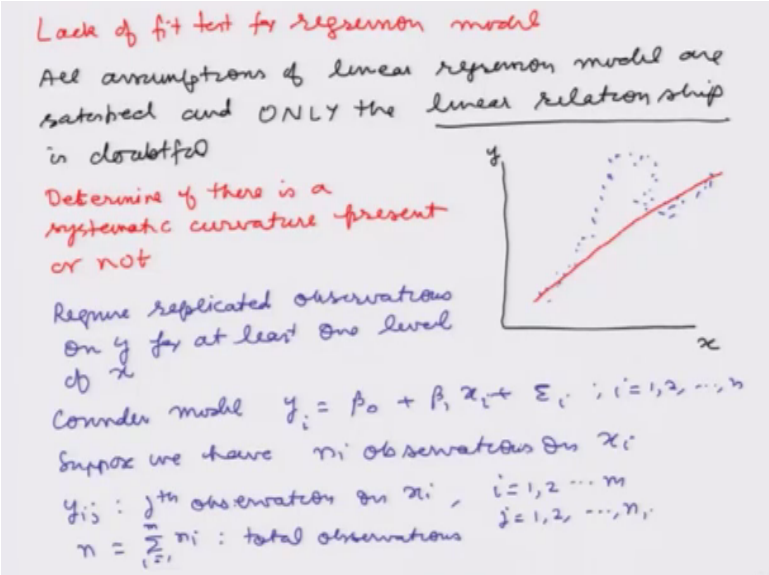
Regression Analysis and Forecasting
Prof. Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology-Kanpur

Lecture – 17
Diagnostics in Multiple Linear Regression Model

Welcome to the lecture you may recall that in the earlier in lecture we had discussed about different types of diagnostic test and diagnostic statistics which help us in checking different types of assumptions and various aspects of multiple linear regression modeling on the basis of given sample of data, continuing in the same line in this lecture we are going to discuss some more diagnostic which helps us in checking the violation of basic assumption of multiple linear regression model and some other aspect.

(Refer Slide Time: 00:54)



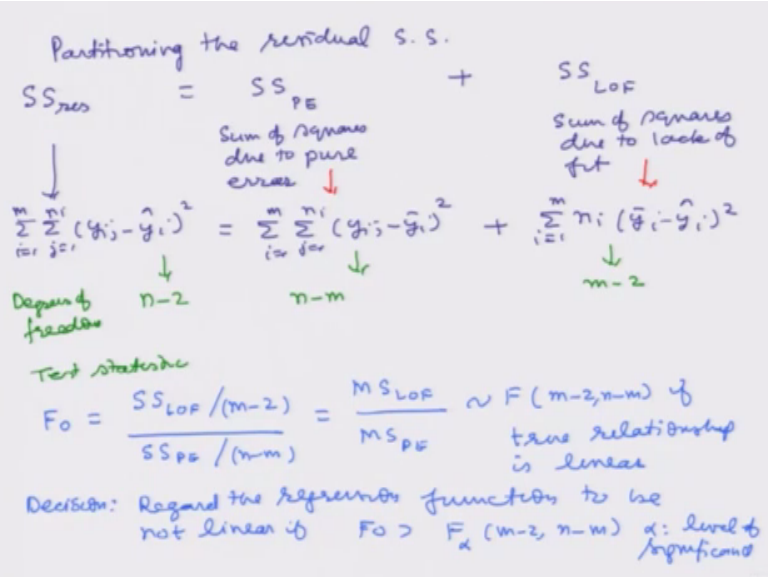
So let us start with the first aspect, and we are going to discuss here lack of fit test for regression model. What do you mean by lack of test? In this case we are assuming that all assumptions of linear regression model are satisfied and only the linear relationship is in doubtful. So we are going to test about a situation where we want to discuss whether the model has got a linear relationship or not.

If try to draw such a situation this may look something like this that we have a situation in which we are going to have scatter diagram, something like this, so in this case if you try fit a

regression line that may go like this. So in this case we are not really confident whether the linear regression model can be fitted or not, so our basic objective here is to determine if there is a systematic curvature present or not. In order to develop test for such a diagnostic require replicated observation on y for at least one level of x.

