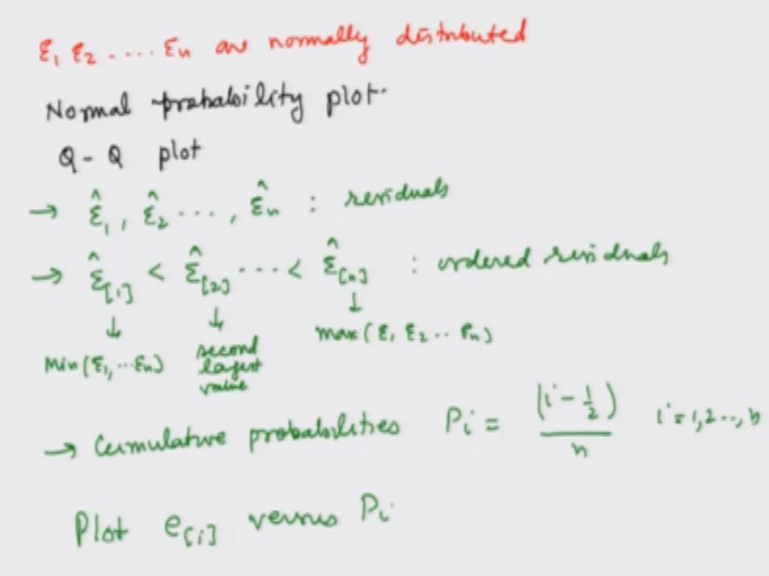
Regression Analysis and Forecasting
Prof. Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology-Kanpur

Lecture: 18
Diagnostics in Multiple Linear Regression Model

Welcome to the lecture in the last lecture we had discussed some graphical tools for diagnostic the violation of assumptions of the linear regression model. Now continuing on the same lines we will discuss here another violation of assumption, in any linear regression model we assumed that all the random errors, epsilon1, epsilon2 epsilon n, they are normally distributed.

(Refer Slide Time: 00:37)



And suppose we want to test this assumption, now what is the problem? The problem is that you are assuming the normal distribution for the entire population and we have got here only a small radiators set on xi and yi, so we would like to verify this assumption using some graphical procedure, and we have to use only the data on xi and yi is whatever are available to us.

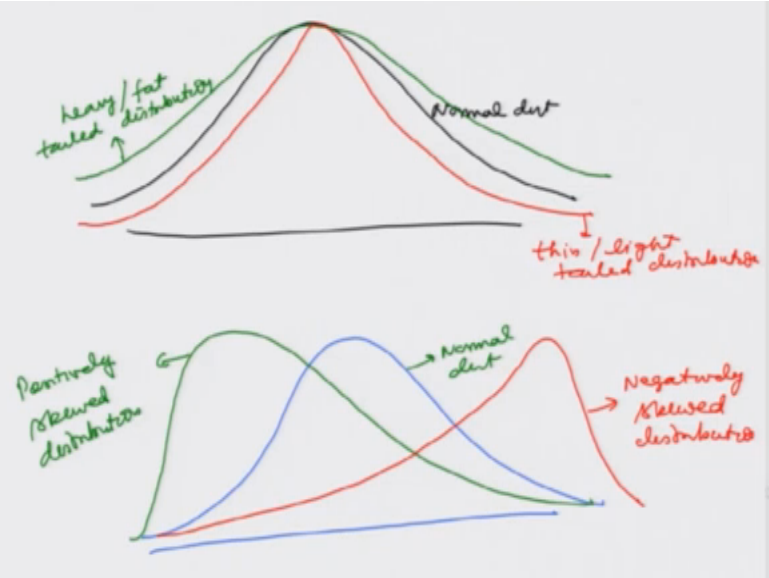
So this can be done using the normal probability plot and this is also called as q-q plot, q-q plot means quantile-quantile plot. So in this case what we do is the following, that in the first step we obtain, we fit a model and we obtain the residuals. In the second step, we try to order these residual and we denote them something like epsilon hat and 1, 2, 3, 4 up to n inside a

square bracket and the meaning of this thing is that, this is the minimum value among all epsilon1, epsilon2, epsilon n and this is the second largest value.

