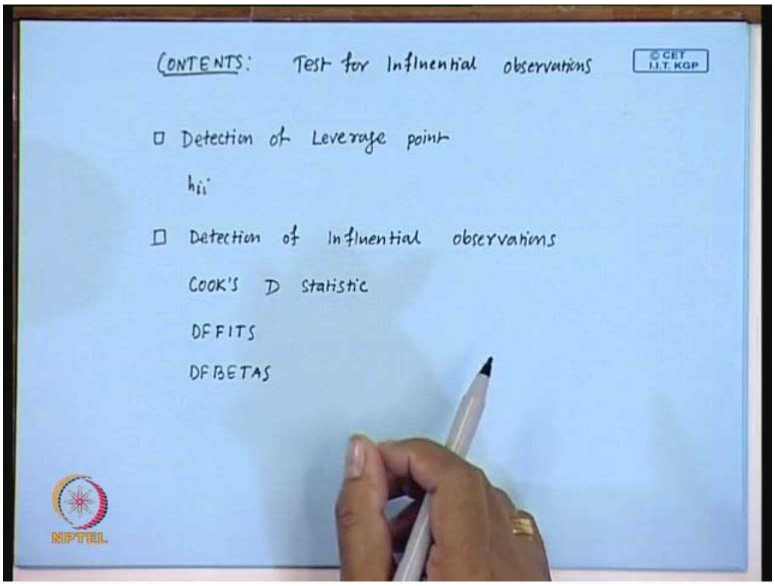
Regression Analysis
Prof. Soumen Maity
Department of Mathematics
Indian Institute of Technology, Kharagpur

Lecture - 20
Test for Influential Observations

(Refer Slide Time: 00:26)



CONTENTS: Test for Influential observations

□ Detection of Leverage point

h_{ii}

□ Detection of Influential observations

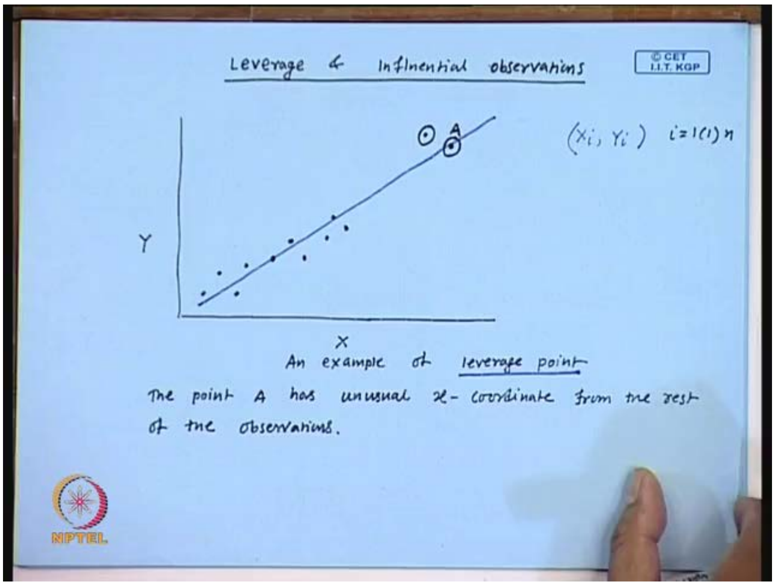
Cook's D statistic

DFFITS

DFBETAS

Hi, this is my first lecture in module 6, that is Test for Influential Observations. And here is the content of this module, test for influential observations. First we will be talking about detection of leverage point using h i i, where h i i is the i th diagonal element of the hat matrix h. And then we will be talking about detection of influential observations, using Cook's D statistics, and then the DFFITS that is difference of fits and this is DFBEATS that is the difference of beta values. So, the objective of this module is to present different techniques to detect the influential observation. So, we have learned about the leverage point and the influential observation in the previous module, again I will just repeat those things one we will recall those things once more.

(Refer Slide Time: 01:47)



Leverage & Influential observations

© CET I.I.T. KGP

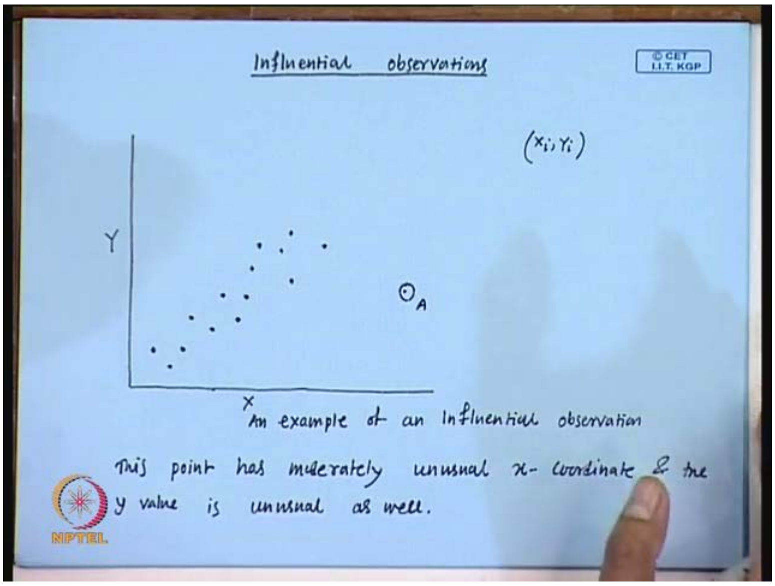
$(X_i, Y_i) \quad i=1(1)n$

An example of leverage point

The point A has unusual x-coordinate from the rest of the observations.

Here is the definition for leverage point, so I you have several observations like (X i, Y i) for i equal to 1 to n, and here you know is the scatter plot for the observations. And the point, so these are the points in the scatter plot and as you can see, the point A here has unusual x coordinate from the rest of the observations. So, this point A is called leverage point, so it could be, so not necessarily that if you fit a module to the given data or if the this point is exactly lying on the fitted module, but not necessarily I mean this is also a leverage point for example. So, what I want to say is that, it is the point is on the is lying on the trend of the observation, but this has unusual x coordinate, so this type of points are called leverage point. And next we will be talking about influential observation, so here is an example of influential observation.