So we are going to illustrate this test using a simple linear regression model and we consider here a model y = beta0 + beta1 x + epsilon and have availability of n sets of observations which are going to satisfy this model. Now suppose we have ni observations on xi, so now let this yi j denote the jth observation on xi, so i goes around from 1 to m and j goes around 1 to ni and n = here summation i goes around one to m ni this is denoting the total observations.

(Refer Slide Time: 05:01)



So now let us try to develop this test. Now let us try to understand the test, this test is developed by partitioning the residual sum of squares into two components and this is denoted as a SS res this is the residual sum of a squares and this is going to be partition into sum of squares due to pure error than this Is denoted by SSPE and another component which is the sum a squares due to lack of weight.

So this is actually sum of square due to pure error and this is sum of squares due to lack of it. You know that the sum of squares due to residual it is something like I goes from one to m, j goes to from one to ni, yij - y hat i whole square. Yes, you have to keep in mind that we have

observation on y at 2 levels that i and j so the earlier definition of sum of a square due to residual that summation i goes around one to n, yi - yi hat whole square that has to be extended with double summation.

And this can be broken in to something like i goes from one to m, j goes from 1 to ni, yij - y bar i whole square + sum of square due to lack of it and this is given by i goes from 1 to m ni y bar i - y hat i whole square so if you try to see here the sum of square due to pure errors given by this and sum of a square due to lack it is given by this. Well, I am skipping the minor detail, but by looking at this expression you can now see that it is not difficult to obtain.

This is just obtained by adding and subtracting y bar i in the expression yij - u hat I here. Now when we talk about the distribution then this sum a square due to residual this has got degrees of freedom as n - 2 and sum of a square due to pure error this has got degrees of freedom n - m and sum of square due to lack of it this has got degrees of freedom m - two. Why it is here two, because we are estimating to parameters beta 0 and beta1.

Now based on that we can define the test statistics, well again I would say here that I am not considering here the minor details; well there is a strong theoretical proof available for this test. The test statistics is denoted by see here F naught and this is given by sum of squares due to lack of it divided by its degrees of freedom upon sum of a square due to pure error divided by its degrees of freedom, and if you try to see the sum of a square divided by degrees of freedom that are nothing but the mean square due to LOF, the lack of fit and mean square due to pure error.

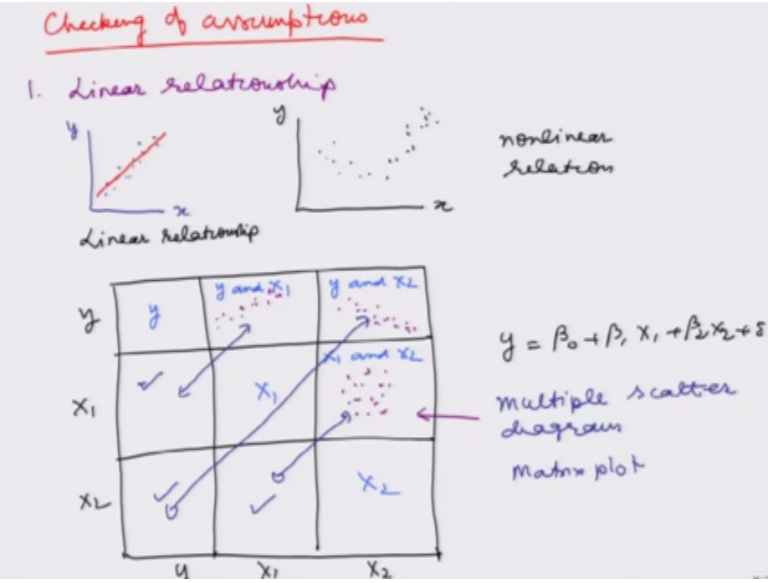
And this follows F distribution with degrees of freedom m minus two and m minus m if true relationship is linear. So in order to compute the lack of it what we have to do, we simply have to compute the statistic F naught and then we have to make a decision here that regard the regression function to be nonlinear if F zero is greater than the tabulated the value of F at alpha level of significance with m - 2 n, n - m degrees of freedom where alpha is the significance.

