.4	0.5	0.6	0.8	0.9	1.0
VIF	1	1.25	1.67	2	2.5	5	10	$\infty$
1/VIF	1	0.8	0.6	0.5	0.4	0.2	0.1	0.0

Dr J Maiti, IEM, BT Kharagpur
33

The  $\beta$  values are this and variance inflation factor is 1.092 both cases 1.092 if  $R_j$  is 0 if  $R_j$  is 0 then VIF will be equal to 1 and we do not want  $R_j$  apart from value apart from 0 value.

Mean we want that 0 value that is the best value because  $R_j = 0$  means no correlation means no regression is not valid regression that means, that means  $X_j$  is not dependent on the other independent variable so you have to create this type of variance inflation factors for each of the variables then you see this here what happen your  $R^2$   $R_j^2$  this  $R_j^2 = 0$  mean VIF 1 if it is 0.2 1.2 like this then there is another concept called tolerance is nothing but just reverse tolerance is 1 by variance inflation factor.

So if you use tolerance or variance inflation factor both are same ultimately here what is happening you are getting within a 0 to 1 scale 0 to 1 scale here any value is possible so as if we get in terms of 0 to 1 scale it is easier for us to interpret so now then what will be the VIF value that should be considered you are getting me for basically we say that if the VIF value is 10 or more this mean high collinearity high relationship so 10 or more 10 is the cutoff value it should not be 10 or more 5 also 5 is the warning limit you can think of so that means if tolerance is 0.1 or less or variance inflation factor 10 or more that is not desirable but when if it is 5 and then it is warning case.

(Refer Slide Time: 33:04)


## Multicolinearity: Eigen-value & MCN

$$R = \sum_{j=1}^p v_j \lambda_j v_j^T$$

One or more  $\lambda_j$  values equal or close to zero indicate multicollinearity

$$MCN = \frac{\lambda_1}{\lambda_p}$$
$$VIF_m \leq MCN \leq p \sum_{j=1}^p VIF_j$$

$MCN < 100$ : Not serious  
 $MCN > 1000$ : Very serious



Dr J Mohi, IEM, IIT Kharagpur

14

Then another issue is that another is the Eigen value criteria what is this Eigen value criteria in Eigen value criteria so all of you know that the correlation matrix we are talking about correlation matrix.

(Refer Slide Time: 33:22)

$R =$  Correlation matrix of  $X_{n \times p}$ .  

$$= \begin{bmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{12} & 1 & & \\ \vdots & & \ddots & \\ r_{1p} & & & 1 \end{bmatrix}_{p \times p}$$

Using spectral decomposition:  

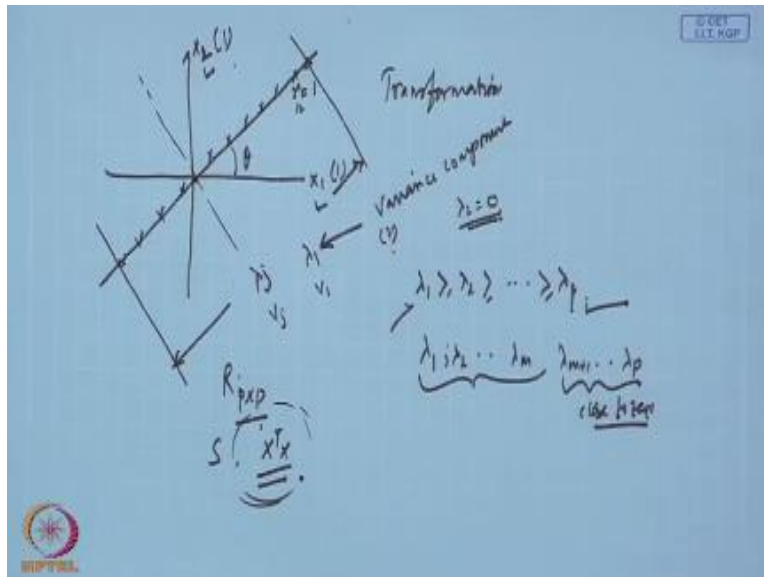
$$R = \sum_{j=1}^p v_j \lambda_j v_j^T$$

