Multivariate Procedures with R

Prof. Shalabh

Department of Mathematics and Statistics

IIT Kanpur

Week – 09

Lecture – 42

Goodness of Fit and Testing of Normality

Hello friend, welcome to the course Multivariate Procedure with R. So, you can see up to now we have discussed different aspects of multiple linear regression model and whenever we try to take a decision that we have to fit a statistical model using the regression analysis. First we try to judge whether a linear model can be fitted or not. So, you have seen that we have used different types of plots, scatter diagram, matrix scatter plot and there are different ways, different types of graphics, two dimension, three dimension etc. We try to use and we try to finally take a call whether a linear regression model can be fitted or not. Now, if you try to recall that when we considered the matrix scatter plot or even the scatter plot and we want to fit there a linear model, there is a linear trend, but do you think that all the data points are lying exactly on the same line? No.

As a statistician I can see that there is a linear trend, possibly a linear equation can be fitted, but some points are above the line, some points are below the line. Ideally what should happen? That all the points should lie exactly on the same line, but this is not happening. So that means if you have two data sets in which in one set the points are lying close to the line and in another set the points are lying quite away from the line, what do you think about the goodness of fit? In which of the case you will get a model which is a better fit? Definitely the case in which the points are lying closer to the line that will give us a better fit. But the question is now I can see it in the plot, but how to quantify it? Now if I try to extend this question whenever a real data set comes then it does not follow always a linear trend.

If you try to even plot a matrix scatter plot between y and individual xi's all the plots are not exactly linear. And then sometime we try to make different type of transformation to linearize the data. Sometimes we take log transformation etc. So which makes the data to be compatible with the fitting of a linear regression model? So for the same data set different people can fit different types of linear models, but the ultimate question is which of the model is good? So in order to test or to know the goodness of fit of a model we have a statistics which is called as R square statistics or coefficient of determination which essentially tries to compute the degree of linearity or degree of linearity to the correlation coefficient and they try to indicate that in which case the correlation coefficient is less or higher. So in this case what is really happening? We have one variable y which is the outcome variable and we have here a group of variable x1, x2, … xk and we want to see the relationship between y and this group of variable is linear or not.

So that means I want to find out the correlation coefficient between a variable y and a group of variable x1, x2, … xk. So this type of correlation can be found through the concept of multiple correlation coefficient. So this concept of multiple correlation coefficient is used in the setup of multiple linear regression model to just the goodness of fit of a fitted model. So in this lecture we are going to talk about it. The next question we are going to address in the lecture is that you can have seen that all your test of hypothesis analysis of variant confidence interval estimation and many statistical technique they depend on the basic assumption that your data is coming from a normal population.

If there is a deviation that the data is not exactly coming from the normal population then the tools which we are going to use here which are developed for the normal population may not work or there can be some deviation in their statistical inferences. So whenever you get the data the first observation or the first decision you have to make does this data coming from a normal population how to get it done. So for that we have some graphical techniques yes we have some analytical techniques also like goodness of fit statistics but we are going to use here the Q-Q plot to judge whether my data is coming from a normal population or not. Definitely in real life it is very difficult to find out a data which is 100% normal there will be some deviation and using this normal Q-Q plot we can judge that how much is the deviation from the normality and if normality is not followed then what are the different type of distribution which possibly the data is following based on that we can modify our statistical technique which we are going to use. So that is what we are going to do in this lecture.

So let us begin in this lecture and try to understand first the goodness of fit that how to measure the goodness of fit in a multiple linear regression model which is fitted using the ordinary least square estimator or the maximum lateral estimator. So let us begin our lecture. Okay so now first we talk about goodness of fit and then we will talk about the concept of testing of normality in the setup of multiple linear regression model. Right so as I said that if you want to judge the goodness of fit so if I try to take here two data set like this one in which they are like this and another data set where the variability is here more like this and in both the cases you can find out the similar straight line but definitely in the case number 1 the points are lying more closer to the line than in the case number 2. So I can say that in the first case the model is fitted better than the case 2 and if you try to see we are trying to find out the degree of correlation between X and Y in some way the one observation is on X axis and another set of observation is on the Y axis.