So now using this test we can conduct or we can verify whether the regression function is following a linear trend or not, and this test essentially is a part of the software outcome, so

whenever we are going to deal with the software I will try to show you later on that the sum of a square due to pure error and sum of a square due to lack of fit they are the two components of the sum of a square due to residual and they are very clearly mentioned.

So now we come to another aspect and we would like to check the assumptions using some graphical techniques.

(Refer Slide Time: 11:00)



Up to now we have discussed the different type of diagnostics which are based on certain statistical format, so first of all I would try to test what about the linear relationship. Once we have a sample of data the first step is to verify whether the data is following a linear relationship or not and this gives us the first hand information. So in the case of a simple linear regression model it is not difficult as we have earlier discussed that we simply need to plot here a scatter diagram between x and y.

And if we are getting a trend like this one, yes one can see here that possibly a linear model can be fitted over here and the other hand in case if we are getting a relationship like this one in which we have got a curvilinear relationship this shows that well here possibly a nonlinear relationship is more suitable and this indicates the linear relationship. Here you have to be little bit careful.

For example in this particular case of nonlinear relationship one can always transform the model so that it becomes a linear relationship also, but that will be our secondary step and in another case for example when we are trying to deal with the multiple linear regression model then we try to use the multiple scatter diagram. So I try to draw here a multiple scatter diagram with when we have suppose, two independent variables.

So in the case of multiple scatter diagram it tries to give us a sort of a scatter diagram among all sorts of variable. For example, here if I take see here y, here x one, here x two and here y, x one and here x two so that here I am trying to consider here a model like y = beta0 + beta1 x 1 + beta2 x 2 + here epsilon. Now in this case we will try to see on the diagonal elements we are going to get a relationship between y and y x1 and x1 x2 and x2 which has no meaning for us.

Now here if you try to see, here in this block we will be getting a relationship between y and see here x1 and here I will be getting a relationship between y and x2. In this block I will be getting here a relationship between x1 and x2. Now we try to look at the scatter diagram for example it is giving us like this diagram and it is giving us a diagram like this one. So this is possibly showing that the linear relationship between y and individual independent variable see here in this case x one and x two has got a linear trend.

So we can expect that the joint relationship between y and x1 x2 may also be linear, one point where I would like to make all of you cautious is that these are only the indicative, because we are essentially interested in finding out the joint relationship of x1 x2 with respect to y or when we have k independent variable then the joint relationship of x1 x2 x k with respective y.

Up to three dimension possibly I can make a plot, but when am going to more than two independent variables like a three e independent four independent variables and we try to plot them with respective y then making four dimensional plot or five dimensional plot or say higher order plot it is not possible. So we have to depend on this two dimensional plot and by drawing the information this two dimensional plot we try to conclude about the joint relationship of all independent variable with respective y.

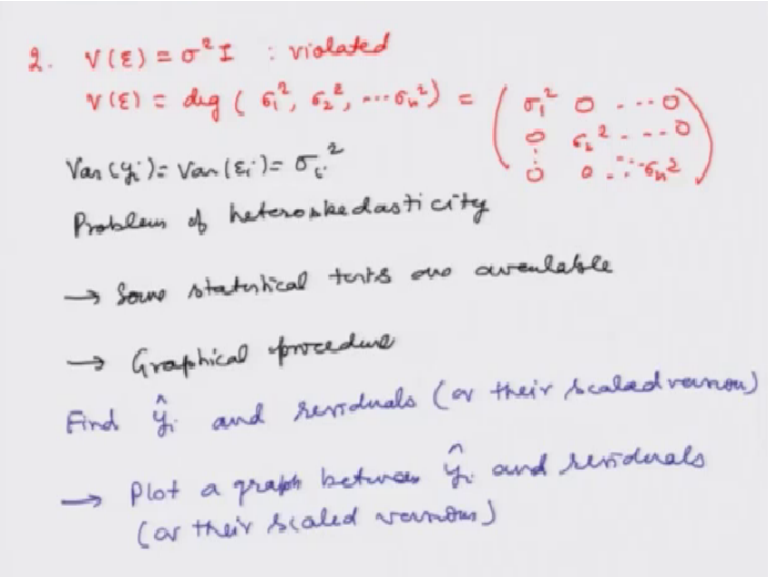
So these only provide us a sort of indication whether the linear relationship can be fitted with the given set of data or not. Now in this block you will see we are going to get a relationship between x one and x two and what we expect that they are not going to show us any trend, because we have assumed that all x one x two x k they are independent by assuming that rank of x matrix is equal to k.