$\lambda_j =$   $j$ -th eigenvalue  
 $v_j =$   $j$ -th eigen vector

Of  $x \ n \ x \ p$  which will be so this is  $p$  cross  $p$  matrix now using spectral decomposition this  $r$  can be written like this that I can write that  $v_j \lambda_j v_j^T$  equal to  $1$  top where  $\lambda_j$  is the  $j^{\text{th}}$  Eigen value and  $v_j$  is the  $j^{\text{th}}$  Eigen vector so you can any this  $p$  cross  $p$  matrix, this matrix can be decomposed its Eigen value and Eigen vector components this can be that mean if I know Eigen values and Eigen vector I can re constructor because this one is  $p \times 1$  this one is  $1 \times 1$  this one is  $1 \times p$  so if you multiply this two ultimately  $p \times p$  matrix you will be able to recreate now what is the meaning of this Eigen value here when you do the spectral decomposition Eigen value here.



(Refer Slide Time: 35:20)



This is something like this suppose you consider two variable case suppose  $X_1$   $X_2$   $X_1$  if they are dependent you may get a structure like this for the perfect dependent case will be like this so here  $r_1$   $r_2$  equal to 1 so what we mean to say here that we do not require  $X_1$  and  $X_1$  to measure if we transform the axis by certain degree this theta degree then what will happen you will get another dimension which will capture the totality of the data given here now so that means if I can do some manipulation here transformation so you rotate this  $X_1$  and  $X_2$  by  $\theta$  you are coming to this place.

And here this axis is having the variability from here to here this variability is captured by  $\lambda$  and the direction is captured by  $v$  so what we mean to say we are trying to say here that only 1 dimension is required if the structure is like this only 1 dimension is required to measure this and that one will be  $\lambda_1$  and then  $v_1$ . So,  $\lambda_1$   $v_1$  is sufficient enough to capture this data, because if I go in along this line my variability, here is this, but what is my variability along perpendicular to this line 0. So,  $\lambda$  represents the variance component, so if my structure is like this then  $\lambda_1$  and I have taken two variables, which in the standardized case suppose this 1 and this is 1.

So, then both the variability 1, 1 is captured by this, so this will become 2 because the total variability is 2 here for the two variables, so other dimension there will be 0, so what will happen  $\lambda_2$  will become 0. Then the two variable situations when you decompose the R matrix into its Eigen value, Eigen vector and if you find out that one of them is 0  $\lambda$  value is 0, then it is a perfect correlation case. So, similarly if there are p such variables, so you will be getting  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$  this way you extract.

So, the first component will have the maximum followed by like this, so it may so happen that when the R, is basically if it is it is basically  $p \times p$ . So, we assume that R or S or  $x^T x$  when we do regression we assume that  $x^T x$  is full rank, now what will happen if you find out that out of this p  $\lambda$  Eigen values. Suppose  $\lambda_1, \lambda_2, \dots, \lambda_n$  these are basically having values not equal to 0, but  $m+1$  to  $m+p$  these are close to 0.

So, what will happen in that case, basically rank deficient that means this is not full rank and this 0 are representing that there are large number of variance  $m+1$  to  $p$  that large number of independent variable, so called independent they are not independent. So, there is multicollinearity this is what is known, is known an Eigen value criteria, so you take the R that is the correlation matrix go for your spectral decomposition that means Eigen value, Eigen vector decomposition then finds out the Eigen values if you find out that some of the Eigen values are close to 0, it simply indicates that your case is not independent, the independent variables are not truly independent, okay.

