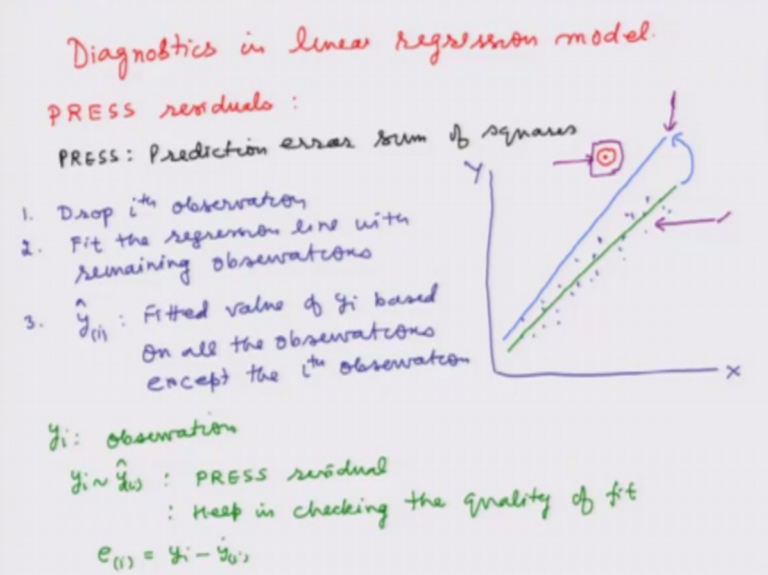
Regression Analysis and Forecasting
Prof. Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology-Kanpur

Lecture-16
Diagnostics in Multiple Linear Regression

Welcome to the lecture you may recall that in the last lecture we had discussed about different aspect of a multiple linear regression model, so now we are going to discuss about different a small topics which are useful in application, so we will basically discuss various type diagnostics in linear regression model.

(Refer Slide Time: 00:37)



First of all we are going to talk about PRESS residual based on that we will define diagnostics which is called as PRESS statistics, first we try to understand what is this actually PRESS, but before that let me tell you the interpretation, the interpretation of PRESS is prediction error sum of squares Now first let us try to understand the problem with the following figure.

Suppose this is my X axis this is my Y axis, and we have some data points like this and here we try to fit a line like this. Now as long all the points lying close to the line there is no issue and the fitted regression line will be a good regression line, now suppose there is one point, which is lying quite away say some where here. Now because of this thing the what will happen, that the

regression line if I try to fit using all the data mean the earlier data and this new data point the regression line may shift to this place

So one can see that the earlier regression line is now shifted in this direction when a value which is quite away from the existing observation is add in the data, now there can be two question that first of all this type of extreme values or this type of unusual values they may be a part of experiment, so in that case we have to use some other statistical tools which can take care of such unusual observation.

The second aspect is that well they are not the part of your experiment, but somehow they are appearing in the data set, so our objective is that how to identify such observations from the given set of data which are unusual and they may be good or they may be bad. So first we are going to discuss on this aspect. So now the next question is how to develop a tool? The idea of developing a tool to identify such observation is very simple.

You can see here that when we have all the observation there is unusual observation we have a line which is given by the green colour. Now when we are trying to add an unusual observation here then the regression line is shifted and this blue color line is the new regression line, so now the idea which we are going to use is the following that suppose we have got a data in which we have got unusual observation.