(Refer Slide Time: 03:43)



Influential observations

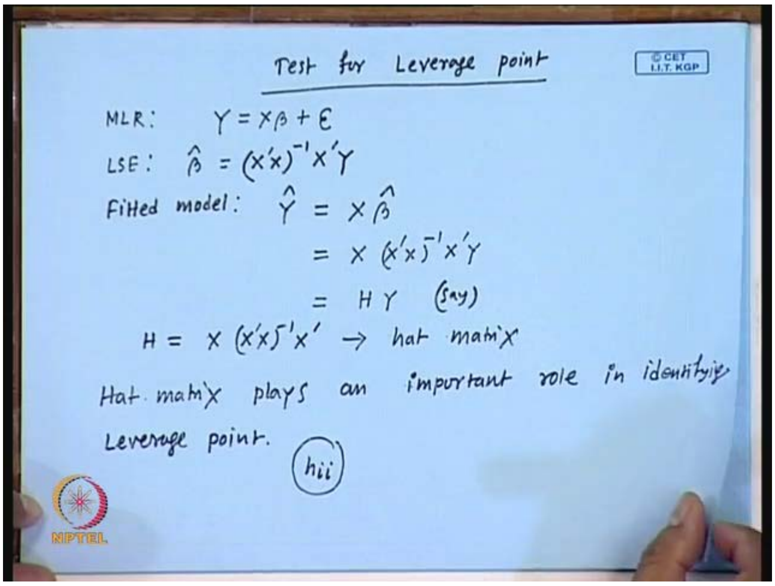
An example of an Influential observation

This point has moderately unusual x- coordinate & the y value is unusual as well.

So, look at this point A here, again you this is the scatter plot for the given observation (X i, Y i) and the point a here is this has moderately unusual x coordinate, and also it has moderately unusually y coordinate. So, this point is has moderately unusually x coordinate, and also it is not in the general trend of the data, so this point is called an influential observation. So, what we learned is that if a point has unusual x coordinate then the point is, but it may lie on the general trend of the data, then the point is called leverage point.

And if a point has moderately unusual x coordinate and also unusual y coordinate, then the point is called influential observation. And the influential observation has significant effect on the model regression coefficients. So, what we will do in this module today is that, we will be talking about several techniques to detect the leverage point and influential point. So, first I will be talking about one technique, which can detect leverage point.

(Refer Slide Time: 05:39)



Test for Leverage point

MLR: $Y = X\beta + \epsilon$

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

Fitted model: $\hat{Y} = X\hat{\beta}$

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

$= HY$ (say)

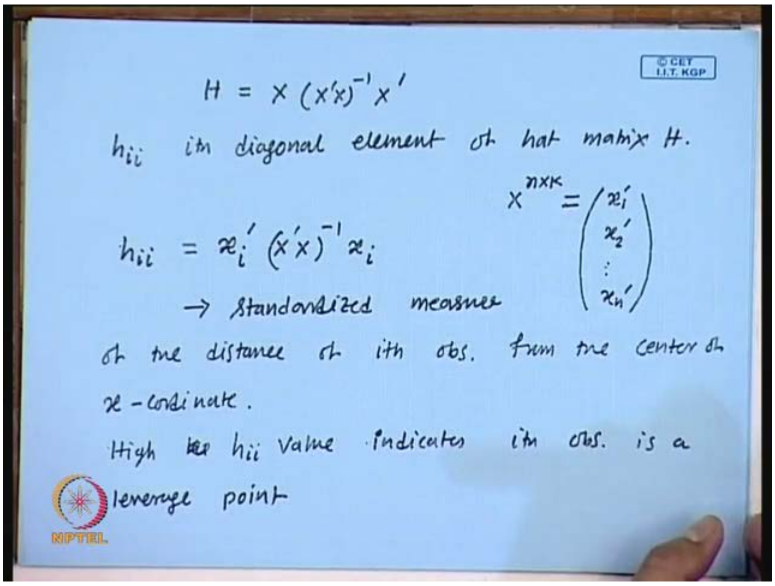
$H = X(X'X)^{-1}X'$ → hat matrix

Hat matrix plays an important role in identifyig Leverage point. $\boxed{h_{ii}}$

So, test for leverage point, so what I will do is that first I will again recall the multiple linear regression model. And in matrix form here is the model Y equal to X beta plus epsilon and the list square estimate of this regression coefficient beta, so this is a vector with k regression coefficient. So, beta is equal to (X prime x) inverse X prime Y, so the fitted model is Y hat, which is equal to X beta hat. So, this is equal to X beta hat equal to (X prime x) inverse X prime Y, and this we can write as H Y, we talked about all these things in the previous module, just I am recalling here the same thing again.

So, here H is equal to (X prime x) inverse X prime, and this matrix is called the hat matrix and this is called hat matrix, because this matrix maps Y to Y hat, so this hat matrix plays an important role in identifying leverage point. So, I told that in fact, you know the i th diagonal element of this hat matrix that is h i i, so h i i is the i th diagonal element of this hat matrix. And this has an important role to check whether the i th observation is a leverage point or not, so let me write what is this h i i.

