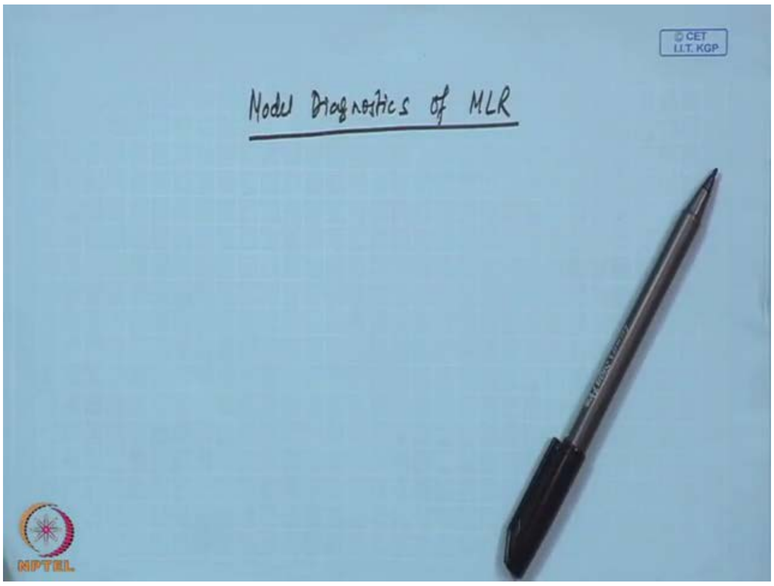
Applied Multivariate Statistical Modelling

Prof. J. Maiti
Department of Industrial Engineering and Management
Indian Institute of Technology, Kharagpur

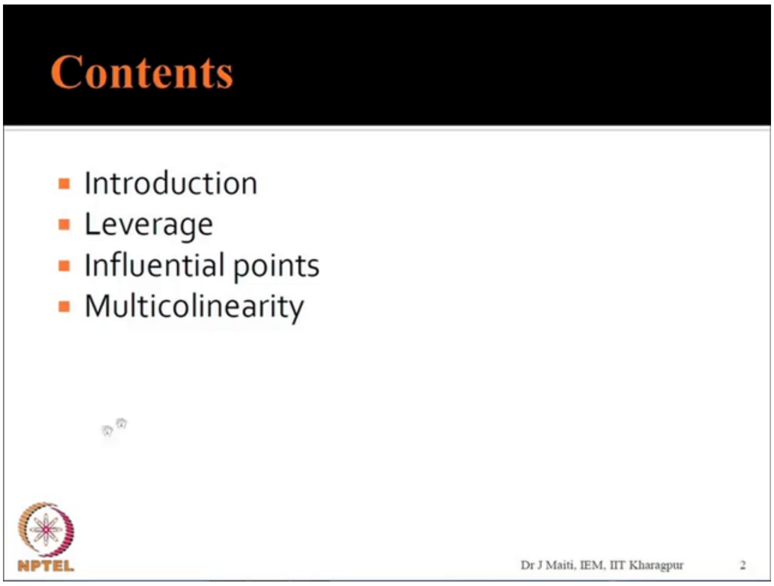
Lecture - 25
MLR - Model Diagnostics

(Refer Slide Time: 00:23)



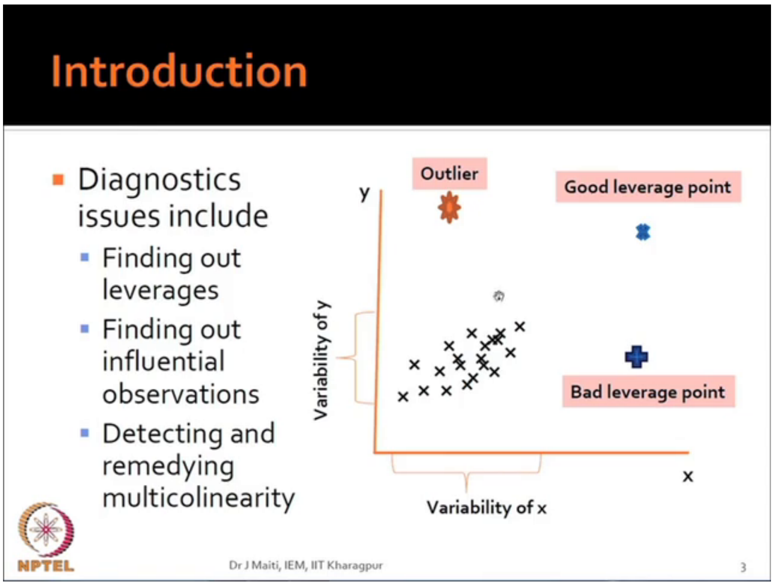
So, we will start. Now, model diagnostics of multiple linear regressions, so under model diagnostics what are the issues we will be covering.

(Refer Slide Time: 00:38)



Leverage points, influential points and multicolinearity.

(Refer Slide Time: 00:46)



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.

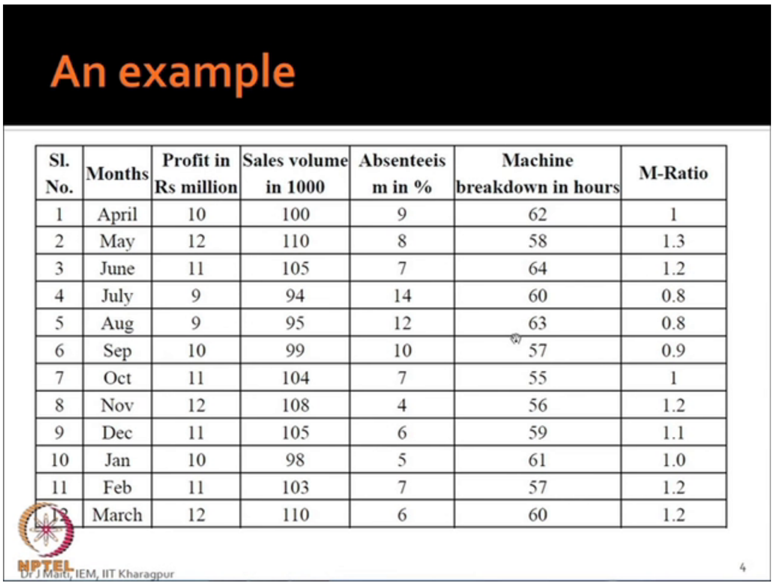
But, along x if we see this is away from the general range of the x and similarly the other one also, this one also, now leverage point is a point which lies beyond the that general mass of x. So, that means outlier is necessarily related to the y related observations and x leverage points related to 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 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 behaviour 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, 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 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 multicolinearity. So, we will find out how to identify these observations what are the remedies to high leverages or high influential points, and what is multicolinearity? And how to detect and remedy multicolinearity.