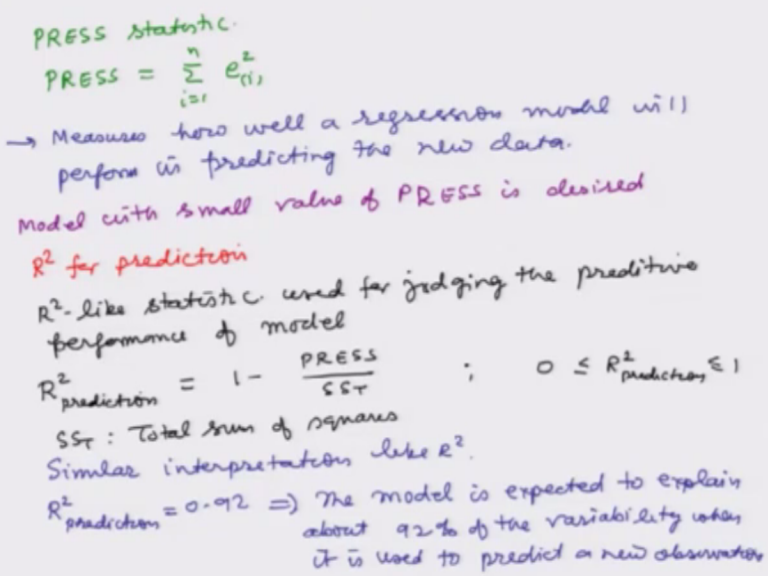
So now if I try to delete this unusual observation from our data set and then we try to fit the regression line then this is going to be the same regression line, which is given here by green colour. So based on this idea we proceed as follows: First step is drop ith observation and in the second step fit the regression line with remaining observations and in the third step using the fitted model, which is obtain after removing the ith observation we compute the predicted value which we are going to denote as y hat and i is written inside the bracket to denote that this is the fitted value of yi based on all the observations expect the ith one.

So now if you observe yi is the observation that we have obtain from the experiment that we have observed. Now if I try to find out the difference between yi and yi this bracket i hat, this

trying to give us a sort of residual, which is indicating the two values of y when this unusual observation was added in the model and when this unusual observation is not include in the model and this is called as PRESS residual.

And this PRESS residuals they help in checking the quality of fit that means whatever model we have obtain on the basics of given set of data this PRESS residual is going to help us. Now we denoted this Press residual as ei, i is written inside the bracket, so this is now we are going to consider has yi yi hat, now based on that we define the PRESS statistic.

(Refer Slide Time: 08:23)



And this is usually donated by PRESS, which is the sum of squares of this PRESS residuals, so now if you try to observe this quantity what are we doing, we are going to delete one observation at a time from the given dataset, we will try to find out the fitted model using ordinary least square estimation or say maximum likelihood estimation and then based on that we will try to calculate the PRESS residuals and this process is going to continue for each observation.

And then we are going to find out this PRESS statistics. The role of a PRESS statistics is this that it measures how well a regression model will perform in predicting the new data. So you can see now we are moving towards the aspect of forecasting and prediction, so this is one statistics that will help us in diagnosing that whether the fitted model is going to work well in case of prediction or not.

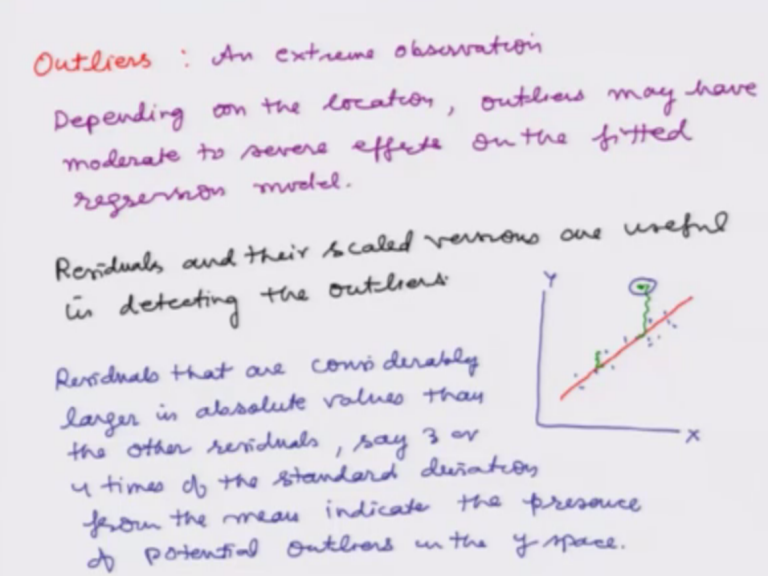
In this case a model with small value of PRESS is desired. Now based on this PRESS statistics we try to define good news of it for the prediction, and we define here R square for prediction. You may recall that in case of multiple linear regression modeling we have defined the coefficient of determination R square and we had use it for judging the good news of a fit of a model.