(Refer Slide Time: 40:22)

## Multicolinearity: Eigen-value & MCN


$$R = \sum_{j=1}^p v_j \lambda_j v_j^T$$

One or more  $\lambda_j$  values equal or close to zero indicate multicollinearity

$$MCN = \frac{\lambda_1}{\lambda_p}$$

$VIF_m \leq MCN \leq p \sum_{j=1}^p VIF_j$

$MCN < 100$ : Not serious  
 $MCN > 1000$ : Very serious



Dr. J. Mehta, IEM, IIT Kharagpur

14

Then there is a multicollinearity number this multicollinearity number is known as MCN.

(Refer Slide Time: 40:29)


## Multicolinearity: Eigen-value & MCN

$$R = \sum_{j=1}^p v_j \lambda_j v_j^T$$

One or more  $\lambda_j$  values equal or close to zero indicate multicollinearity

$$MCN = \frac{\lambda_1}{\lambda_p}$$
$$VIF_m \leq MCN \leq p \sum_{j=1}^p VIF_j$$

$MCN < 100$ : Not serious  
 $MCN > 1000$ : Very serious



Dr. J. Mehi, IEM, IIT Kharagpur

24

Which is basically the largest Eigen value divided by the smallest one, and now, if there are many values close to 0 then definitely this  $\lambda_p$  this one is very close to 0 and it will be very high value MCN will be very high value, so if MCN greater less than 100 it is not a serious multicollinearity problem not serious. But, if it is greater than 1000, it is a serious issue, getting me then there is one relationship mean from the MCN and VIF point of view that VI variance inflation factor. It is M less than MCN less than sum total of, this  $VIF_m$  is the maximum variance inflation factor, so less than this less than p into j equal to all the sum of all the variance inflation factor.

(Refer Slide Time: 41:58)

## An example

Sl. No.	Months	Profit in Rs million	Sales volume in 1000	Absentees m in %	Machine breakdown in hours	M-Ratio
1	April	10	100	9	62	1
2	May	12	110	8	58	1.3
3	June	11	105	7	64	1.2
4	July	9	94	14	60	0.8
5	Aug	9	95	12	63	0.8
6	Sep	10	99	10	57	0.9
7	Oct	11	104	7	55	1
8	Nov	12	108	4	56	1.2
9	Dec	11	105	6	59	1.1
10	Jan	10	98	5	61	1.0
11	Feb	11	103	7	57	1.2
	March	12	110	6	60	1.2



Now, can you not find out the data for data whatever we have given, here we have seen here, this one if it is asked to you that you find out that whether multicollinearity problem is there or not, so what way you proceed, what will be your case.

(Refer Slide Time: 42:20)

$$X_{3 \times 2} = \begin{matrix} & X_1 & X_2 \\ \begin{bmatrix} 9 & 62 \\ 8 & 58 \\ 7 & 64 \end{bmatrix} \end{matrix}$$

$$R = \frac{1}{n} [X^T X]$$

$$R = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$$

Find out eigenvalue

$$\begin{vmatrix} 1-\lambda & 0.8 \\ 0.8 & 1-\lambda \end{vmatrix} = 0$$

For example 9, 62, 8, 58 and then 7, 64 suppose these three data points are given to you, so this is my X this is 3x2, this is  $X_1$  and  $X_2$  what is our aim we want to test the  $X_1$ ,  $X_2$  that multicollinearity issues are there are not, one is that you find out the regression. You regress  $X_2$  on  $X_1$  since there are only two variable  $X_1$  and  $X_2$  is enough and then find out the VIF, other one is I said that can you not find out the R value. How do you compute R here, what you require to do, I told you in early multivariate descriptive statistics class.

I told you first find out R, R I told you that  $x \sim^T x \sim 1/n-1$  and your  $x \sim$  is suppose if I say  $x \sim$  is something like this, yes. So, suppose  $x \sim_{ij}$  is there then  $x \sim_{ij}$  tilde will be  $x_{ij} - \bar{x}_j$  by standard deviation of this, so we require to get  $R^2$  value R value first here, once you get R value suppose for example let R value is I am giving some arbitrary value here, suppose R value is this 1 and this is 1 and let us take 0.8 here, although it is not like this 0.8, so you got the R value, so what is required to do you require to find out the Eigen values, how do you find out Eigen values. Normal processes that characteristic root. Correct, so you how to first consider that  $1-\lambda$  0.8,  $1-\lambda$  this determinant this equal to 0.