So later on I will try to discuss how to verify this thing that whether all x1 x2 x k are independent or not, but here also I can just inform you that by looking at this type of graphics or some part of this multiple scatter diagram we can also conclude whether our x1, x2, x k are independent or not so this is essentially a multiple scatter diagram, and sometime this is also called as matrix plot.

Now what about these 3, 1 this one, this one and this one, so we can see that this is symmetric diagram, symmetric diagram means that whatever is the relationship between y and x1 that is the going to be the same relationship between x one and y, so this and this will have a same picture and similarly this will also have a same picture and similarly this and this they are also going to have a same picture with suitable modification in the trend.

So this is about the checking of assumption of linear relationship between say x and y.

(Refer Slide Time: 17:51)



Now we come to another assumption, one of the assumption we had made that the covariance matrix of epsilon sigma square i and suppose now we assume that this is being violated, and we assume that variance epsilon is given in the form of a diagonal matrix of sigma1 square, sigma2 square, sigma n square so this is like this sigma1 square, sigma2 square, sigma n square, and all of diagonal elements are 0.

Indicating that epsilon1, epsilon2, epsilon n they are independent but they have got different variance in this case you can see that the variance of yi is same as the variance of epsilon i = sigma i square. So now the first question comes that how such situation can be identified in real life? So one factor can be your own involvement with experiments for example if I take a simple example that supposes somebody starts leaning the typing.

Now on the first day, he is given a sheet of paper and he is asked to type it on a typewriter. On the first day he has no idea so possibly he will make suppose hundred mistakes per page, and he is going to take one hour a time. Now he practices say for about a week and after a week he repeats the same test that he given a sheet of paper and he has to type it, so obviously after a week's practice the number of mistakes are going to be lower than on the first day.

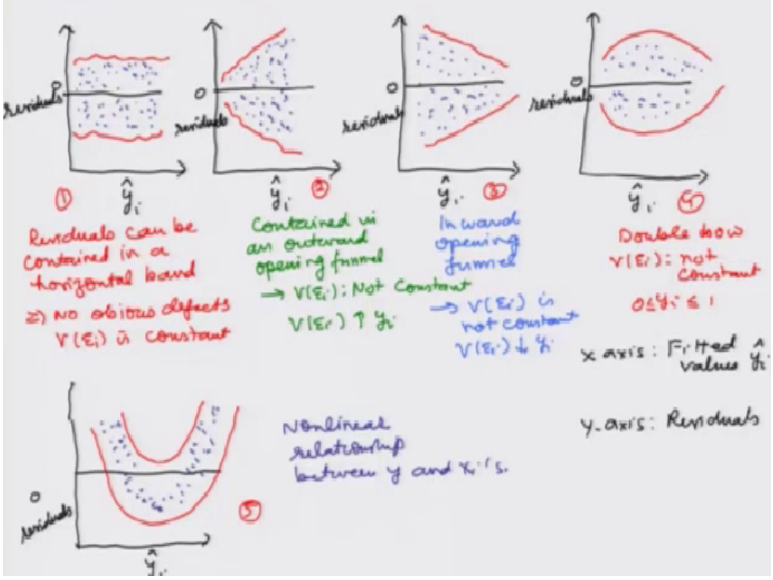
And similarly if he continues to practice more the number of mistakes are going to be lesser so in this case, in case if I try to see that the variability of the number of mistakes that is decreasing as the time is increasing or the time of a number of practicing hours are increasing. Similarly if I take another example that as the salary increases the variety in the consumption of food that also increases.

For example if somebody has a lower salary, possibly he would try to have a simple food at his home as the salary increases the variety in the food increases and possibly when the salary becomes more they may also like to dine out in a good restaurant and so on. So in this case we can see that the variability in the food consumption that is increasing with respective salary. So in both the cases the variance does not remains constant in the first case of typing the variance is decreasing.