Now once we have obtain the model we would also like to know what will be the performance of my fitted model when it is used for prediction, so in order to develop a goodness of fit statistics we try to use the PRESS and we define here a R square like statistic and which is used for judging the predictive performance of model, and this is define as R square. Let me write R square prediction, this is equal to one minus PRESS divide by SST.

Where this SST is nothing but the total some of s squares which was obtain from the analysis of variance table, and this value of R square prediction also lies between 0 and 1 and this also has got a similar interpretation like R square. For example when R square is 1 that means the model is really good for the production this is in an extreme case and similarly if R square is 0 that mean model is not at all good for the purpose of prediction.

Now if a we say that R square prediction = 0 point nine two then this value indicates that the model is expected to explain about 92% of the variability when it is used to predict a new observation or in simple words one can say that the model is a nearly 92% good for the purpose of prediction. Next we are going to talk about outliers.

(**Refer Slide Time: 13:53)**



What is an outlier? Outlier is simply an extreme observation, well now once again we will have here two option the extreme observation is a part of our study and this coming in an natural way from the experimental setup or the second thing can, this is an unusual observation, which is not decide and this going to affect our model which is not acceptable. So in case this extreme observation is rising as part of the experiment then we have to think of some other statistical ways to find out the fitted regression model.

In case this is some unusual observation, so were we would like to have some diagnostic statistics and diagnostic test which can tell us on the basis of given set of data that, which of the observations are extreme observation and they are essentially an outlier. So we are going to work on this line of action. Now depending on the location these outliers may have moderate to severe effects on the fitted regression model.

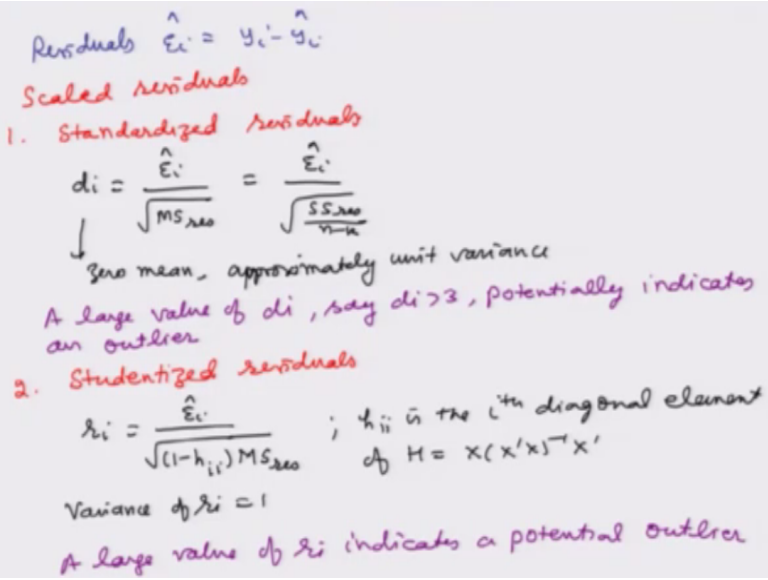
So next question is how to identify them what should we do so that we can develop a diagnostic for identifying such outliers? In such case the residual help us lot, so we are going to consider here the residuals as well their scale version, and then we will see that the residuals and their scaled versions are useful in detecting the outliers. So first we try to understand how this can be done.

We had see earlier that incase if we have some data x and y we try to plot here on XY axis and then we try to switch here a line like this one now suppose there is a some observation, which is somewhere here, so now we try to find out the difference between the observed value of y and fitted value of y. So one can observe that in these points the residual are going to be quite small, whereas in this observation the residual are going to be quite large.

So now looking at these 2 residuals 1 can say that incase if an observation is going to be outlier here then possibly the corresponding residual is going to be larger. So one can say here that residuals, that are considerably larger in absolute values than the other residuals, say 3 or 4 times of the standard deviation from the mean indicate the presence of potential outliers in the y space.

So what are we trying to say here that incase if a particular residual considerably larger, then the question comes how large? So we are saying as a rule of term and question should be 3 or 4 times standard deviation from the mean, incase if this happen so, then possibly that would indicate that the corresponding observation is possibly in outlier. Now when we are talking about the residuals, we had seen that we try to define the residual as epsilon hat which was y i - yi hat.