(Refer Slide Time: 45:07)

The image shows a whiteboard with handwritten mathematical work. At the top, the equation  $(1-\lambda)^2 - 0.8^2 = 0$  is written. Below it, two equivalent forms of the quadratic equation are shown:  $1 + \lambda^2 - 2\lambda - 0.64 = 0$  and  $\lambda^2 - 2\lambda + 0.36 = 0$ . The text "two roots" is written between the equations. To the right, there are some calculations:  $12 = 144$ ,  $14 \rightarrow 196$ , and  $2 \times 2 = 4$ . The quadratic formula is applied to find  $\lambda = \frac{+2 \pm \sqrt{4 - 2 \cdot 0.36 \cdot 1}}{2 \cdot 1}$ . This is simplified to  $\frac{+2 \pm \sqrt{4 - 0.72}}{2}$ , then to  $\frac{2 \pm \sqrt{3.28}}{2} \approx \frac{2 \pm 1.8}{2}$ . An arrow points to the final results:  $\frac{3.8}{2} = 1.9$  and  $\frac{0.2}{2} = 0.1$ . An NPTEL logo is visible in the bottom left corner of the whiteboard image.

Let us see this what will happen, here will we get  $(1-\lambda)^2 - 8^2=0$ , so  $1+\lambda^2-2\lambda-0.64$  that is okay, it is this equal to 0 then  $\lambda^2 -2\lambda+ 0.36=0$  then there are two roots. So,  $\lambda$  will be minus be means  $+2\pm\sqrt{4-2\cdot 0.36} /2.1$  this is the case. So, then  $+2$  plus  $\pm\sqrt{4-0.72}/2$ , so  $2\pm\sqrt{3.28}/2$ , so 3.28 what will be the square root. So, see if it is 12 then 144, if it is 14, no it will not, it will not be 12 this is 328 I think 2 point around 1 point something will it be 2, 2 into 2 is 4 it cannot be. So, it will be less than 2. For example  $2\pm$  may be it will be  $1.8/2$  then it is  $3.8/2$  this one is 1.9 another one is  $0.2/2$  0.1. So, we can we can say that it is basically it is a multicollinearity problem.

(Refer Slide Time: 47:21)

$$X_{3 \times 2} = \begin{matrix} & \begin{matrix} X_1 & X_2 \end{matrix} \\ \begin{bmatrix} 9 & 62 \\ 8 & 58 \\ 7 & 64 \end{bmatrix} \end{matrix}$$
$$R = \frac{1}{n_1} [X^T X]$$
$$\rightarrow R = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$
$$R_{ij} = \frac{x_{1j} - \bar{x}_j}{s_j}$$

eigenvalue

$$\begin{vmatrix} 1-\lambda & 0.8 \\ 0.8 & 1-\lambda \end{vmatrix} = 0$$

Because we have already taken this, essentially then what it is what is what is coming now that multicollinearity issue can also be tested using the R matrix, so if any what will be the value of this correlation coefficient and which will tell you that, yes there is multicollinearity definitely with the two variable case when 0.88 is saying that it is almost that multicollinearity is like this.