So but now we are going to find out the correlation coefficient between a variable Y and a set of variable X1, X2, … Xk. So this is achieved by the concept of multiple correlation coefficient. Right so let R be the multiple correlation coefficient between Y and the group of variable X1, X2, …Xk. And if you try to find out the square of this multiple correlation coefficient this is indicated by R square and this is popularly called as coefficient of the deuterization. Right so the value of R square commonly describes how well the sample regression line fits through the observed data.

And this is treated as a measure of goodness of fit of the model. I am not saying this is the only measure there are other measures also like as AIC that is Akaike information criteria, BIC that is Bayesian information criteria, but I am going to talk about here this coefficient of the determination. Yeah, AIC and BIC can also be obtained from the in the R software. So in this case we are going to consider here a model with intercept term. This is the prime condition.

If the model does not have intercept term remember if there is no intercept term in the model in the model do not use R square which we are going to consider here. This R square definition what we are going to consider here is valid only when there is an intercept term in the model otherwise you have to use something else. So we are assuming here that there is a intercept term here in the model and then we have here 3 - 1 variables X2, X3, … Xk. So now this R square is defined here as say Y transpose X into X transpose X whole universe X transpose Y divided by Y transpose Y. And if you try to recall the definition what we have used in the concept of analysis of variance the quantity in the numerator is simply sum of square due to regression divided and the quantity in the denominator is sum of square due to total.

$$
\begin{aligned}
R^2 &= \frac{y'X(X'X)^{-1}X'y}{y'y}, \quad 0 \le R^2 \le 1 \\
&= \frac{SS_{regression}}{SS_{total}} = 1 - \frac{SS_{res}}{SS_{total}} \\
&= 1 - \frac{e'e}{\sum_{i=1}^{n}(y_i - \bar{y})^2}
\end{aligned}
$$

And that is correct also if you try to recall you this total variation was divided orthogonally into two parts sum of square due to regression and sum of square due to error or say residual. So definitely you would like to find out a model where the random error is minimum and that is how you had obtained it using the principle of ordinary least square estimator. So definitely ideally what should happen that sum of square due to residual should be 0 and then in that case a total variation should be explainable only by the fitted line. So this is the concept of analysis of variance because of which R square can be obtained from the R square from the analysis of variance table also. So now using this result I can also write down here as say SSregression here is TSS or say SStotal - sum of square due to residual divided by SS total and this can be written here as say 1 - SSres divided by SStotal.

Now SSres that can be obtained by here the E transpose E where this is actually summation ei square and ei is the i-th residual that we have obtained using the ordinary least square estimator y - y hat. This was your E vector. And then total sum of square can be obtained by here summation yi - y bar whole square. So you can see here this is the definition which is again depending on the ordinary least square estimator. If you try to see this component here X transpose X whole inverse X transpose y is simply here b.

So you can see here this is how one can say that if you choose a good estimator which is explaining the variation in y variable through the fitted model then this R square should be higher otherwise lower or the value of R square will depend on the estimator that you have used to split the model. So this R square measures the explanatory power of the model which in turn reflect the goodness of it of the model. That means how well the model can explain the variation in y. In statistics we are always interested in understanding the variation due to the random factors. So this R square reflects the model adequacy in the sense that how much is the explanatory power of explanatory variables.

That means if your explanatory variables are good they will explain better and your model will be good if not then the opposite or the reverse will happen. And as I said this is very important for you to learn and remember if constant term or the intercept term is absent in the model then R square cannot be defined. And in such cases although I squared value but still R square can be negative. Well the next question comes what to do in such a case then some ad hoc measures based on R square for regression line through origin have been proposed in the literature but I am not commenting anything on their goodness that how well they are trying to explain the phenomena. Now the limits of this R square are between 0 and 1 and it has got an interpretation that is what you always have to keep in mind.