And in the second case the variance is increasing, in case of a salary versus the food consumption. This problem is actually called as problem of hetrokedasticity so now incase if I want to know whether my data has got a hetrokedasticity problem or not so we have two option there are some statistical tests are available but all such tests they are based on one strong assumption that the problem of hetrokedasticity is being caused by a particular reason.

Anyway we are not going to discuss those analytical tests, but we are interested here in diagnosing such a problem using some graphical procedure. So in this case what we try to do that we find the fitted value yi hat and we find the residuals, in this case we can find either the residuals or their scaled version and in the second step we plot a graph between yi hat and residuals or their scaled versions. Now depending on the situations, we can have different types of pattern and based on that we try to take a proper decision.

(Refer Slide Time: 23:04)



So now I have plotted here some type of pattern, so one you can first have a look over this slide on the X axis I am plotting yi hat and on the Y axis I am plotting the residuals and I have made here five different types of pattern and based on that I will try to show you that how do we conclude about such a thing, so now let us try to come to the picture number one. Now in this case you can see here that the blue dots that are indicating the observation between yi hat and residual.

They can be contained in a sort of horizontal band, so here in this case residuals can be contained in a horizontal band or I would say that observations are fluctuating randomly inside this band and this implies that there are no obvious defects and variance of epsilons i is constant, similarly in the second case if you see we have a trend in which I can contain all the points in a sort of outward opening funnel.

So in this case all the points can be contained in an outward opening funnel, and in this case we can conclude that variance of epsilon i is not constant and we further conclude that variance epsilon I is increasing with observations on yi, this our figure number2. Now similarly in the figure3 this is just opposite to the figure number2, and in which all the points can be contained in an inward opening funnel.

So in this case all the points can be contained in an inward opening funnel and in this case also we can conclude their variance of y epsilon i is not constant and it is just opposite to figure number 2 that variance of epsilon i is decreasing with yi. Similarly in figure number four we can see here that all the points they can be contained in a sort of double bow and this also indicates that variance of epsilon i is not constant.

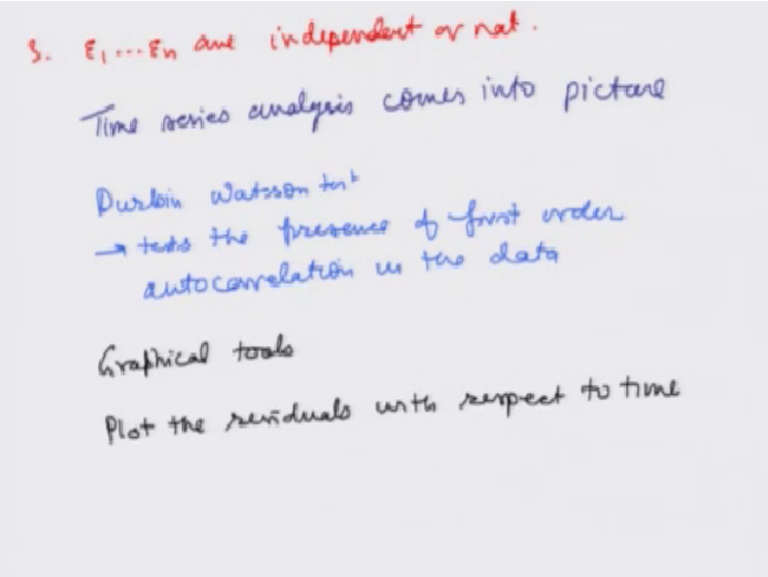
And it is indicating that the variance of y is first increasing and then after that it is decreasing, so possibly indicates that may be yi is variable in which the values are lying between zero and one and one possibility is that this can be a binominal or Bernoulli distributed random variable. Similarly on the fifth graphic, we can see that all the points can be contained in a sort of nonlinear band.

So in this case we can conclude that there is a nonlinear relationship between y and say here xi's and obviously that we conclude here about the variance of yi so these are some graphical procedure which can help us in testing different types of situation where one can conclude if the variance of epsilon i remains constant or not, and these graphics can be repaired using any statistical software.