(Refer Slide Time: 47:58)

Handwritten work on a whiteboard:

$$-0.8 = 0$$

$$0.164 = 0$$

$$-2\lambda + 0.36 = 0$$

two roots.

$$\lambda = \frac{-2 \pm \sqrt{4 - 2 \cdot 0.36 \cdot 1}}{2 \cdot 1}$$

$$MCN = \frac{\lambda_1}{\lambda_2} = \frac{-2 \pm \sqrt{4 - 0.72}}{2}$$

$$= \frac{1.9}{0.1} = 19$$

$$= \frac{2 \pm \sqrt{3.28}}{2} \approx \frac{2 \pm 1.8}{2} \Rightarrow \frac{3.8}{2} = 1.9, \frac{0.2}{2} = 0.1$$

MCN < 100 ← MCN > 1000

Now, what is our MCN multicollinearity number that  $\lambda_1 / \lambda_2$ , so our  $1.9 / 0.1$  this is 190 or 19 this is 19. So, what we have seen that multicollinearity number less than 100 is not a serious issue that is what is given there. If multicollinearity number is greater than 1000, that is a serious issue. So, then what we will do, we will go by this logic as I am able to see that there is 0.8 and one of the, this one this 1.9 this one of the dimensions the variability extent is much higher, second one is much lower.

(Refer Slide Time: 49:04)

$$X = \begin{bmatrix} 9 & 62 \\ 8 & 58 \\ 7 & 64 \end{bmatrix}$$
$$R = \frac{1}{n} [X^T X]$$
$$\rightarrow R = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$
$$\text{Eigenvector eigenvalue}$$
$$\begin{vmatrix} 1-\lambda & 0.8 \\ 0.8 & 1-\lambda \end{vmatrix} = 0$$

*(Note: The handwritten notes also include  $\tilde{x}_j = \frac{x_j - \bar{x}_j}{s_j}$  and a matrix  $X_0 = \begin{bmatrix} \tilde{x}_{11} & \tilde{x}_{12} & \dots \\ \vdots & \vdots & \ddots \end{bmatrix}$ )*

So, should I go for regression or we will simply dimension we reduce the dimension and then we will go for regression what will be your issue.

(Refer Slide Time: 49:13)

$$(1-\lambda)^2 - 0.8^2 = 0$$

$$\Rightarrow 1 + \lambda^2 - 2\lambda - 0.64 = 0$$

$$\Rightarrow \lambda^2 - 2\lambda + 0.36 = 0$$

two roots.

$$\lambda = \frac{+2 \pm \sqrt{4 - 2 \cdot 0.36 \cdot 1}}{2 \cdot 1}$$

$$MCN = \frac{\lambda}{\gamma_L} = \frac{+2 \pm \sqrt{4 - 0.72}}{2}$$

$$= \frac{1.9}{1.1} = \frac{1.9}{1.1}$$

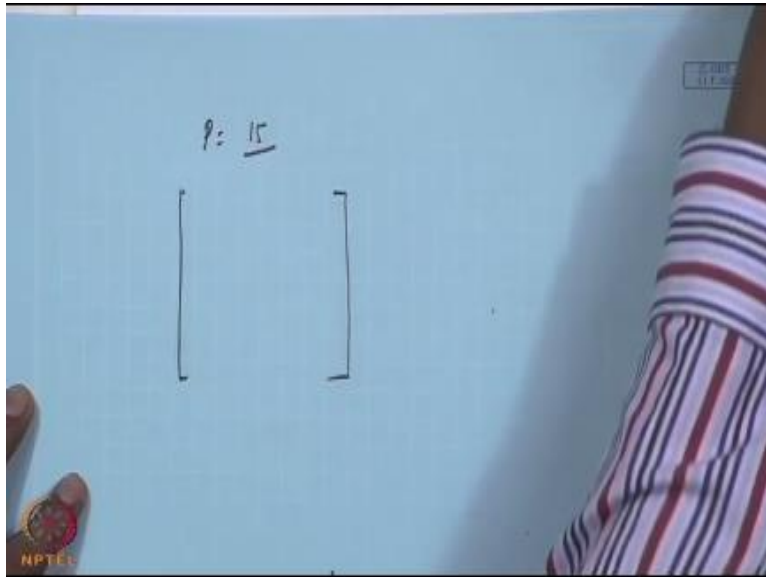
$$= \frac{2 \pm \sqrt{3.28}}{2} = \frac{2 \pm 1.8}{2} \Rightarrow \frac{3.8}{2} = 1.9, \frac{0.2}{2} = 0.1$$