(Refer Slide Time: 06:54)



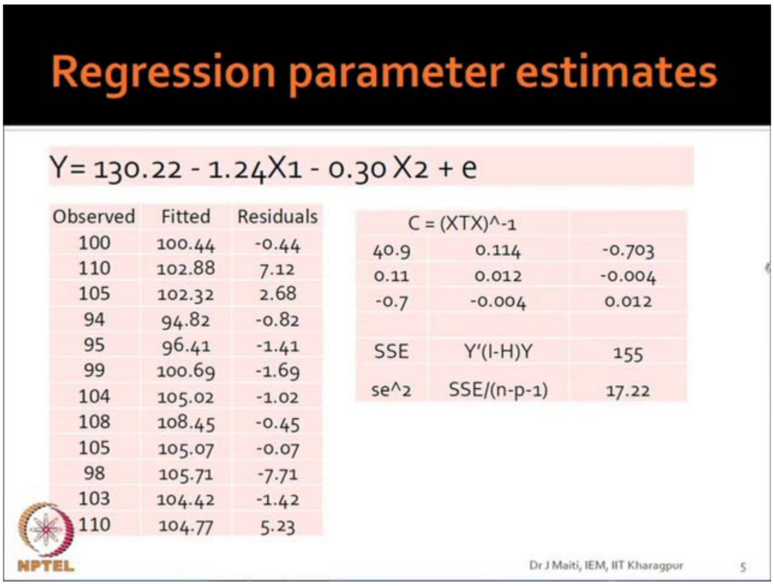
Sl. No.	Months	Profit in Rs million	Sales volume in 1000	Absenteeism in %	Machine breakdown in hours	M-Ratio
1	April	10	100	9	62	1
2	May	12	110	8	58	1.3
3	June	11	105	7	64	1.2
4	July	9	94	14	60	0.8
5	Aug	9	95	12	63	0.8
6	Sep	10	99	10	57	0.9
7	Oct	11	104	7	55	1
8	Nov	12	108	4	56	1.2
9	Dec	11	105	6	59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

This is our example and these are the fitted values and regression lines parameter tests.

(Refer Slide Time: 07:03)

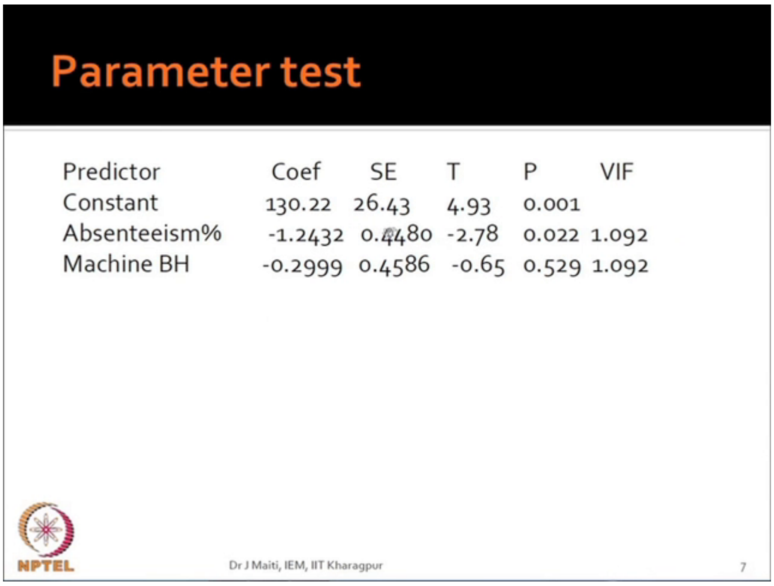
Now, see the outliers or leverages based on scatter plot it is visible that is there any outliers difficult I think it is even. This first figure it is, it is not clear that outlier is there or not or residual what I can say influence our observations, second one you see that sales volume versus bishop is machine breakdown 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: 07:52)



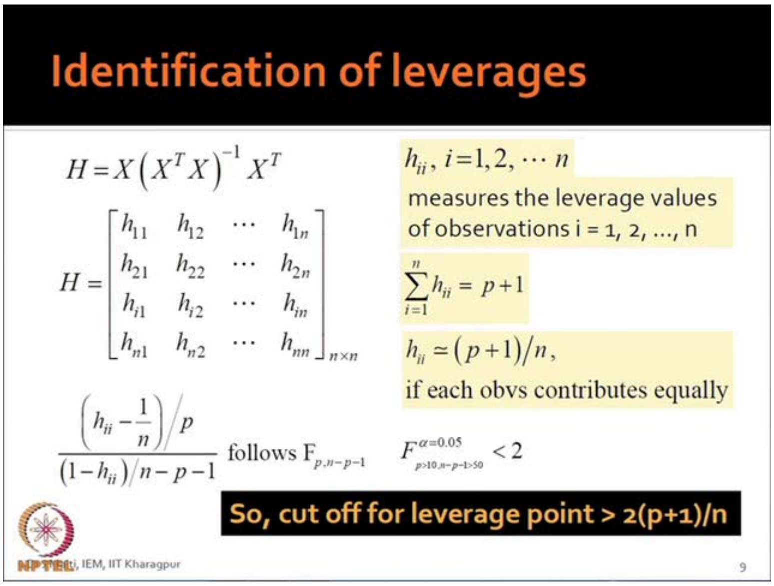
This regression equation we have not considered this M ratio we have consider X 1 is the absenteeism X 2 is the breakdown hours.