(Refer Slide Time: 09:07)



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

h_{ii} ith diagonal element of hat matrix H.

$$X^{n \times K} = \begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{pmatrix}$$

$$h_{ii} = x_i'(X'X)^{-1}x_i$$

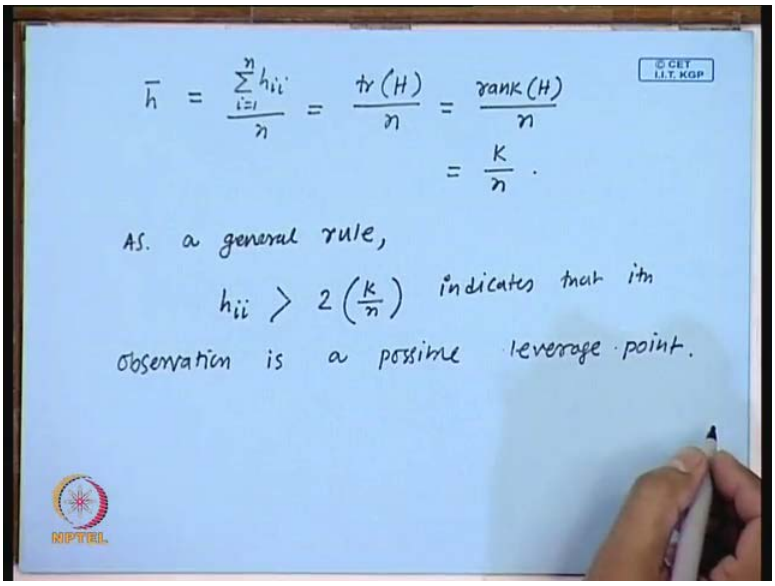
→ Standardized measure of the distance of ith obs. from the center of x-cordinate.

High h_{ii} value indicates ith obs. is a leverage point

So, H matrix is equal to, H is equal to (X prime x) inverse X prime, then what is h i i, h i i is the i th diagonal element of hat matrix H. So, X is n cross K matrix and let me just write that the n rows, n rows or say for example, x 1 prime, x 2 prime, x n prime are the n rows of this X matrix. Then in terms of x i the i th row, I can write h i i equal to x i prime (X prime x) inverse x i, so see this part is independent of i, I mean whatever be the value of the i this does not change.

So, what does this mat measure is that, this h i i is this one is a standardized measure of the distance of i th observation from the center of x coordinat. So, what you have observed is that the i th diagonal element of the hat matrix h i i, this one measures the standardized distance of the i th point from the center of x coordinate. So, and we call a point leverage point, if it has unusual x coordinates, so the obviously, if h i i is large for a particular i, then that indicates the i th point is a leverage point. So, the conclusion here is that the criteria for i to be a leverage point is that. So, usually the high h i i value indicates i th observation is a leverage point, so and I said that, so let me just compute the average value of h.

(Refer Slide Time: 13:36)



$$\bar{h} = \frac{\sum_{i=1}^{n} h_{ii}}{n} = \frac{tr(H)}{n} = \frac{rank(H)}{n}$$
$$= \frac{K}{n}.$$

As. a general rule,

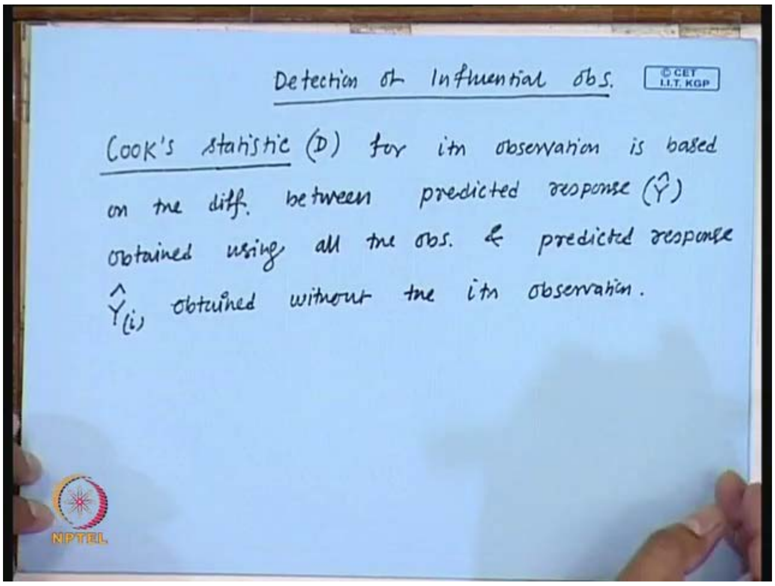
$h_{ii} > 2\left(\frac{K}{n}\right)$ indicates that ith observation is a possible leverage point.

So, I said that if the value of h i i is large, then that indicates that i th observation is leverage point, so what I am trying to do is that, I am trying to find the average value of h. And for a particular h i i is very larger may be more than the double of the average value, then we say that the point i is a leverage point. So, h bar is equal to h i i, i equal to 1 to n by n, so this one is nothing but the trace of the hat H matrix, trace means the sum of the diagonal elements.

So, the trace of H by n and for this matrix H, the trace of H is equal to rank H by n and the rank of H this matrix is K, so the average value of H is K by n. So, as a general rule h i i greater than 2 times K by n indicates that i th observation is a possible leverage point, so next we will be talking about some technique to detect the influential observation. So, see in case of leverage point, a point is said to be leverage point, if it has unusual x coordinate and h i i measure the distance of the point i th point from the center of x coordinate.

So, then obviously, if the h i i value is large, then that indicates the i th point is a leverage point, but in case of inferential observation, recall the definition of influential observation that it has a point is said to be influential observation, if it has moderately unusual x coordinate as well as moderately unusual y coordinate. So, here we need to take care of both the x coordinate and the y coordinate, and and cook has suggested a statistic to do this; so next we will be talking about Cook's statistics to detect influential observation.