MCN < 10% → MCN > 10%

$1/2 = 1/4$   
 $2 \cdot 2 = 4$   
 $0.1$

Basically, see if I go actually although 0.1 in, here we are getting 0.9, because of this two variable case I think this 0.8 where is not at all a simple issue I think we should not go by this, that is what I personally feel using that analysis, by this logic, now this 0.8 is it is a reasonable correlation coefficient. Sir, we can perform dimension reduction. This is given by dimension reduction, okay then see then what we will do suppose, here it is p for two variable case, but there are P variables.

(Refer Slide Time: 50:01)



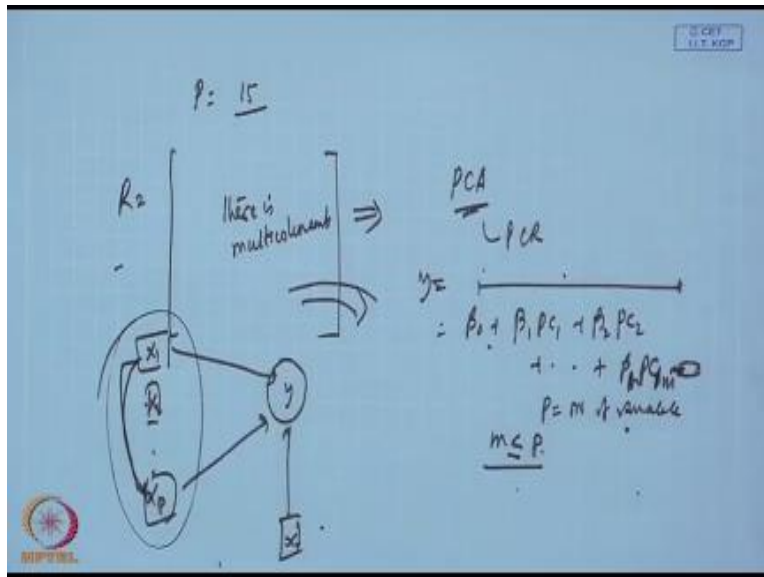
Suppose P is greater than 15 variables or 20 variables, so under this case you will be having a big correlation matrix and using this correlation matrix seeing this value you cannot judge.

(Refer Slide Time: 50:15)

The image shows a whiteboard with handwritten mathematical work. At the top, there are two quadratic equations:  $1 + \lambda^2 - 2\lambda - 0.64 = 0$  and  $\lambda^2 - 2\lambda + 0.36 = 0$ . The second equation is identified as having 'two roots'. The quadratic formula is applied to the second equation, yielding  $\lambda = \frac{2 \pm \sqrt{4 - 2 \cdot 0.36 \cdot 1}}{2 \cdot 1}$ . To the right, a small calculation shows  $12 = 144$  and  $2 \times 2 = 4$ , with a result of  $0.1$  written below. Below the main equation, the Mean Value (MCN) is calculated as  $MCN = \frac{\lambda}{2}$ . This leads to  $MCN = \frac{1.9}{2}$  and  $MCN = \frac{0.1}{2}$ . The value  $1.9$  is circled, and a note indicates  $MCN < 100$ . Another calculation shows  $\frac{3.8}{2} = 1.9$  and  $\frac{0.2}{2} = 0.1$ , with  $1.9$  circled and a note indicating  $MCN > 100$ .

Because even in the two variable case if I see 0.8 and then I discard. But, from multicollinearity number point of view, it is saying that you should go, you should go for, should go for regression without bothering for multicollinearity, so you cannot judge just by seeing the R matrix, fine.

