Multivariate Procedures with R

Prof. Shalabh

Department of Mathematics and Statistics

IIT Kanpur

Week – 09

Lecture – 39

Model Fitting with R Software

Hello friend, welcome to the course multivariate procedure with R. So, you can recall that in the last lecture we had discussed various concepts related to the multiple linear regression model we have introduced different terminologies on a release square a scimitar fitted values residuals etc. And these quantities are playing to are playing an important role when we try to find out the statistical model and we want to judge its just the validity of different assumptions. But surely when you get a real data you cannot compute those values so easily because the data may be too big and it may take a longer time. So, we try to take the help of software. So, in this lecture today I am going to repeat the earlier lecture whatever I had taught you through the R software.

So, I will try to take the same example which I introduced in the last lecture about the marks of the students and then I will try to fit here a multiple linear regression model using the R software. I will try to divide the output of that software into smaller components where I can show you that which component is indicating what or it is corresponding to which of the part which I explained you in the last lecture. So, let us begin our lecture and we try to understand that how we can obtain a multiple linear regression model in the R software. So, now we are going to talk about the model fitting with the R software in this lecture in the multiple linear regression modeling.

So, I try to consider here the same data which I explained you earlier that we have got 20 students on say 20 students about their obtained marks Y out of 250 and they are assumed to be dependent on 3 variables X1, X2, X3. X1 is the number of hours in a week that a student has studied, number of assignment the student has submitted in a month and X3 is the number of hours which a student has played in a week. So, these observations are given here under these three column X1, X2, X3. So, as I explained to

you earlier let me try to give you a quick review that when we try to say about the student number 1, then I can say that the student number 1 has studied for 34 hours in a week, the student has submitted 3 assignments in a month and the student has played for 15 hours in a week and the student number 1 has obtained 180 marks out of 250. And similarly for the second student student number 2, the student studied 12 hours in a week, submitted only 1 assignment in a month and played for 13 hours in a week and the student got 116 marks and so on.

So, this is my data set. Now, I want to find out here the multiple linear regression model between Y and X1, X2, X3. Now, before we try to move forward, we have to take the first decision that we want to fit here a model with intercept term or without intercept term. And this is a place where you will see that statistics does not give you everything, but many things you have to observe as a social scientist that what is happening around you. Now, if you try to see here in this case, I am considering the outcome Y to be dependent on these 3 variables, number of hours of study, number of assignments and the number of hours of play.

Do you think that if all X1, X2, X3 are equal to 0, whether this Y is going to be 0? Means in my opinion, no. Because it is possible that a student has, means a learn from some online courses, the student might already have done all these things in the past. Although a student has completed the questions in the assignment, but the student has not submitted it and suppose the student has not played, but the student was watching some film or something else called the entertainment. So, that means well these 3 variables, number of hours of study, number of assignment and number of hours of play, they definitely play an important role when in the life of a student to get good marks, but definitely there are many other factors which also play and there are alternatives to these variables which the student might have satisfied. So definitely in this case, I cannot say that expected value of Y that is the average marks will become 0 if X1, X2, X3 are all 0.

So definitely with this point of view, I would like to consider here a model with interceptor. So in order to fit a multiple linear regression model in R, we have a command here lm() which is about the short form of the linear models, l of linear m from models and its uses is actually like this. It has many options, but I will try to entertain here only those options which we are going to use here. The base command here is lm, then we have to give the formula. That means I have to give the input of response variable and explanatory variable in a particular way because I need to inform the R software also to identify that this is the response variable and this is the input variable or the explanatory variable.

Then I have to give the data. Data can be given in different formats. And there are many other options. I will try to take here only certain number of options which we are going to use here. So formula is as I said, this is an object of the class which is a symbolic description of the model to be fitted.

So and then data, data can be given in the form of data frame, list or any other option also. And you can see here then we have here a coefficient. This will give us a vector of coefficient that is the value of our nearly least square estimators. Then we have option for residuals. Then we obtain the fitted values and then we will also conduct the analysis of variance in the context of this linear model.

And the way we are going to give the formula, this I will although it is mentioned here, but I will try to explain it through the example which is easier to understand. So now suppose for this data set which I have taken here, I consider here a model yi is equal to beta0 + beta1 xi1 + beta2 xi2 + beta3 xi3 + epsilon i, i goes from 1 to 20, that means 20 observations. I try to means compile all the observations on x1, x2 and x3 as well as y in this four data vectors x1, x2, x3 and y, right. And from there, yeah, this is the screenshot and from there first I have to investigate whether there really exist a linear relationship between y and x1, x2, x3. So if you can recall when we did the graphics, then we had talked about the matrix scatter plot, right.

So we try to create here the matrix scatter plot and then it will give us the pairwise comparison of the values and then we can take a call whether a linear model is to be fitted here or not. So in order to obtain the matrix plot, we have the command here pairs() and then you have to give here the data in this format pairs and then tilde y + x1 + x2 + x3. This is a special format in which you have to give the data, right. And then yeah, main is the title of the curve that this is matrix scatter plot. So if you try to now look at this graph, first you try to look at this diagonal elements, right.

So it is like here something on the say here y axis, these are the y, x1, x2, x3 variables. On the x axis also, we have here y, x1, x2 and here x3. So if you try to see on the diagonal elements, they are trying to give you the graphics between their own variables, right. So for example, this here at 1, it is trying to give you the graphics y between y and in the here box number 2, it is trying to give you here the graphic between x1 and x1. In box number 3, it is trying to give you here the graphic between x2 and x2 and in box number 4, it is trying to give you the graphic between x3 and x3.