(Refer Slide Time: 17:29)



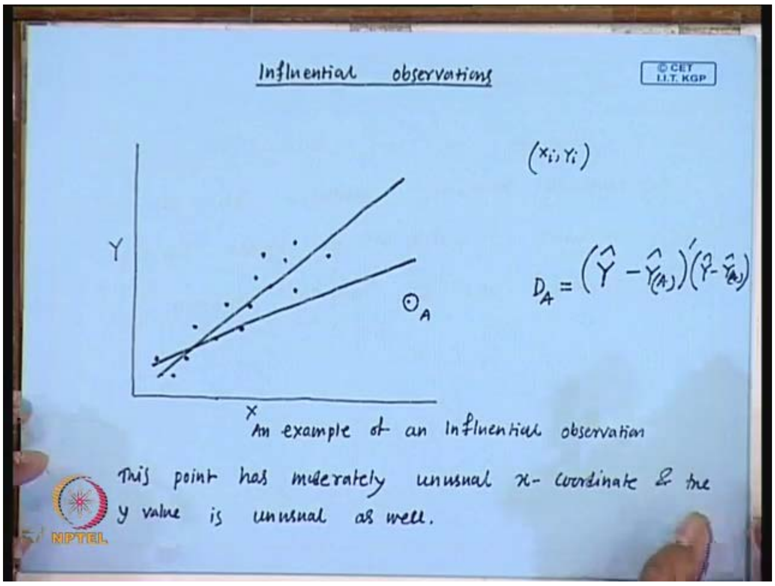
Detection of Influential obs.

Cook's statistic (D) for ith observation is based on the diff. between predicted response ($\hat{Y}$) obtained using all the obs. & predicted response $\hat{Y}_{(i)}$ obtained without the ith observation.

So, detection of influential observation, so we will be talking about Cook's statistic and it is denoted by D, so it is the distance between two things. So, what this Cook statistic does is that, it measures the distance between the fitted response value, obtained using all the observations that is the usual fitted, I am talking about the Cook's statistics of the i th observation. So, what is does is that, it measures the distance between fitted observation obtained using all the data or all the observations.

And the fitted response obtained without using the i th observation, I mean using all the observations except the i th observation, so let me write down the Cook statistic formulae. So, Cook statistic for i th observation is based on the difference between predicted response, we call it now Y hat obtained I mean, this is obtained using all the observations and predicted response Y i. So, this one is predicted response, which is obtained using all the observations except the i th observation obtained. So, this one is obtained without the i th observation.

(Refer Slide Time: 21:00)



Influential observations

© CET I.I.T. KGP

(x_i, y_i)

$$D_A = \left(\hat{Y} - \hat{Y}_{(A)}\right)'\left(\hat{Y} - \hat{Y}_{(A)}\right)$$

An example of an influential observation

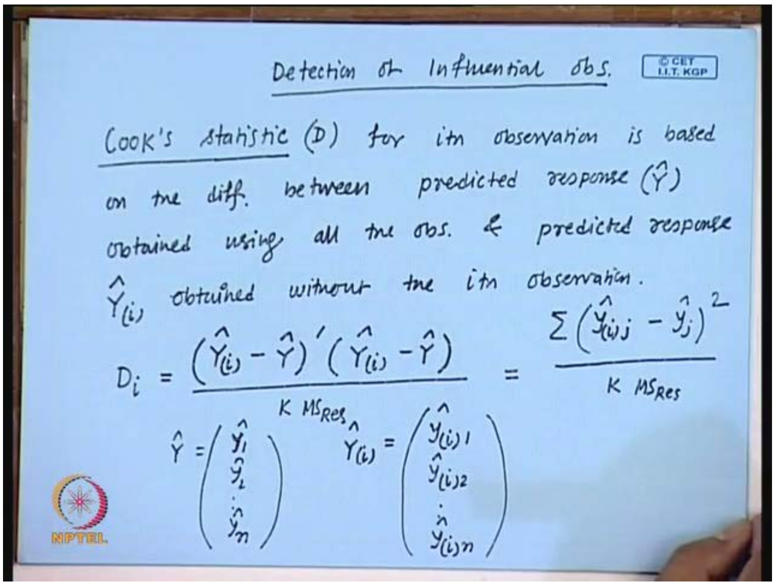
This point has moderately unusual x- coordinate & the y value is unusual as well.

So, what I want to say is that, what the Cook statistic does is that, suppose you want to compute the Cook statistic for the observation A here; that means, we call it D A. So, what that D A is that, so D i is the Cook statistic for the i th observation, so D A is the Cook statistic for observation A. First you compute the distance, I mean you have to compute the euclidean distance between Y hat and Y A hat, so Y hat is the fitted response based on all the observations. So, if you consider all the observations, then the model will get influenced by this observation, and the fitted model may be look like this.

Now, so this is what the Y hat, once you have this fitted model you can get Y 1 hat Y 2 hat, Y 3 hat, Y n hat everything and this is the vector and vector of fitted response values. And now you compute Y A hat, so this is the fitted or predicted response value, obtained without using observation A, but using all the other observations. So, if I do not use this observation, then my model will look like this, fitted model will look like this, so from here I will get Y 1 A hat and all these things Y n A hat.

So, (Refer Time: 23:07) this a vector, this is a vector you find the euclidean distance between these two vectors that can be obtained using this way, this is Y hat minus Y A hat. So, this is what want to mean by the distance of, I mean this Cook statistic measure the distance between the predicted response obtained using all the observations, and the predicted response based on all the observations except the i th observation. So, this is how we get the Cook statistic for the i th observation and similarly, you do it for all the observations, so you will get D 1, D 2, D n, so this is how we get the D i.

(Refer Slide Time: 23:59)



Detection of Influential obs.