(Refer Slide Time: 50:37)

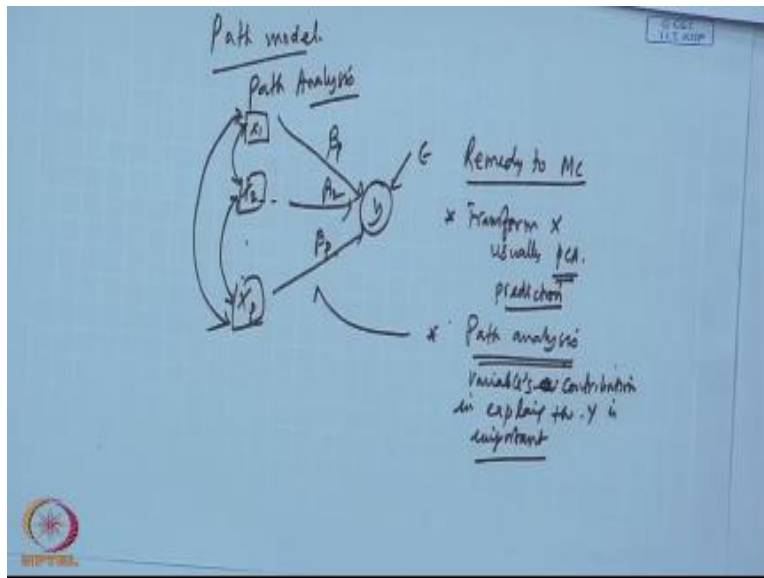


Then the solutions are what are the solutions, solution is one is the principle component analysis when there is multicollinearity. So, principle component regression you can go for, so principle component regression that mean you reduce the dimension as you are saying then find out the what are the  $\lambda$  values that are significant values and only those components you take then your  $y$  is for those components you find out the regression line the  $\beta_0$ . Suppose  $\beta_1 PC_1 + \beta_2 PC_2 + \beta_p PC_p$  and so on  $PC_p$  we will not go, we will go for  $\beta_m PC_m$ ,  $P$  is the number of variables where  $m$  definitely is less than equal to  $P$ .

Now, if you go for PCA what will be the problem, problem is that, now your original variables are missed that one, then but you may be interested, no I will not do like this. I want to keep the structure regression equation structure is like this  $x_1$  and  $x_2$  this is the structure  $x_p$   $x_0$ , now what happen they are dependent suppose this is dependent with this is dependent it is, okay. So, what do you want if you go by PCA this structure will lost this independent variable original variable will lost. So, you may be interested to that first you find out of these many variables, what are the independent variables, what are the dependent variables strict sense.

If you still find that, no these are still independent variable they cannot be treated as dependent variable.

(Refer Slide Time: 52:43)



Then you can allow them in modeling, you can allow them to co vary, in the sense I will do like this only, in regression you are doing this. But, here you can allow the covariance structure to be, getting me, so you do not go for transformation you will simply allow the covariance structure to be kept as it is. Then estimate this estimation is also possible this estimation we will be understanding through path model or path analysis. So, then remedy multicollinearity what we have discussed remedy to multicollinearity, multicollinearity one is the transform the data transform X that is usually we go for PCA.

Here, PCA is possible only if your, if your interest is prediction because you do not bother about the original data is transformed to what scale and whatever these things. But, we want to predict then fine any electrical engineering you know that some of that prediction, using that is principle component regression that is done. If you want to keep the independent variable. Sir, name of the variables same. Same and you want, you do not want to lose the nature of the variables then you go for path analysis, this path analysis what it will do it will estimate same regression parameters

this regression parameters. But, it will allow the independent variable to co vary amongst them, this co variant structure will be taken into consideration and then these two analysis it is better to go for path analysis when variable explanation is an important issue, variables contribution in explaining the y is important, so I think we can stop now.

**NPTEL Video Recording Team**

**NPTEL Web Editing Team**

**Technical Superintendents**

**Computer Technicians**

**A IIT Kharagpur Production**

**[www.nptel.iitm.ac.in](http://www.nptel.iitm.ac.in)**

**Copyrights Reserved**