(Refer Slide Time: 19:19)



Sometime it is useful to work is scaled residual, so now we try to talk about some scaled residual. So we are going to talk about two types of scaled residuals first we are going to talk about standardized residuals. The standardized residuals they are defined as di is equal to epsilon

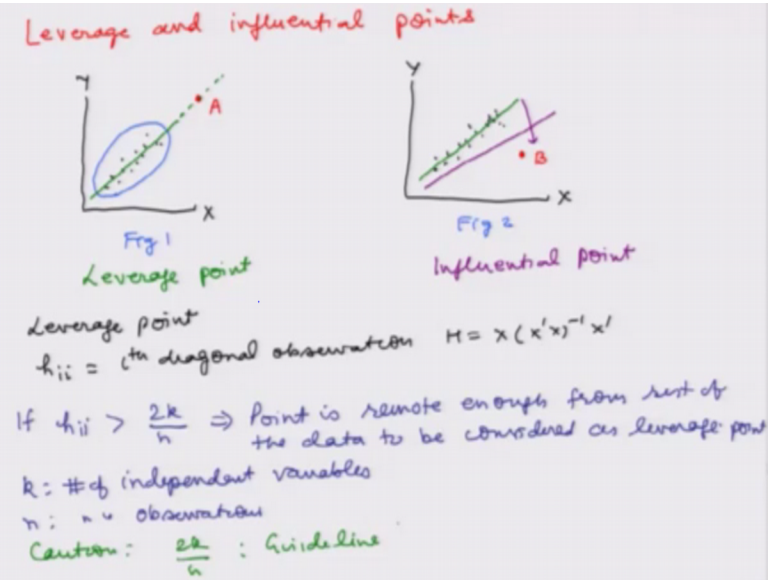
i hat divided by MS res and this same as epsilon i hat divide SS res divide by n - k in the case of multiple linear regression model y = x beat + epsilon with epsilon explanatory variables.

And you may recall that the value of MS res that is mean square due do residual or say SS res that is some of square to do residual that can be obtain from the analysis of variance table, this di's have got 0 mean and they have approximately unit variance. Now this standardize residual helps us in diagnosing a potential outlier. So we say that a large value di, say di greater than three potentially indicates an outlier.

The next is scaled residual that we are going to talk about is studentized residuals this is defined as ri which is epsilon i hat divided by 1 - hii times MS res were this hii is the ith diagonal element of hat matrix h which is x x transpose x whole inverse x transpose, and the variance of ri is equal to here one and this also has got the similar interpretation like di that large value of ri indicates a potential outlier.

These standardize residual or studentized residual they can be very easily calculate when we are trying use as statistical software

(Refer Slide Time: 23:17)



Next topic we are going to discuss is about leverage and influential point, so obviously the first question comes what is a leverage point and what is an influential point, so let us try to make

here to graphics and we will try to explain this concept through the graphic. These are your X axis, Y axis and we have got here a dataset something like this and here also like this. Now we are going to consider here 2 different points 1 point is lying somewhere here.

Let us denote it has point A and there is another point here which is lying somewhere let us try to denoted by here point B. So now if you try to observe supposed this our figure number 1 and this is figure number two. Now, one can observe that in figure number 1, there is a point A which has lying quite far away from the remaining points. The most of the sample observation they are scattered somewhere here and this point is lying quite far away.

But once we try to fit here a line, the line is going to be like this incase if I try to use all the observation except the observation at A. So when I try to include this A point in my line the regression line is not going to be changed, so in this situation this point A is lying quite away from the observation, but the regression line is not affected, so this point is called as leverage point.

Now we consider the figure number two, now here you can see that incase if I try to use the observation expect the observation at point number B then the regression line is going to be obtain something like this and again this point B is a lying quite away from the existing point, but there is a difference, the difference is this that in case if I try to include this point B in the given sample of data, then the new line will be somewhere here.

So what we observe that this point B is trying to attract the fitted regression line towards itself, so this is called as an influential point. So one can see that if a point is a leverage point, then the fit regression line is not going to be changed and possibly based on that other model property are not much changed. Whereas incase if we have got influential point, then the regression is going to be changed and that is attract towards the influential point or towards the direction in which the influential point is lying.