When R square = 0 this indicates that the model has got the worst fit. It is not well fitted at all. It is very bad model. Worst model one can say. When R square = 1 then this indicates that this is the best fitted model. And if there is any other value of R square between 0 and 1 for example if I say it is coming out to be R square = 0.95 then it indicates that 95 percent of the variation in the response y is explained by the explanatory variables x1, x2, … xk. Or in simple words I can say that the model is nearly 95 percent good. So similarly if you try to take any other value of R square between 0 and 1 that is going to indicate the adequacy of the fitted model.

Now if you really want to know that in a given setting whether R square = 0.65 is going to indicate the good model or a bad model that decision I leave it on the statistician who is trying to handle the data. There are many constraints on which we have to take such a conclusion. But this is the way we try to interpret the value of R square. And then R square has a drawback also that if more explanatory variables are added to the model then R square increases. Why this is drawback? Even if you try to add some irrelevant variable which have no relation with y even then the R square will increase and that may indicate as if the model is getting better but it will not happen.

So this so adding irrelevant variables in the model will give an overly optimistic picture because R square will be high you will be very happy that my model is becoming better which is actually which will not be the case right. So in order to correct it that this R square should not increase incorrectly when there are some variables added in the model we have defined adjusted R square which is denoted here as R bar square or it is written here as a Adj R square and that was also given in the software. And this is defined like this 1 - SS divided by n-k divided by SStotal divided by n-1. If you try to see these are the degrees of freedom which we have discussed earlier in the analysis of variance and this can be written as 1- n-1 upon n-k into 1 - R square. So you can see that here there is a one to one relationship between R square and R bar square also and R bar square is called is called as say adjusted R square and usually you will see in the software that adjusted R square will always be less than say R square right.

$$\bar{R}^2 = 1 - \frac{SS_{res} / (n-k)}{SS_{total} / (n-1)}$$
$$= 1 - \left(\frac{n-1}{n-k}\right)(1 - R^2).$$

So do not get confused but you would try to see both the values if there is a big difference between the values of R square and R adjusted R square then it is indicating that there is something wrong and then you have to be careful watchful. But my basic message at the end is that before we try to use it over a set of data that R square is a good measure but you have to handle it very carefully it has certain limitations also right. So now I try to take here the same example which I have considered in the past that we have the data of 20 students and we have collected the observation on their marks indicated by Y which are going to believe that they are dependent on the X1 number of hours of in a week of study X2 which is the number of assignments submitted in a month and X3 number of hours of play and this data is here given like this. So the first step data indicates that student number 1 has study 34 hours in a week he has submitted 3 assignment and the student has played 15 hours in a week and the student has got 180 marks out of 250. And similarly the student if you try to look at Y20 that means 20th observation then the student has studied 34 hours in a week submitted only 1 assignment in a month and they played say 19 hours in a week and got 197 marks out of 250.

So now we already have obtained the regression model but I would like to now you show you that how to get this R square and adjusted R square it is not difficult. So as you know that we have used the command here lm() to obtain the linear regression model then we have used the command here summary in which we had obtained the confidence interval and test of hypothesis in that output there was another outcome about R square and adjusted R square which I had said at that time that I will take it in the further lecture. So this is the place where I am now going to address it. So if you try to use the summary command with lm then you will find the value of R square and adjusted R square on the other hand if you only want to find out the value of R square or say adjusted R square then you have to use the summary command with this lm() formula and then you have to write dollar R dot square all in lower case. And similarly if you want to find out the value of only adjusted R square you do not want to find out other values in the summary command then the command is you use the summary command with the lm() formula and then use here the dollar and then adj.r.squared all in lower case.

So this will give you the value of adjusted R square. So let us try to do it. So I try to store this data on x1, x2, x3 and y in this the beta vectors and now if you try to see here I am trying to use here this summary command. So now you can see here in this output I will just try to cross the part which you already have done. I will just use the all in colour so that you can visible.

You understand what is this? You understand what is this? This is about the residual. You understand this is about coefficient, estimate, standard error, p value that we already have done in the earlier lectures. So this part you already have now done. Residual standard error with 16 degrees of freedom this you already have done in the case of analysis of variance. Then f statistics with degrees of freedom and p values that you already have done in the case of analysis of variance.