And similarly this is the maximum value between epsilon1, epsilon2, epsilon n, so these are essentially our ordered residuals. Then in the third step we try to find out the cumulative probability and the cumulative probability for the ith observation, this is completed by this expression i-1 over 2 divided by here n, i goes from here one to n. I would like to inform you here that some software may use some other value instead of 1 by 2.

So that may be i- something divided by n, but our basic idea is to compute the cumulative probability. Now then we try to plot the order residuals versus pi that is the cumulative probability and when we try to do so, we can have different types of situations and before we try to describe those situations let me try to explain you the concept of heavy tailed distribution light tailed distribution positively skewed and negatively skewed distribution.

(Refer Slide Time: 04:01)



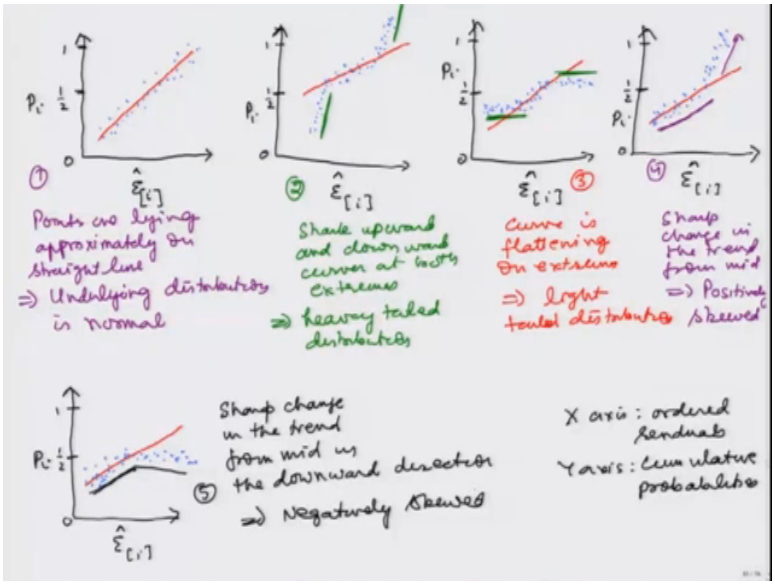
Suppose I try to draw this normal distribution here, suppose this is our normal distribution. Now if there is other distribution whose tails are fatter than the normal distribution and the curve is like this means all other properties like symmetry and everything is maintained. So you can see here that this distribution as a tail, which is heavier than the tails of normal distribution and so this is called as heavy or fat tailed distribution.

Similarly if I try to take another distribution whose tails are thinner or lighter than the tails of the normal distribution then this is called as thin or light tailed distribution and similarly on

the other hand if I try to draw normal curve, which is here symmetric and suppose there is another curve which is more scatter on the left hand side something like this so this is our normal distribution which is symmetric and on mean and this is called as positively skewed distribution.

Similarly if I have got a curve which is more scattered on the right hand side something like this, so this is called as negatively skewed distribution so using that q-q plot we can identify weather the observations are coming from a distribution which is a heavy tailed distribution, light tailed distribution or normal distribution and also we can find out whether the parent distribution is positively skewed negatively skewed or it is symmetric.

(Refer Slide Time: 06:22)



So I have plotted here five figures and I would like to explain you that how we are going to interpret these things, so here you can see with this blue dots I am plotting the observations between ordered residuals epsilon hat i and cumulative probability pi and with the red line I am trying to denote that what is the fitted line, so now depending on the scatteredness of blue points around the fitted line we can make different types of conclusions.

For example here in this case of figure number here one, we can see here that the points are lying approximately on the straight line. So this implies that there is no problem and the underlying distribution is normal or in simple words the observation which we have got y1, y2, y n from a normal distribution and similarly when we come to here figure number here 2.

We can see here that there is a sharp upward and downward curve at both extremes, you can see here that on this extreme there is a sharp trend and here another sharp trend. So this implies that the observations are coming from a distribution which has got a heavy tails or observations are coming from a heavy tailed distribution. Now similarly in the figure number three one can observe that the curve is flattening on extremes.