The question is this, these points can be usual point as a part of your experiment or they can be some unusual point and incase of unusual point we would not like to have them in our regression

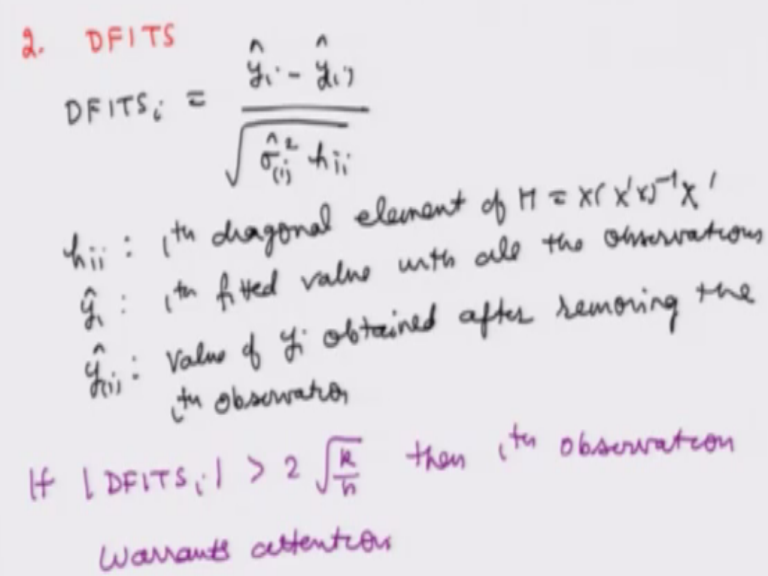
analysis, but the question is that how to identify them whether the point is a good point or a point is a "bad point", so we are going to discuss several diagnostic here possibly they will help us in diagnosing whether a point is a leverage point are it is an influential point

And finally whether we want to retain it in our regression analysis or not. So first of all we try to talk about the leverage point, and let us try to denote say as earlier hii be the ith diagonal observation of the matrix h = x x transpose x whole inverse x transpose. Now the rule is very, very simple if hii is greater than two k upon n, recall that k is the number of independent variables and n is the number of observations.

So incase if hii is greater than two k by n this implies that point is remote enough from rest of the point to be considered as leverage point. So now what we have to do, we have to compute this hii for all the observation and then we have to simply compared by two k upon n and incase if this in equal to satisfies or not based on that I can decide whether a particular point is a leverage point or not

What we have to be little bit cautious, that this limit, this two k upon n right, this is only a sort of guideline, because it depends on the number of independent variable as well as the on the number of observations.

(Refer Slide Time: 30:11)



Now we come on the aspect of a influential observation. In this case we are going to talk about two statistics first statistics, we called has DFBETAS and this statistics indicates how much the regression coefficients changes if ith observation were deleted and this is defined as DFBETAS ji + beta hat j - beta hat ji inside the bracket divide by sigma square i hat, i inside bracket cjj. Now first we try to understand what is the meaning of all this symbols?

You see here beta hat j. it is the estimate of jith regression co efficient using all observations and beta hat ji is obtain like this that eliminate ith set of observations, and obtain the estimate of beta j with remaining observations and similarly sigma square i hat this the value of a sigma square hat which is computed after deleting the ith observation, and the rule now here is very simple that if absolute value of beta sji is greater than two by square route of n.

Then the ith observation warrants examination, similarly there is another statistics which is a called as DFITS, and DFITS for the ith observation is define as yi hat - yi hat inside bracket and sigma square i hat inside bracket and hii. S in this case as usual hii is the ith diagonal element of hat matrix, and yi hat is the ith fitted value with all the observations and yi bracket hat this is the value of yi obtained after removing the ith observation.

And in this case again the rule is very simple that if DFITS i is greater than twice of square root of k upon n then ith observation warrants attention. One thing in this DFBETAS and DFITS we have to keep in mind that the guidelines which we have given, here as in this case it is twice of square root of k by n and earlier twice of square root of n these are only there guidelines, and in practice you may have to look into your data and then you have take a corrective decision.

So we stop here and in the next turn we will try to consider some more diagnostic and some graphical detection procedure, till then good bye.