So the only part which is left here is this. Now I am highlighting it in circling in red colour. Multiple R square and adjusted R square which are given here. So you can see here that here if you try to see in this output this is the place where you get the value of this multiple R square or say R square in general or the adjusted R square.

So R square is here 0.9995 and adjusted R square here is 0.9994. So as I said this value will always be less than the value of say this R square. And in case if you try to see here that how are you going to interpret it and can I will try to show you on the software also. So R square was defined here as SSregression divided by SStotal.

If you try to compute it from the ANOVA table also you will get the same value here. And this is obtained here like this. This multiple R square that is the name and this adjusted R square was defined here like this it is obtained here like this. So now you can see here it is not very difficult to obtain it. And if you only want to find out the value of this R square then as I said you simply try to use the command here summary then lm y it killed x1 + x2 + x3 and $ r.squared which will give you here the value 0.9995016. And if you want to get here the adjusted R square then you try to use the summary command with lm() command with dollar adjusted that adj.r.squared and you will get here this value. So this will give you this R square and this summary command will give you this value. So you can see here it is not very difficult here and you can see here this I have obtained here but anyway I will try to show you it on the software also. It is not difficult it is very easy. So let me try to first copy this data you can see here this is my data and then I try to come here this here command here say lm with this summary command you can see here it is here like this.

I am highlighting this part where you can see this is the value here multiple R squared and then we have the value of here adjusted R squared. So you can see here it is not very difficult to find out and in case if you want to find out only the standalone value then I have to use here this summary lm and lm command with this dollar R dot squared and you can see here it is giving here like this. So whatever is the value here 0.99955 this is mentioned here also. And if you try to use find out here the adjusted R square here then you can see here the standalone value of adjusted R square is obtained here 0.99944 which is obtained here like this one. So you can see here it is not very difficult to find out these values and to get the required information about the goodness of fit about the fitted model. So now after this let me try to address the another aspect here which is about the effecting of normality. So as I told you in the beginning that whenever we get a data usually the assumption is let X1, X2,… Xn or let Y1, Y2,… Yn be a sample from normal population with certain mean and certain variance. They may be known or unknown that is a different aspect. So the first job is that we want to test this assumption whether this assumption is correct or not that random errors are normally distributed or not.

So the next question is how to get it done. So we have got some analytical test also about goodness of fit but here I am going to demonstrate very simple technique which is called as Q-Q plot. Q-Q means quantile-quantile plot right. So we try to create here a normal probability plot using the residuals right. If you remember you had obtained the residuals as a Y-Yhat. They are the difference between the observed and fitted values.

So because we assume that epsilon 1, epsilon 2, epsilon n and they are coming from say normal 0, sigma square. So we want to use this as if we have obtained the value of the random errors which is impossible task. But anyway in order to construct this statistics we are going to think in this direction. So we try to obtain the ordered mean. Whatever is the minimum value out of e1, e2, … en this is indicated as say e inside the square bracket 1 in the subscript.

And this is here En is here the largest value, largest or say maximum value of e1, e2, … en right. So this is minimum of e1, e2, … en and this is here maximum. So we try to arrange it and then we try to compute the cumulative probability by this expression pi = i-1 upon 2 upon n. And then we try to plot between P i and this ordered residuals i goes from here 1 to n. So this is called as Q-Q plot which is Q means here quantile.

Surely I would like to address that in different software they try to compute the cumulative probability in different ways. But anyway my objective is to have some idea. So what type of idea we can get from this Q-Q plot. So I will try to give you here some idea which is from the simple statistical concept. So we have a concept of heavy or fat tailed distribution and similarly we have a concept of light tail or distribution.

So what we try to compare is that we try to compare the tail of the distribution with the normal. Right. So if I try to make here I can just say here suppose this is my hair normal and now I have here suppose one more distribution whose this tails are more than the tails of this hair normal. So this is here normal. So then I would say that this is a heavy tail distribution means all other properties are similar to normal but only that tails of that distribution are not matching and they are more than the normal distribution.