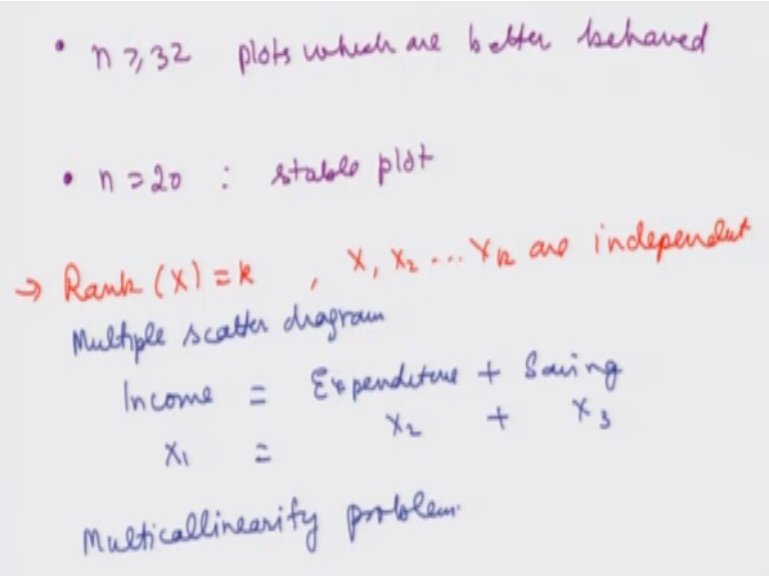
How you can see here that for example if you try to observe over here, here the curve is really flatting and this implies that the observations are coming from a light tailed distribution. Now in this figure number four we can observe that all the points here are initially line close to the straight line and suddenly from the mid there is sharp change in this directions, so one can observe here that there is a sharp change in the trend from mid.

This indicates that the observations are coming from a distribution, which is positively skewed and similarly in this figure number 5, one can observe that the points are initially lying close to the straight line and suddenly there is a sharp change in the trend from the mid but in the downward direction. In this case there is a sharp change in the trend from mid in the downward direction.

So this implies that the observations are coming from a distribution which is negatively schemed okay so now from this figures now one can conclude about the parent population weather this is normal or say heavily tailed or say light tailed or it is a positively skewed population or a negatively skewed population and this figures can be plotted using any statistical software and they are available by a click.

One thing again I would like to emphasize that in these cases one need some experience and practice to correctly infer from the plot about the parent population distribution.

(Refer Slide Time: 11:26)



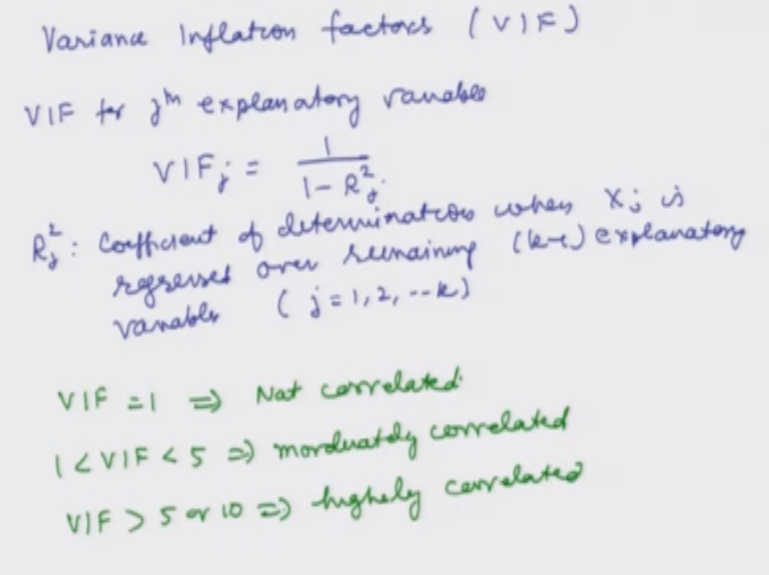
But there are some thumb rules that if I try to take the sample size which is more than 32 this produces a plots which are, I would called better behaved and if I try to take here n=20 observation this usually produces a stable plot, but if we have observed number of observation is smaller than twenty then it is difficult to obtain a good statistical plot. This was all about the normality of the distribution.

Now we come on the last assumption and we would try to see how to verify it. So one of the basic assumption what we had made that rank of x=k, this means all x1, x2, xk they are independent. One option is to check whether they are independent or not is to use the multiple scatter diagram, that we had discussed in the earlier lecture when we were trying to plot the multiple scatter diagram between y x1 and x2 in a linear regression model with two independent variable.