(Refer Slide Time: 08:02)



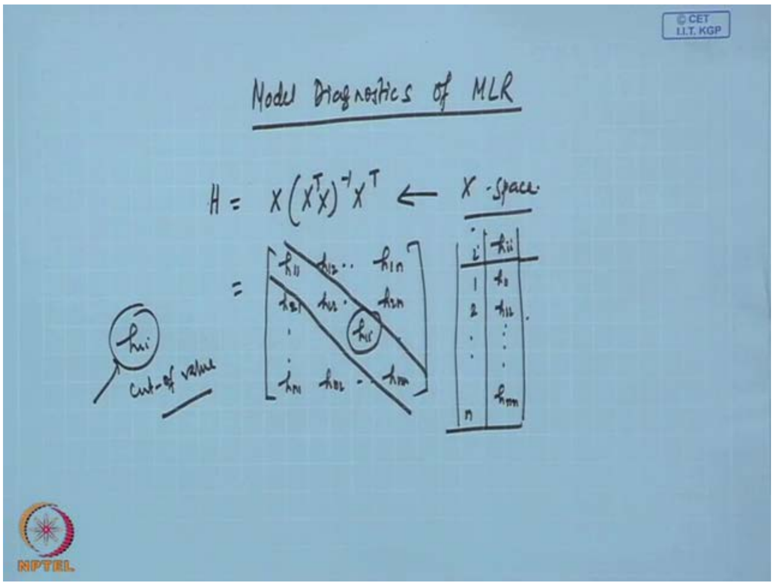
If you see the regression coefficients, now that absenteeism has affect because P value is 0.022, but absenteeism case it is 0.53, so it is that has no effect and which is also rebuilt in this picture there is no effect. So, if we include M ratio what will happen, ultimately your regression fit will be better r square 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: 08:47)



So, in order to identify leverage points you have to understand the hat matrix.

(Refer Slide Time: 08:57)

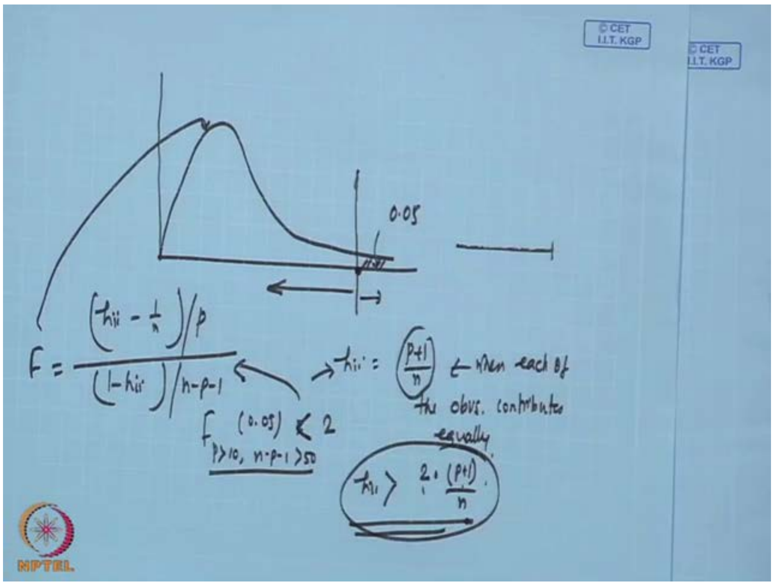


So, I think we have described hat last class not last, but one, this is the hat matrix see ultimately this one all related to x space, so when you are talking about leverages it is related to the space created by the x matrix. This one you have already seen that this is basically h 1 1 to h 1 1, 1 2, 2 1 n, h 2 1, 2 2, 2 n like this I think h n 1, h n 2 to h n n and there will be somewhere h i i. So, leverage values are the diagonal elements, so in the head matrix for we have i of i equal to 1 to n observations and you find out the h i i values these are known as leverage values.

So, h 1 1, h 2 2 like h n n they are all leverage values, now what will be the value of h i i that h i i value what will be the cut off value for h i i 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 distribution of the h i i you will find out that h i i minus 1 by p divided by p and 1 minus h i i by n minus p minus 1 this quantity follows f distribution with p n minus p minus 1 degrees of freedom.

Now, if your p is greater than 10 and n minus p minus 1 greater than 50 means what we are saying if you take large observations as well as your p is little more and for alpha equal to 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 minus p minus 1 when this condition satisfies, so it will be less than equal to 2.

(Refer Slide Time: 12:17)



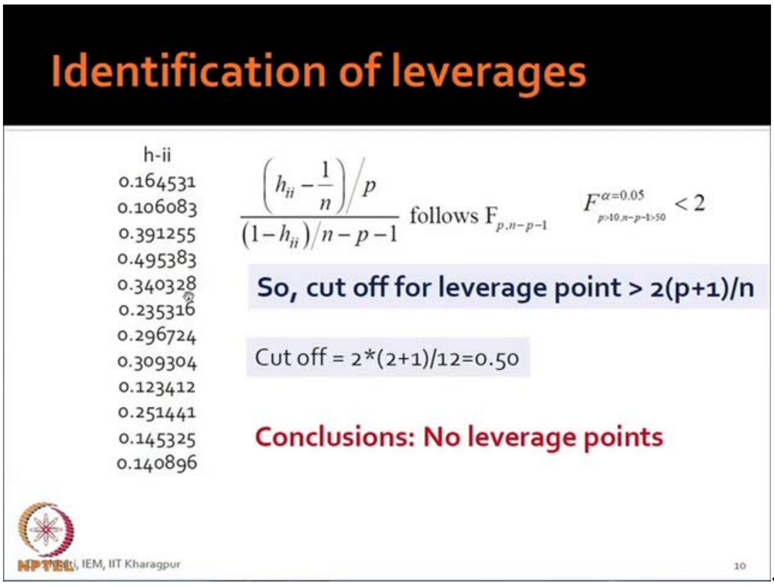
So that what we mean by this when say that whether it is influential or not, that mean you are considering chi square distribution and you will be considering this region that is 0.05. So, you are saying it is within this side then it is not influential when it goes to this side this is influential observation, so that mean we want to find out h i i value for this point. It is it is given that that h i i minus 1 by n divided by degree of freedom and 1 minus h i i 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, but the most widely used is that cut off for leverage point that will be greater than 2 into p plus 1 by n 2 into p plus 1 by n p plus 1 is the number of parameter to be estimated. 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 i i measures it measures the leverage values and the sum total of h i i this will be equal to the number of parameters.

So, then if all our all points are equally influencing then what will happen h i value be equal for all the points and that value will be p plus 1 by n. So, that mean what we mean to say if they are equally h i i equal to p plus 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 greater than 10 n minus p minus 1 greater than 50. If we take alpha equal to this point this will always less than 2 irrespective of any other p when this condition satisfy.