©CET I.I.T. KGP

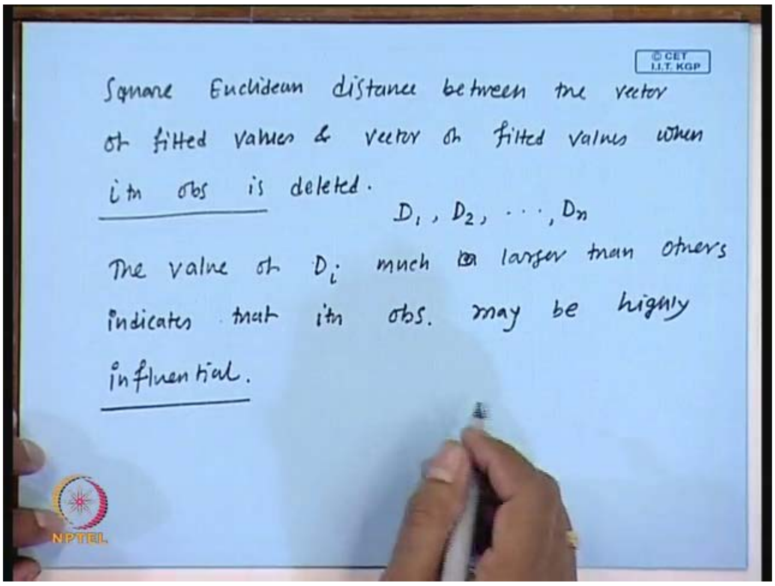
Cook's statistic (D) for ith observation is based on the diff. between predicted response ($\hat{Y}$) obtained using all the obs. & predicted response $\hat{Y}_{(i)}$ obtained without the ith observation.

$$D_i = \frac{(\hat{Y}_{(i)} - \hat{Y})'(\hat{Y}_{(i)} - \hat{Y})}{K\, MS_{Res}} = \frac{\sum (\hat{y}_{(i)j} - \hat{y}_j)^2}{K\, MS_{Res}}$$

$$\hat{Y} = \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{pmatrix} \qquad \hat{Y}_{(i)} = \begin{pmatrix} \hat{y}_{(i)1} \\ \hat{y}_{(i)2} \\ \vdots \\ \hat{y}_{(i)n} \end{pmatrix}$$

The cook statistic for the i th observation D i is equal to Y i hat it does not matter which one you writing first, Y hat prime Y (i) hat minus Y hat and you divide this by K times M S residual. So, you need to understand that this is a vector, so this Y hat is nothing but y 1 hat y 2 hat and y n hat, and you need to understand that Y i hat is the predicted observation obtained using all the data except the i th observation. So, this this vector is I may write it in this form y i 1 hat y i 2 hat like that, y i n hat, so if you write then this can be also written as this is equal to y i, may be let me use j here minus hat minus y j hat square, not difficult understand by K M S residual. So, this is the Cook statistic for the i th observation.

(Refer Slide Time: 26:28)



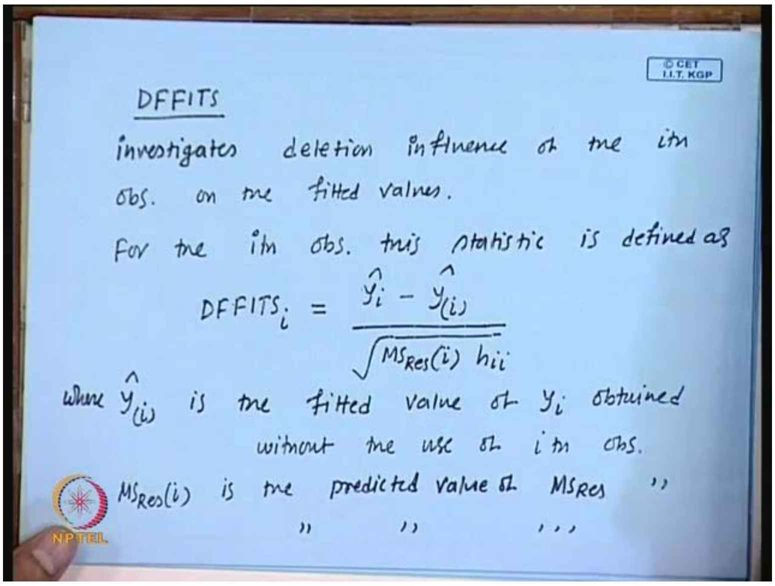
Square Euclidean distance between the vector of fitted values & vector of fitted values when ith obs is deleted.

$D_1, D_2, \cdots, D_n$

The value of D_i much larger than others indicates that ith obs. may be highly influential.

And this can be also treated as, the square euclidean distance between the vector of fitted values and vector of fitted, I mean response values when i th observation is deleted. And the rule to say that one observation is an influential observation is that the value of D I, so say you need to calculate all this D 1, D 1 stand for the Cook distance for the first observation, D 2 like up to D n. So, the value of D i much larger than others indicate that i th observation may be highly influential, so this is not difficult to understand that, if D i is going to be large for the influential observation. If i is not an influential observation, there is not going to be much difference between the fitted value using all the observations, and the fitted value without considering the i th observation. So, next we will be talking about two more statistics to detect influential observation.

(Refer Slide Time: 29:38)



DFFITS

investigates deletion influence of the ith obs. on the fitted values.

For the ith obs. this statistic is defined as

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{\sqrt{MS_{Res}(i)\, h_{ii}}}$$

where $\hat{y}_{(i)}$ is the fitted value of y_i obtained without the use of ith obs.

$MS_{Res}(i)$ is the predicted value of MS_{Res} " " " " "

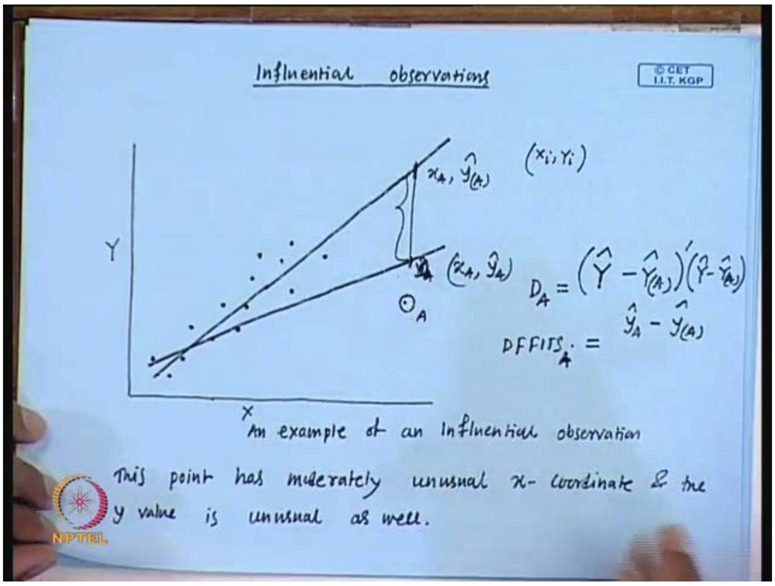
So, the next one is called DFFITS, so this is also called difference between fits statistics, so this one investigates deletion influence of the i th observation, observation on the fitted values. So, Cook statistics also does the same thing, it investigate the deletion influential influence of the i th observation on the fitted values, but here the statistic is different. So, for the i th observation this statistic is defined as D F, difference between fits i, which is equal to y i hat minus y i hat and we make it standardized by M S residual (i) h i i.