One thing I would like advice that in real life whenever we are trying to interpret such figures it requires some experience also because these figures may not exactly be like inward opening funnel or say outward opening funnel or bow by adding or deletion of some of the points the curvature may change, so we need to have some practice and some experience before we can draw correct statistical inferences from this pictures.

After this let us try to consider another situation in which we are going to talk about that we want to verify the epsilon one, epsilon two epsilon are independent or not.

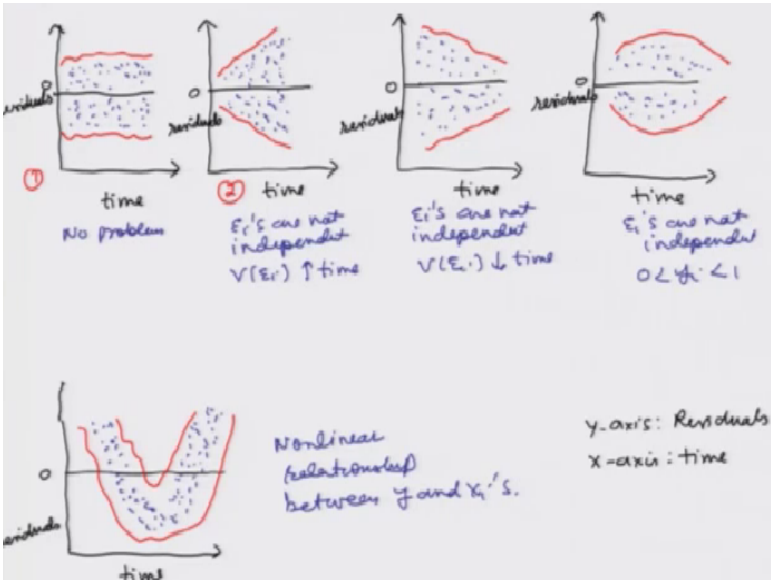
(Refer Slide Time: 28:48)



So now these types of thing they are quite possible in practice and in such a case the entire time series analysis comes into picture, and all sorts of concepts like autocorrelation, autocovariance they come in to picture. So it is not really possible for me to discuss all the details over here, but certainly I would try to give you some basic guidelines. The first thing is that there is a test what is called as Durbin-Watson test.

And this Durbin-Watson tests the presents of first order autocorrelation in the data and similarly we have some other test also we are more interested in knowing about some graphical tools here. In order to construct here this graphical tool we try to plot the residues with respect to time, and we have a set of similar pictures that we have considered in the earlier case and based on that we try to take some correct statistical inferences.

(Refer Slide Time: 30:54)



So if you try to see I have made here five figures and based on that you can see here, but in the figure number1 all the points can be contained in a horizontal bank and here on the X axis I have considered the time and in the Y axis I am considering the residuals or even their scaled versions can also be considered. In the figure number two we can see here that like in the earlier picture all the points can be contained herein a sort of outward opening funnel.

And in this case all the points can be contained in an inward opening funnel in this case all the points can be contained in a double bow shaped and in this case all the points can be contained in a nonlinear band. So similar to the conclusion that we had drawn in the case2, we can also draw here the same conclusion for example here I would say that there is no problem and this type of picture when all the points can be contained in a horizontal band this indicates that all the epsilon one, epsilon two, epsilon n, they are independent of each other.

In this case I would say that there is a problem and epsilon one, epsilon two, epsilons i are not independent and in this case rather we can conclude that variance of epsilon i is increasing with respective time and the similar conclusion can be drawn in this case also when all the points are contained in an inward opening funnel that epsilons i's are not independent, but variance of epsilon i is decreasing with respect to time.

And similarly here also the epsilon i's are not independent, but rather we expect that yi's are lying between0 and one and this can possibly be denoting a Bernoulli or binominal distributed random variables and in this case we conclude that there exist a nonlinear relationship between y and xi's and these graphics again can be obtained using any statistical software. So now we stop here and in the next lecture I will try to discuss about the diagnostic tests using graphical procedure for some other assumption like normality, till then good bye.