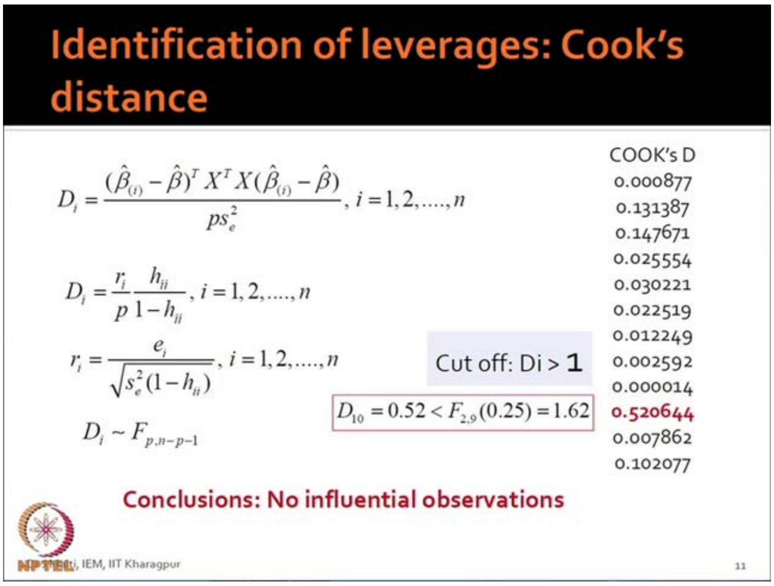
So, that is why they are saying that if you multiply this by 2, so what will happen h i i this greater than you are multiplying by 2 into p plus 1 by n, this is the average value this one. So, from average how much you are going this side depending on this value that is why 2 is multiplied here, so if your value h i value, any h i value which is more than 2 into p plus 1 by n that is leverage value.

(Refer Slide Time: 15:56)



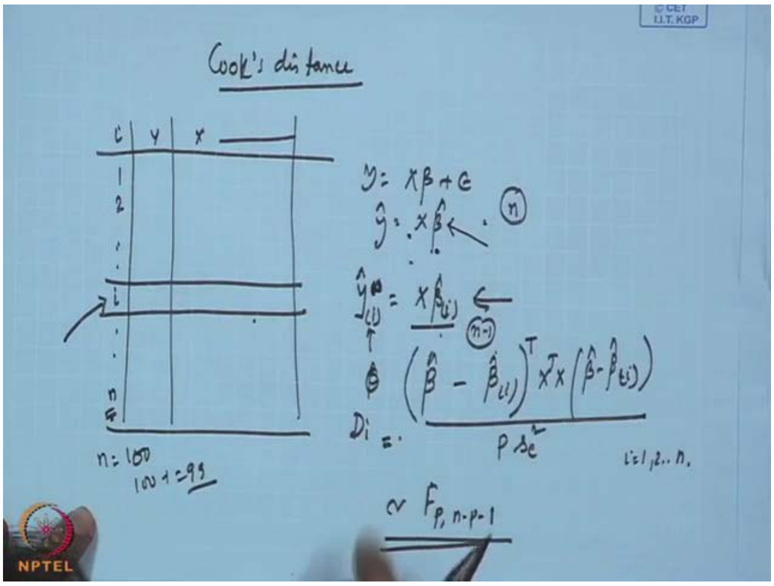
Now, see this for our case, our case you see that h i i values are observation 1 to observation 12 and h i i values are given these are all the diagonal values of the head matrix. Now, what will be the cut off value, cut off value will be we have 2, how many parameters we are estimating 3, what is your sample size n. So, p is 2 plus 1 into 2 by 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: 16:55)



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 i i values and the formulation what we have discussed so far that you can use cook distance also.

(Refer Slide Time: 17:31)



Cook’s distance, what is the procedure in Cook’s distance the procedure is like this, so i equal to 1 2 i dot 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 plus epsilon and using all that observations you have computed y cap equal to x beta cap. Now, what is our interest, our interest is we want to know is the i-th 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 equal to 100 or 100 minus 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 minus 1 data point, the i-th 1 this i the i-th 1 is eliminated. Then what we are saying this should not the difference between this beta cap n minus beta cap, i this difference should not be much really it should that in effectively 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 cap minus beta i cap transpose x transpose x then beta cap minus beta i cap this divided by p into s square and definitely i equal to 1 to n.

So, he created one statistics this type of statistics, so what happen you eliminate the i-th observation do the second round in regression modelling, find out the beta values. You have several x values these are all matrix vector values or matrix of the order p cross p plus 1 cross 1 and then you create this type of statistics that D i equal to this, this one follows f distribution with p n minus p minus 1 degrees of freedom.

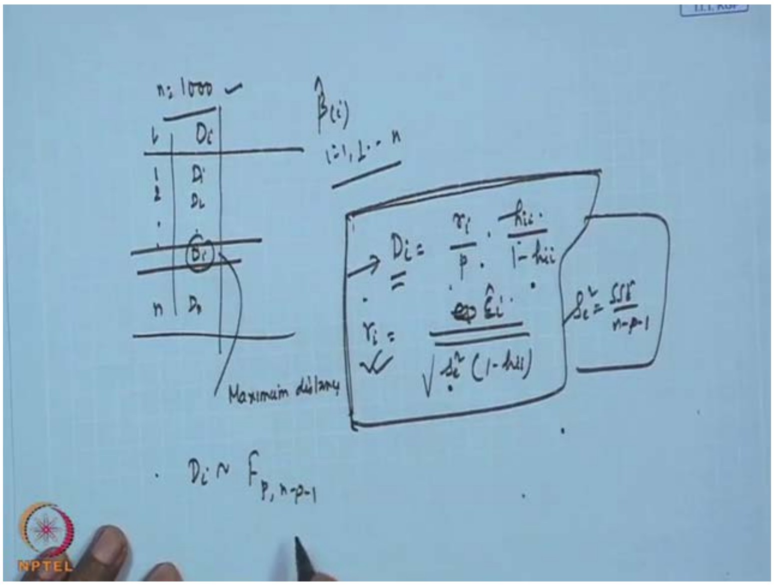
Student: Sir 1 minute, when we will eliminate the i-th observation means x as well as y.

That is total observation, yes.

Student: So, that time this x matrix 1 we consider from that matrix also we have to eliminate that?