So, here you are only considering the i th observation I mean difference between, so what is this this is, this y i hat is the is the fitted value of y i obtained using all the observations. And this guy this y i bracket hat is the fitted value of y i obtained without using the i th observation, so this is y i hat, where y i hat is the fitted value of y i obtained without the use of i th observation. So, here, so this is using without using the i th observation means using all the observations, except the i th observation.

And similarly we have one new term, I need to introduce this notation that M S residual, M S residual (i) is also is the estimated is the predicted value of M S residual obtained without the use of i th observation. So, here also generally M S residual estimates sigma square, which is sigma square is variance of epsilon, which is unknown, but here the M S residual bracket i is the M S residual only that is obtained using all the observations except the i th observation that is the difference between M S residual and M S residual i.

And of course, if i th observation is going to be an influential observation, then this quantity is large, let me just try to explain this one using this example might be.

(Refer Slide Time: 35:03)



Influential observations

$D_A = (\hat{Y} - \hat{Y}_{(A)})'(\hat{Y} - \hat{Y}_{(A)})$

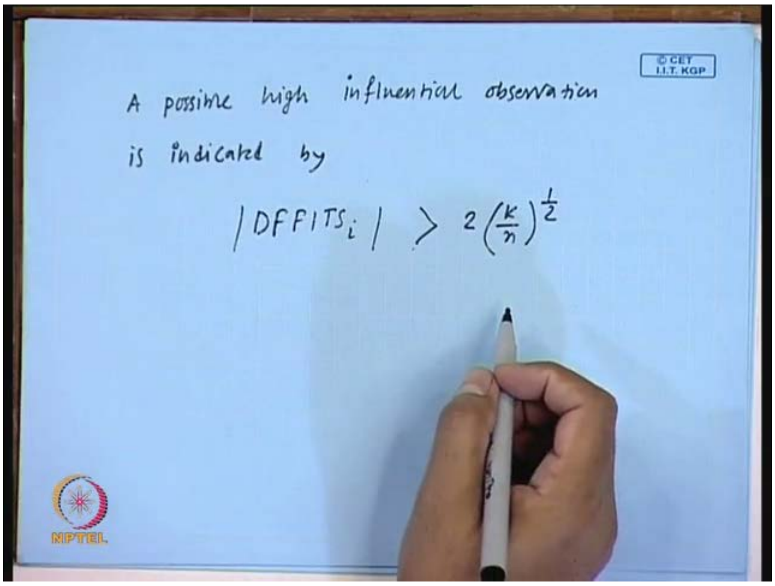
$DFFITS_A = \hat{y}_A - \hat{y}_{(A)}$

An example of an influential observation

This point has moderately unusual x-coordinate & the y value is unusual as well.

So, what I told is that my this difference of fits i, so I am doing it for A for example, here, so what is this, this is I said that y A bracket hat this is based on all observations minus y A bracket hat, and it has been standardized in some way. So, what is y A hat here; that means, this is using the fit using all the data, so this is my y A hat I mean this distance, this height is y A hat, this is the point x A, y A hat and what is y bracket A hat is that. So, this is the fit using all the observations except the observation A. So, this is the point x A, y A bracket hat and this is the point x A, y A hat the difference is this much and this difference is going to be large if the point is influential, quite clear because you take any other point and compute the difits for that point it is not going to be, so large.

(Refer Slide Time: 36:58)

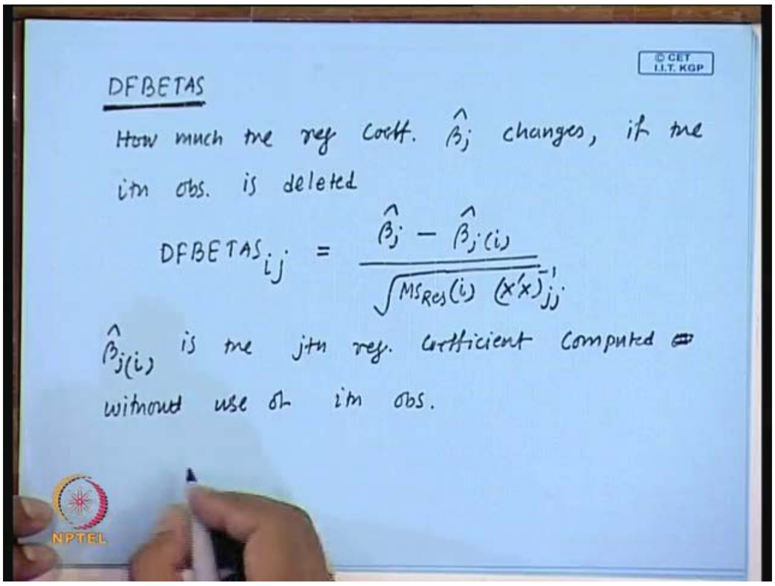


A possible high influential observation is indicated by

$$|DFFITS_i| > 2\left(\frac{k}{n}\right)^{\frac{1}{2}}$$

Now, the testing criteria is that a possible high influential observation is indicated by, if this statistic value for some observation i difference of fits, FITSi, if this quantity is greater than 2 times K by n. I am not going into the detail of how to get this critical value, but just is enough to know that if the DFFITS statistic for the i th observation is greater than this quantity for some observation, then the observation can be treated as influential observation. So, next we will be talking about one more statistic that also, this new statistic that also measure the influence of deleting the i th observation from the data set.