So in case If we are getting a random pattern among all x1, x2, x3, x4 pair wise then that would indicate that they are independent and this type of situations can arise in practice for example if somebody is considering three variables income, expenditure and saving. So he or she might be taking them as x1, x2, x3, but if you try to observe there is a relationship that x1=x2+x3.

And this case we have a problem and this problem is called as multicollinearity. This multicollinearity problem is one of the problems, which has not got a satisfactory solution up to now, but there are different techniques which are available and they can help us in diagnosing the problem of multicollinearity in the data.

(Refer Slide Time: 13:55)

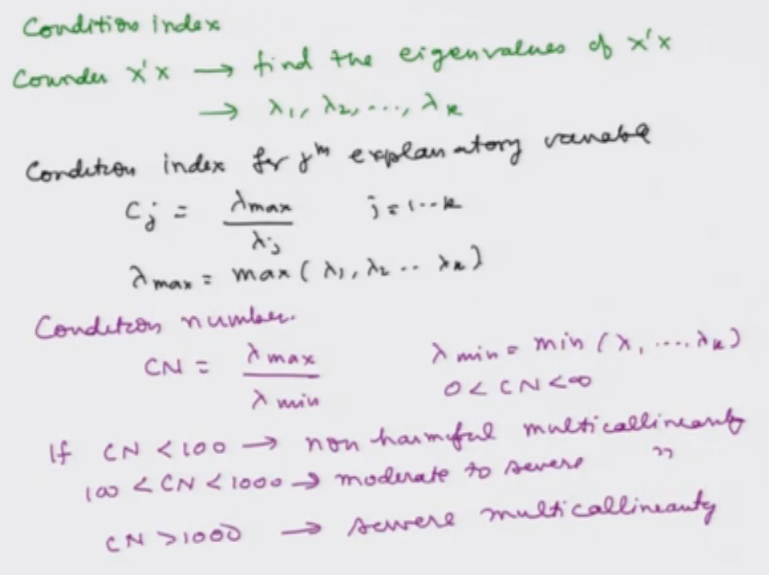


So one of the popular diagnostic to test about multicollinearity is to use the variance inflation factors and they are briefly denoted as VIF. So this VIF for jth explanatory variable is defined as like VIF j is equal to one over one minus R square J, where R is square j is the coefficient of determination when the jth explanatory variable xj is regressed over remaining k-1 explanatory variable, j goes wrong one to k.

So now this variance inflation factor they are is easily available in any software outcome and they are part of outcome of a regression analyses and they have the following interpretation that if we are getting variance inflation factor which =1 this indicates that the explanatory variables are not correlated. In case if we are getting a value of variance inflation factor between 1 and 5 this indicates that the variables are moderately correlated.

And if you are getting g the value of variance inflation factor say more than 5 or 10 this indicates that the explanatory variables are highly correlated.

(Refer Slide Time: 16:02)



Similarly there are some other popular diagnostics which are like as condition index and condition numbers. So in order to find out the condition index what we try to do here that we consider the matrix, X transpose x and then we try to find out the eigenvalues of X transpose x and suppose this eigen values or characteristic roots they turns out to be lambda1, lambda2, lambda k. Then based on that we define the condition index for jth explanatory variable cj = lambda maximum upon lambda j.

And j goes from one to k and where lambda maximum is the maximum value among lambda1, lambda2, lambda k, and similarly we define the condition number as CN, which is equal to lambda maximum over lambda minimum where lambda minimum is the minimum value among lambda1, lambda2, lambda k. Now how to interpret it you can see here that this condition index is lying between 0 and infinity.

So this condition number and condition index they have got a similar interpretation, so we can conclude that if condition number is coming out to be less than hundred then this would be indicating that the explanatory variables are not correlated or they are correlated very little, and in this case we can call that we have non-harmful multicollinearity in the problem.

Similarly if condition number is lying between hundred or say one thousand this possibly indicates that the explanatory variables are moderately correlated and this case there is moderate to severe multicollinearity and if condition number is greater than one thousand this possibly indicates that there is a severe multicollinearity. Now again I would say that these

guidelines or these numbers they are only indicative and they are not something like a very hard and fast rule.

So now we have finished all the diagnostics test whatever we can do in the given time frame and now in the next lecture I will try to emphasize on the use of software and after that we will come on the forecasting, till them goodbye.