Total that x y total as you said the i-th observation including x and y you eliminate, 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 D i 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: 22:52)



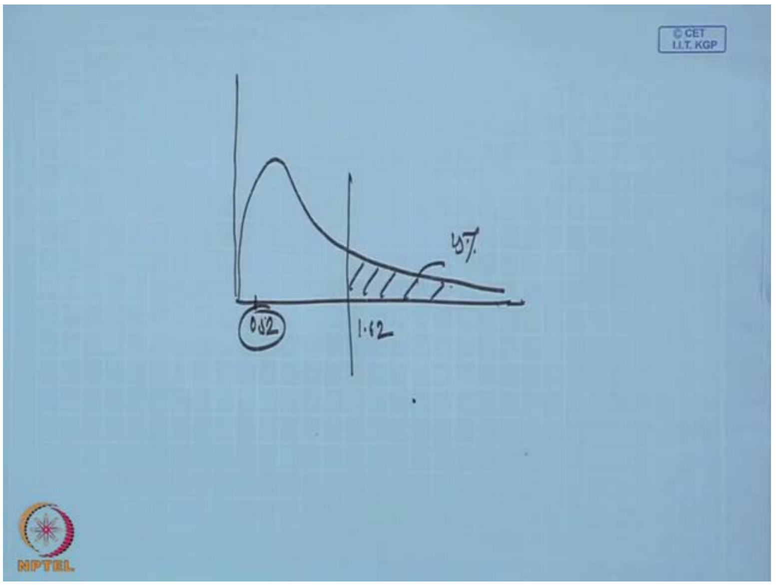
Now, question is there are suppose 1000 data points, so should I go for that 1000 plus 1 when every time you are eliminating 1 the first one second one like this so many observations. So, you will be having so many regression, fittings regression equation you have to develop you have to calculate beta 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 h i i by 1 minus h i i.

So, h i i 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 by square root of s square 1 minus h i i e i is the basically the error one, error one. So, i given that like this epsilon i cap you know s square, s square is SSE by n minus p minus 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 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, yes not coming yeah.

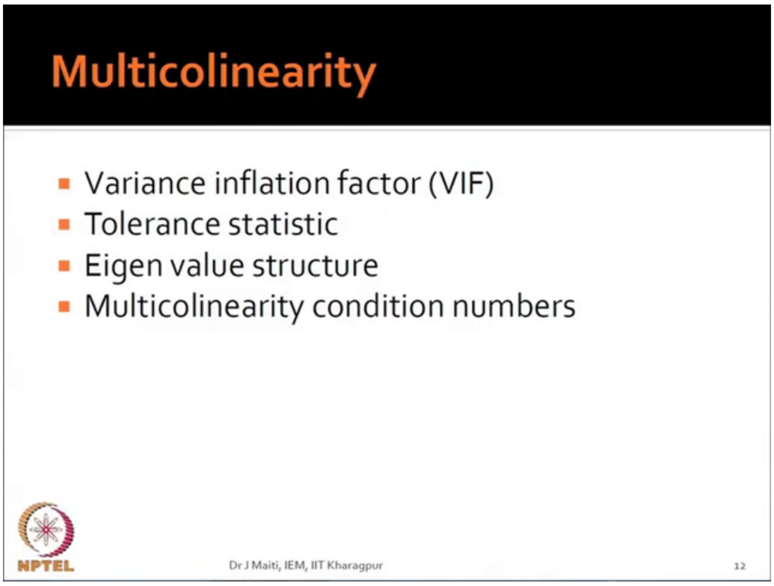
So, using this you find the Cook's distance, so ultimately 1 to n then D i value you are getting, so if I say D 1 D 2 like this D n then you find out the maximum one which one is maximum let the D i, this one is the maximum distance. So, for this maximum distance what we say that D i follow F p, n minus p minus 1, so I will take the maximum distance and then what is p value in our case. In our case p is 2 n minus p minus 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 closer.

(Refer Slide Time: 26:41)



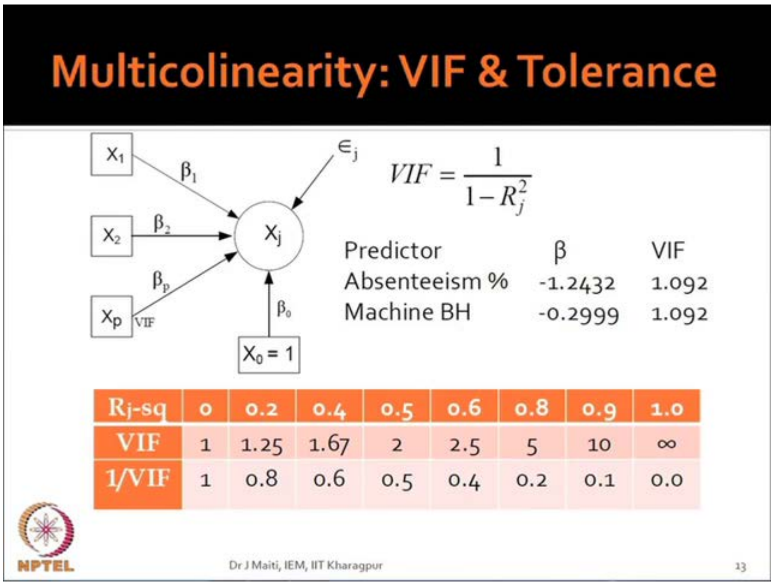
It is if I see the F distribution table sorry, graph like this we have taken 25 percent this is 25 percent t and this value is 1.62, but your maximum value, here it is 0.52, so it is not at all a influential point.

(Refer Slide Time: 27:16)



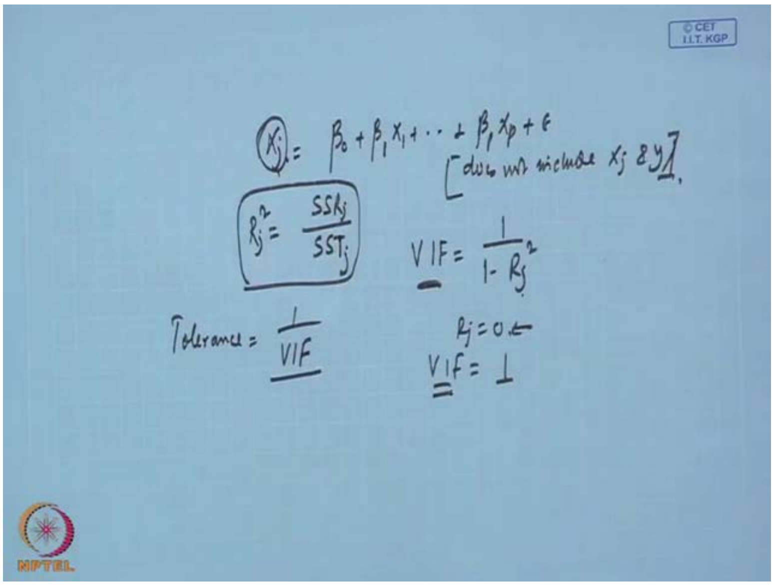
Then we will go for multicolinearity, now multicolinearity as I told you multicolinearity is a, is an issue where independent variables are not truly independent there is, there is dependence structure amongst the independent variables. Under such condition what will happen if there is linear case, linear dependence case the determinant of this x transpose 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, multicolinearity has to be tested and multicolinearity can be tested through four different procedures and these are known as variation inflation factor, tolerance statistic, Eigen value structure and multicolinearity condition number. So, what we will discuss we will first discuss the variance inflation factor.