(Refer Slide Time: 38:40)



DFBETAS

How much the reg coeff. $\hat{\beta}_j$ changes, if the ith obs. is deleted

$$DFBETAS_{ij} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{MS_{Res}(i)\,(X'X)^{-1}_{jj}}}$$

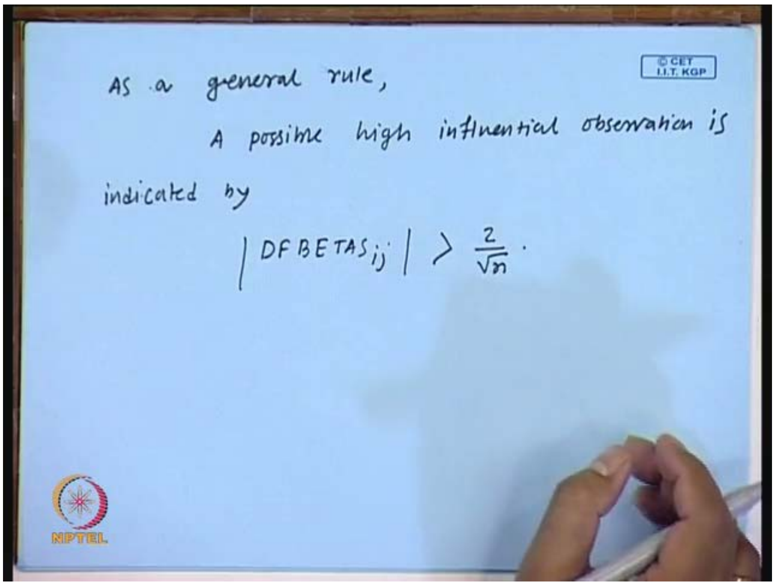
$\hat{\beta}_{j(i)}$ is the jth reg. coefficient computed without use of ith obs.

So, this one is called difference between the betas, and here it measures instead of measuring the difference of fits, I mean the difference of two fitted values, it measures the difference of two, I mean the estimated value of the regression coefficient beta j. So, here what we compute is that, we compute how much the regression coefficient beta j hat changes, if the i th observation is deleted.

So, instead of looking at the difference in fitted value here, this DFBETAS, it looks the change in regression coefficient beta j, so this DFBETAS for the i th observation and the j th regressor i j is equal to beta j hat minus beta j hat (i). And it has been standardized using this M S residual i and what is this and X prime x inverse j j, just let me tell what is this beta j, beta j hat is the least square estimate of beta j obtained using all the observations. And this beta j hat i, this beta j hat i is the j th regression coefficient computed without use of i th observation.

So, what you have to do here is that and what is this, this is the what is X prime x inverse and j j means is the jth diagonal element of X prime x inverse, and so this difference of betas, I mean DF betas is calculated for each i and for each j. So, here i runs from 1 to A and also j runs from 0 to K minus 1.

(Refer Slide Time: 43:02)



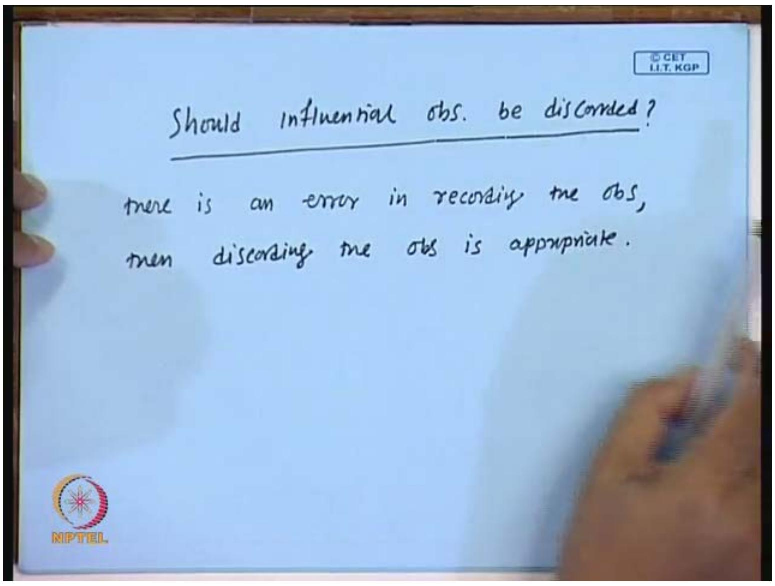
As a general rule,

A possible high influential observation is indicated by

$$\left| DFBETAS_{ij} \right| > \frac{2}{\sqrt{n}} .$$

And as a general rule a possible high influential observation is indicated by DF, difference of betas i j, if this value is difference is greater than 2 by root n, again I am not going in to the detail of how to get this critical point. So, these are the different techniques to detect influential observation or the leverage point. And next we will be talking about once an influential observation is detected, then whether this influential observation should be discarded or not. So, what we will do after detecting an influential observation in the given set of data.