Similarly if the tail is less than the normal distribution like this one then this is called as say here say light tail distribution. Right. So this is how we try to classify and similarly the normal curve is always symmetric like this thing. But if this is skewed on the left hand side that means there is more data on the left hand side than the right hand side then it is called as positively skewed curve and when there is more data on the right hand side here like this one then it is called as negatively skewed curve. So we try to identify this type of distribution from the normal plot or the Q-Q plot.

Right. So we are now going to make these plots. Right. So first I will try to show you that how are you going to interpret it. So on the x-axis I have made the I have taken the ordered residuals and on the y-axis I have taken the pi which is the cumulative probability. So you know that cumulative probability will always be between 0 and 1.

So now there will be here line that will be plotted. Right. That is the straight line between the ordered residuals and this cumulative property and now for a given set of data it will compute the probability on the basis of data obtained and it will try to plot it on this curve. So these data points are here indicated by here these circles or small dots and then we have to see that how these points are concentrated around this line. So if you see that all these points are lying approximately on the straight line like here in this case then you can say that the underlying distribution that is thus the population from where the sample is obtained is normal. So normal means approximately normal.

Ideally means all the points should lie exactly on the line. Then if you try to see this type of here curve that the sharp upward and downward curves at both the axis. You can see here it is going here sharp and then becomes yellow and then again at the end it is a sharp trend. So this indicates that the underlying distribution from where the observations are coming is a bit tailed. That is the tail of the underlying distribution are thicker than the tails of the normal distribution.

Right. And similarly if you try to look at this curve it has this flattening type of points like as they are not that this and then suddenly it becomes steeper and then it becomes once again say flat. So this figure has a flattening at the extremes of the curve. So this indicates that the underlying distribution is light. That means the tails of the underlying distribution from where the observations are originating are thinner than the tail of the normal distribution. And if you have a figure line which has a sharp change in the direction of the trend in an upward direction from the mid.

You can see here it is coming like here this and then suddenly it started going like this. This indicates that the underlying distribution is positively skewed. And now we try to plot here the normal probability plot for the data what we have considered in that example. So the command here is the same command which we have used earlier plot. But now I have to give here the object that is lm() inside the parenthesis y ~ x1 + x2 + x3.

Now you have said this which = 2 it will try to give you here the line or dotted line and so on. So you can see here this is here the line and the points are here like this you can see here. Now it is your capability to decide that how you want to interpret it whether it is normal or something else. This I am leaving up to you because yeah it is something like looking at the same x-ray different doctors give different opinion and all of them are correct.

So it depends on your practice. So the best thing is if you try to generate some artificial data set from the normal distribution and then try to plot this QP plot and try to see and then try to get some more data sets from heavy tailed distribution, light tailed distribution and try to plot these things right. So here if you try to recall that yes I will try to write down this command here as say here this say here like here this and yeah we already have entered the data on y say here x1, x2 and here x3 right. And now you have to use a that command here plot lm() y ~ x1 + x2 + here x3 right. And then you have to get here which = here I will try to show you that by changing which what happens right.

If you try to see here you are getting here the plot like this one right. And if you try to just change here this which = 3 suppose like this here you can see here now you are getting a different type of curve. So but anyway you have to see into the into the help of plot function that what that what is the correct value for you to choose for the which right okay. So now we come to an end to this lecture and you can see that it was a pretty simple concept that how you can judge the goodness of it and how you can judge the normality of the observations right. So but the more important part is that how are you going to understand it and how are you going to take the correct observations or the correct statistical conclusion by looking at these values and looking at these plots. Now I will say this graphical tool and this analytical tool both should work together and they must give you the correct information if you have used the correct statistical tool.

I will say that you try to take some data set try to artificially generate normal data as you know how you can generate the normal observations or the random observation from a normal distribution and try to make this Q-Q plot and see that how it is looking like. Then you try to change the variance trying to change different distribution and try to see how it looks like right. And then try to add some unnecessary variables in your model and try to see what happened to the R square value. So that way you will gain more confidence more experience so that you can become a better data scientist. So you try to practice it and I will see you in the next lecture till then goodbye.