(Refer Slide Time: 28:28)



What is variance inflation factor, variance inflation factor 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 y. 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, X p then you are making a regression equation, so your regression equation is.

(Refer Slide Time: 29:24)

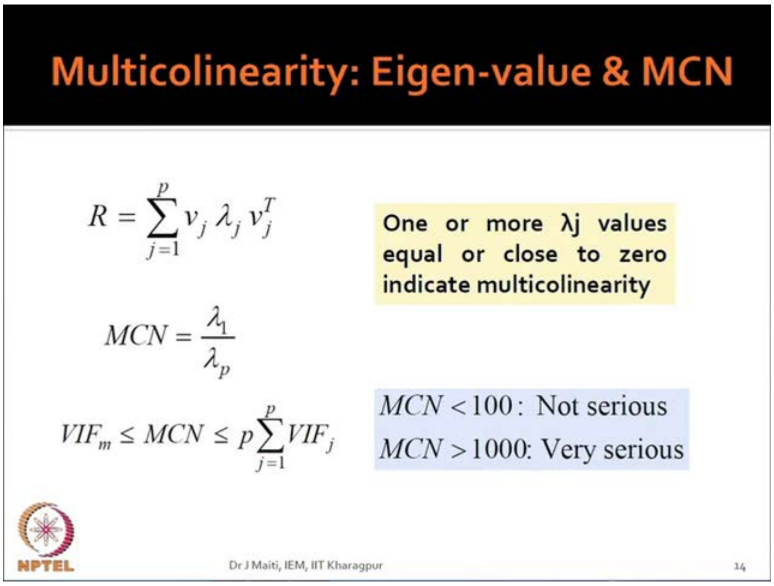


Now, X j is beta 0, beta 1 X 1, so like this beta p X p plus epsilon this does not include X j and y then you find out the R j square, so that will be S S R j by S S T that is for the j-th variable. So, then you create variance inflation factor equal to 1 by 1 minus R j square for example in our case absenteeism and machine breakdown 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 square R j square this R j square 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.

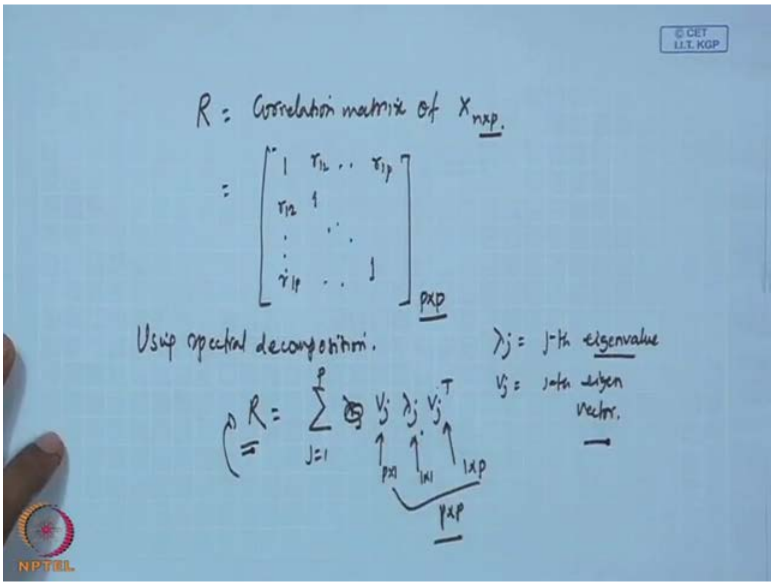
Here, what is happening you are getting within a 0 to 1 scale, 0 to 1 scale, here any value is possible, so as if we get in terms of 0 to 1 scale it is easier for us to interpret. So, now then what will be the VIF value that should be considered you are getting me for basically we say that if the VIF value is 10 or more this mean high collinearity, high relationship. So, 10 or more 10 is the cut off 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: 33:03)



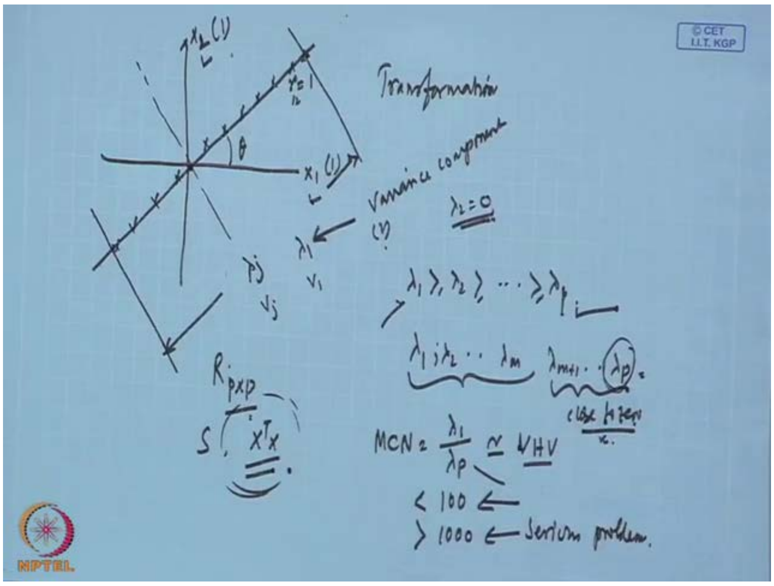
Then another issue is that another is the Eigen value criteria, what is this Eigen value criteria, in Eigen value criteria, so all of you know that the correlation matrix.