(Refer Slide Time: 44:45)



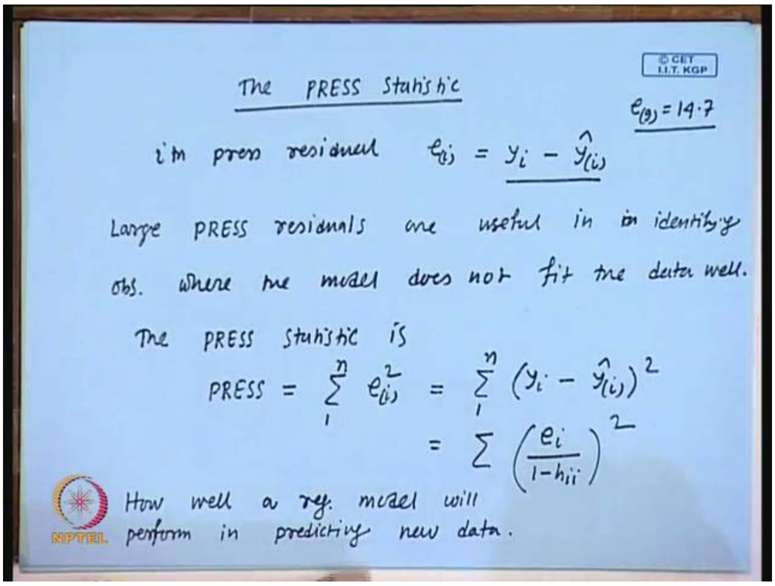
Should influential obs. be discarded?

there is an error in recording the obs, then discarding the obs is appropriate.

So, the question is, should influential observation be discarded, so here the recommendation is that we need to take care of the influential observation, I mean first we need to check, whether there is any error in measurement for that particular observation. And if you see that, if you see that there is an error in recording, the observation then the discarding, the observation is appropriate. Otherwise, now if you see, if you analysis reveal that there is a the observation is a valid observation.

Then there is no justification of discarding the discarding a valid observation only we need to taspecial care of that influential observation, so that is all about the influential observation or how to detect and how to take care of the, if there is an influential observation, what to do with that influential observation. And in the last module, because of the time constraint I could not talk about one thing that is called the press statistic, this is not part of this module, but I have some time today. So, what I want is that, I just want to talk about the press statistic now and then I will stop, so the press statistic, so note that this is not a part of this module.

(Refer Slide Time: 47:43)



The PRESS Statistic

$e_{(9)} = 14.7$

i th press residual $e_{(i)} = y_i - \hat{y}_{(i)}$

Large PRESS residuals are useful in in identifying obs. where the model does not fit the data well.

The PRESS statistic is

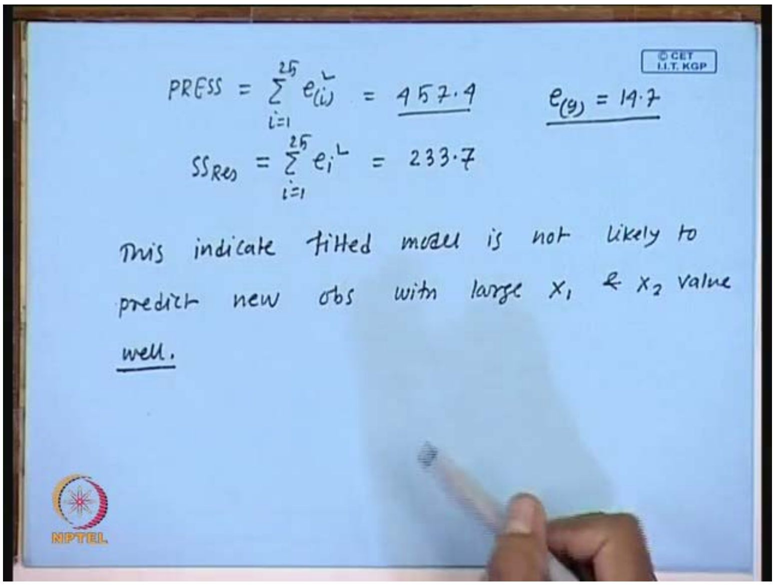
$$PRESS = \sum_{1}^{n} e_{(i)}^2 = \sum_{1}^{n} (y_i - \hat{y}_{(i)})^2$$
$$= \sum \left(\frac{e_i}{1-h_{ii}}\right)^2$$

How well a reg. model will perform in predicting new data.

The press statistic, so this is one thing I wanted to talk in the previous module that is in module adequacy checking, let me just recall the i th press residual, is e i which is equal to I already talked about this y i hat. So, that this is the observed value and this is the fitted value of y I, obtained without the use of i th observation and of course, here large press residual are useful in identifying observations, where the model does not fit the data.

So, if you can recall the 9th observation in the, that delivery time data there e 9 was very liar, very large, e 9 was something like 14.7. And so this indicates that, this fitted model does not, I mean the model does not fit the 9th observation and anyways, so the press statistic is press, which is equal to e i square 1 to n. So, this is nothin, but y i minus y i hat square from 1 to n, and this also can be written as we learned in the previous module this can be written as e i by 1 minus h i i square. So, what this measure is that, it measures how a regression model will perform in predicting new data. So, this is how the press statistic is used and just I want to recall the previous example that delivery time data, there you can check that.

(Refer Slide Time: 51:46)



$$PRESS = \sum_{i=1}^{25} e_{(i)}^2 = \underline{457.4} \qquad \underline{e_{(9)} = 14.7}$$

$$SS_{Res} = \sum_{i=1}^{25} e_i^2 = 233.7$$

This indicate fitted model is not likely to predict new obs with large X_1 & X_2 value well.

The press value is equal to e i square, i equal to I think 1 to 25 that is 457.4 and see the press is a counter part of S S residual, when you consider all the data. So, this is some just e i squre, i equal to 1 to 25, which is equal to 233.7 and as I told you that for the 9h observation in that example e bracket 9 was 14.7. So, e 9 square is almost the half of this press value, so the higher press value indicates that the model cannot perform very well in predicting the new data.

So, specially for the data where the regressor values are large. So, this indicate, fitted model is not likely to predict a new observation with large X 1 and X 2 value, because I do not have the example with me at this moment, I have it. So, (Refer Time: 54:02) this is the 9th observation, so you can see that for here the X 1value and the X 2 values are very large and that is why the e 9 value, and the e 9 value is very large.

So, this indicates that the fitted model is not likely to predict new observations with larger X 1 and X 2 value, so this is what about the press statistics, and also as I told this is not the part of this module. So, in this module we have learned about how to detect an influential observation, and once the influential observation has been detected, what to do with that observation and I have to stop now.

Thank you.