(Refer Slide Time: 33:22)



We are talking about correlation matrix of x n cross 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 transpose j equal to 1 to p where lambda j is the j-th Eigen value and v j is the j-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 reconstruct r because this one is p cross 1, this one is 1 cross 1, this one is 1 cross p. So, if you multiply this two ultimately p cross p matrix you will be able to recreate, now what is the meaning of this Eigen value, here when you do the spectral decomposition.

(Refer Slide Time: 35:20)



Eigen value, here 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 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, 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 1 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, two variables which in the standardized case suppose this one and this is one. 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.

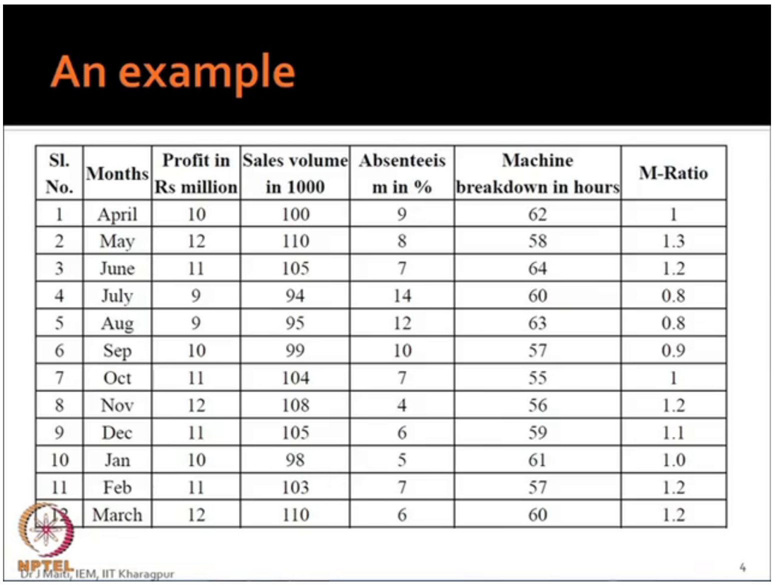
So, in the two variable situations when you decompose the R matrix into its Eigen value, Eigen vector and if you find out that one of them is 0 lambda value is 0, then it is a perfect correlation case. So, similarly if there are p such variables, so you will be getting lambda 1 greater than equal to lambda 2 greater than equal to 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 r is basically if it is it is basically p cross p. So, we assume that R or S or X transpose X when we do regression we assume that X transpose 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, lambda n these are basically having values not equal to 0, but M p plus 1 to M p these are close to 0, 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 v M plus 1 to p that large number of independent variable, so called independent they are not independent. So, there is multicolinearity this is what is known, is known an Eigen value collected here, 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. Then there is a multicolinearity number this multicolinearity number is known as MCN which is basically the largest Eigen value divided by the smallest one. Now, if the 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 n MCN greater less than 100 it is not a serious multicolinearity 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 V I 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: 41:49)



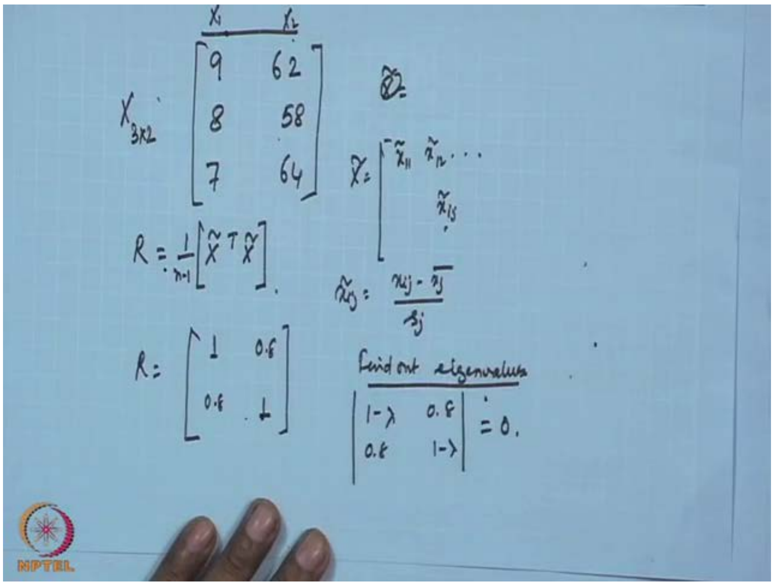
An example

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
12	March	12	110	6	60	1.2

Dr J Maiti, IEM, IIT Kharagpur

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 multicolinearity problem is there or not, so what way you proceed what will be your case.

(Refer Slide Time: 42:19)



For example 9, 62, 8, 58 and then 7, 64suppose these three data points are given to you, so this is my X this is 3 cross 2, this is X 1 and X 2 what is our aim we want to test the X

1, X 2 that multicolinearity issues are there are not 1 is that you find out the regression. You regress X 2 on X 21 since there are only two variable X 1 and X 2 is enough and then find out the V I F, 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 tilde transpose X tilde 1 by n minus 1 and your X tilde is suppose if I say X tilde is something like this yes.

So, suppose X i j tilde is there then X i j tilde will be X i j minus X j bar by standard deviation of this, so we require to get R square 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 one and this is one 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, find out the Eigen values, how do you find out Eigen values.

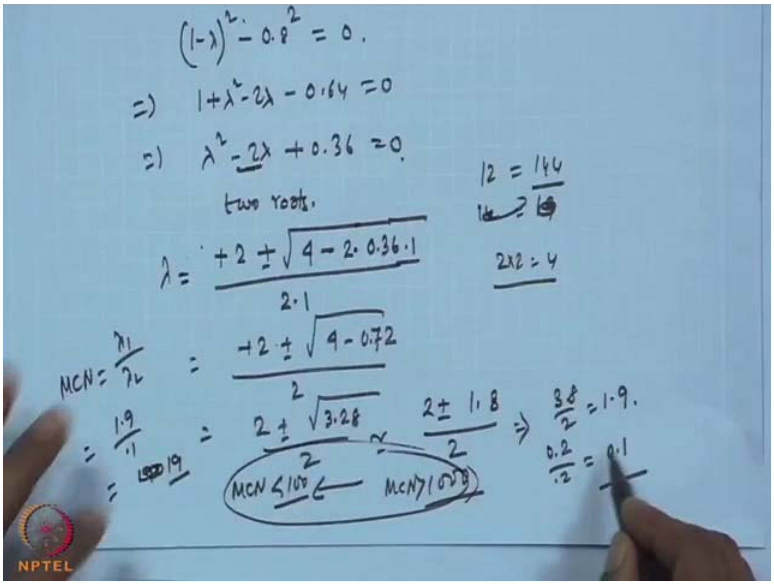
Student: Normal process.

Normal processes that characteristic root.

Student: Yeah.

Guess I show you how to first consider that 1 minus lambda 0.8, 0.8 1 minus lambda this determinant this equal to 0.

(Refer Slide Time: 45:05)



Let us see this what will happen, here will we get 1 minus lambda square minus 8 square equal to 0, so 1 plus lambda square minus 2 lambda minus 0.64 it is this equal to 0 then lambda square minus 2 lambda plus 0.36 equal to 0 then there are two roots. So, lambda will be minus b means plus 2 plus minus root over this square mean 4 minus 2 into 0.36 divided by 2 into 1, into 1 this is the case. So, then plus 2 plus minus root over 4 minus 2, 72, 0.72 by 2, so 2 plus minus root over 3.28 divided by 2, so 3.28 what will be the square root.

So, see if it is 12 then 144, if it is 14 190, 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 plus minus may be it will be 1.8 by 2 then it is 3.8 by 2 this one is 1.9 another one is 0.2 by 2 0.1. So, we can we can say that it is basically it is a multicolinearity problem because we have already taken this essentially then what it is what is what is coming. Now, that multicolinearity 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 multicolinearity definitely with the two variable case when 0.88 is saying that it is almost that multicolinearity is like this. Now, what is our MCN multicolinearity number that lambda 1 by lambda 2, so our 1.9 by 0.1 this is 190 or 19 this is 19. So, what we have seen that multicolinearity number less than 100 is not a serious issue that is what is given there if multicolinearity 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.

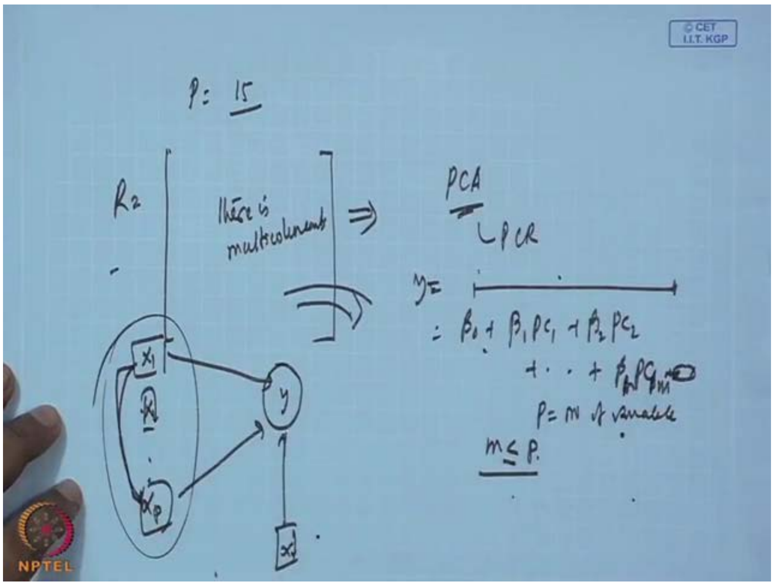
So, should I go for regression or we will simply dimension we reduce the dimension and then we will go for regression what will be your issue. Basically, see if I go actually although 0.1 in, here we are getting 0.9, because of this two variable case I think this 0.8 where is not at all a simple issue 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.

Student: Sir, we can perform dimension reduction.

Dimension reduction it is better.

Then see then what we will do suppose, here it is p for two variable case, but there are P variables.

(Refer Slide Time: 50:00)

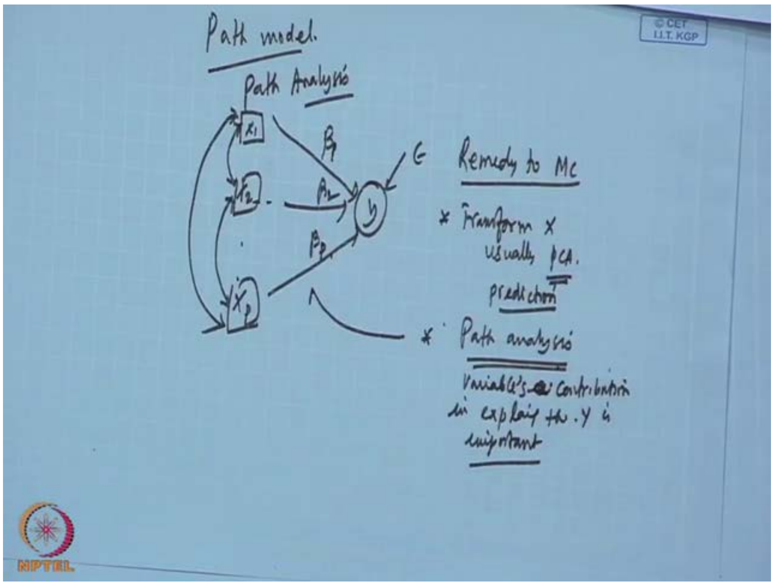


Suppose P is greater than 15 variables or 20 variables, so under this case you will be having a big correlation matrix. Using this correlation matrix seeing this value you cannot judge because even in the two variable case if I see 0.8 and then I discard. But, from multicolinearity number point of view, it is saying that you should go, you should go for, should go for regression without bothering for multicolinearity, so you cannot judge just by seeing the R matrix, fine. Then the solutions are what are the solutions, solution is one is the principle component analysis when there is multicolinearity.

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 that are significant values. Only those components you take then your y is for those components you find out the regression line the beta 0. Suppose beta 1 P c 1 plus beta 2 P c 2 plus like this beta P P c P and so on P c P we will not go, we will go for beta m P c m P is the number of variables where m definitely is less than equal to P.

Now, if you go for P C A 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 1 this is the structure X p X 0, now what happen they are dependent suppose this is dependent with this is dependent. So, what do you want if you go by P C A 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 materials. Strict sense if you still find that, no these are still independent variable they cannot be treated as dependent variable.

(Refer Slide Time: 52:39)



Then you can allow them in modelling, 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 multicolinearity what we have discussed remedy to multicolinearity, multicolinearity one is the transform the data transform X that is usually we go for P C A. Here, P C A 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.

Student: 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 covariant structure will be taken into consideration and then these two analysis. Is it is better to go for path analysis when variable explanation is an important issue variable’s contribution in explaining the y is important, so I think we can stop now.