And now if you try to look in the off-diagonal elements, one if you try to see here in the box number here 5, box number here is y on this axis and here this is x1. So it is trying to give you here the idea that what is the trend between y and x1, that how you can see here this is like nearly there is a linear trend. You see 100% linearity is very difficult to obtain in real life and yeah and so but we try to partition it into different components and then we try to take a conclusion, right. So and similarly is the off-diagonal elements you can see here, this is the same as here 5 because it is trying to give you here the relationship between say here x1 and here y. So yeah means so we need to investigate only the off-diagonal elements either on the upper diagonal or on the lower diagonal, right.

So I will simply try to mark here the box number so that you can understand that these two boxes are giving us the same information. The only thing is this in one graphic that is between x and y and say another is between y and x, right. So that does not make any difference. So now we come to the box number here 6. You can see here it is not so easy to see that whether there is a linear trend or not, right.

So we try to just leave it at the moment, yeah. If you want you can say here this type of line can be there but it is very difficult to say whether line is like this or like this, right. And you can see here this is here the same box number 6 on the lower diagonal also. Now we come to here box number here 7. Seven you can see here this is the relationship between here say here y and here x3, right.

And you can see here this is here this type of line as well. So you can see that approximately there is a linear trend, right. So and the same thing is happening here also in box number 7. So you can see here from box number 5 this is the relationship between y and x1 which is linear, almost linear. Box number say here 6 which is the relationship between y and x2.

It is I cannot say very clearly what is happening but in box number 7 this relationship is nearly linear that is a good linear trend, right. So using this row I can see here that how is the relationship between y and individually xi's. But then remember your basic objective. Your basic objective was to have the relationship between y and x1, x2, x3, joint relationship of x1, x2, x3 with y should be linear. So now it is the experience and our capability to decide whether we want to move ahead with the linear model or not.

And remember one thing this matrix scatter plot is trying to give the scatter plot

between y and individual xi's and scatter plot may be non-linear but the model may be linear because the model is said to be linear when it is linear in the parameter. So do not depend only on the graphics but also try to use your statistical knowledge to decide whether there can be a linear model or a non-linear model. So that is my sincere advice to you all. And so now we try to take a final call. Okay, it looks that at least two variables are linear and third may also be linear with respect to y.

So that is why we will go ahead and we will try to fit here a linear regression model. Now before I try to move forward let me try to show you here what other boxes are trying to give you the information. Do you remember that you had made one assumption that rank of x is equal to k that means all x1, x2, xk they are independently distributed. So now if you try to look into the here box number here 8, box number 8 is trying to give you here a plot between here x1 and here x2 and similarly here box number 9 is trying to give you here a plot between say here x1 and here x3 and same plot is obtained here in the lower diagonal matrix. So you can see here it appears that there is no trend and the values are really scattered.

You see 100% achieving the independence in as possible only in theory. In practice you will always find some non-zero value of correlation coefficient but a lower value of correlation coefficient can be considered as the variable r independent. So you can see here these points are here, here, here, here like this, here, here like this. So I can take them as nearly independent. So similarly if you try to look into here the last here this here box in the box number here 10.

This is also the plot between here x2 and here x3. So you can see here the points are here like this and so I can say that the points are nearly independent. So and the same thing is happening in the lower diagonal matrix. So this is how you are going to take the conclusion about the model assumptions and there are many such graphical ways. As I said in the beginning that graphical ways and this analytic tool they work together.

So now I try to fit here my model on the basis of given set of data. My model here is yi is equal to beta0 + beta1 xi1 + beta2 xi2 + beta3 xi3 + epsilon i and my command here is lm and now you can see here how I am trying to give here the formula. So whatever is my response variable y this I try to write first y and then this here symbol with delta this symbol then I try to express my explanatory variables 1 + explanatory variable 2 up to here like this. So you can see here I have a 3 explanatory variable which I have indicated by capital x1, x2, x3. So they are written here x1 + x2 + here x3.

So that is what I told you earlier that I will try to explain you when I try to execute the command and it is very easy to understand. So now it is here the outcome. This outcome is indicating something which is now our duty to understand and which we want to understand from our basics fundamental. So this is the model here which you have fitted. Now if you try to see what is this thing.

This is written here intercept and then here this is the value here 8.769. So this is the value of beta0 hat which is 8.769. So this is the value of the intercept term which is obtained on the basis of given set of data. Then I have here the variable x1 whose corresponding value is obtained here as a 1.997 which is here the value of beta1 hat. So this is the ordinary least square estimator which you had obtained earlier. Then similarly for x2 we have this value 3.918 as a beta2 hat and for here x3 this value here is beta3 hat.

So now since you have used here the expression beta is equal to x transpose x whole inverse x transpose y which is your least square as well as maximum likelihood estimator. So now this is here at 3 by 1 vector b1, b2, b3. So this value here is here yeah there should also be here intercept term. So this value here is here b0, this value here is b1, this value here is b2 and this value here is b3. And if you try to recall your maximum likelihood estimation also, so maximum likelihood estimation you had given the symbol beta tilde as say x transpose x whole inverse x transpose y.

So that will be something like I can say b1 tilde, b2 tilde, b3 tilde and say here beta0 tilde. So this value will become here b0 tilde, this value will become b1 tilde, this value will become b2 tilde and this value will become here b3 tilde. So this is how you can see that it was useful for us to know those expressions so that I can identify that whose values are being computed here. What is the expression behind these values whose value have been obtained. So now you can see here based on these values you can write down here your model yi is equal to 8.769 + 1.997 x1 + 3.918 x2 and 6.106 x3.

So how you are trying to write down this model? This beta0 hat is coming from here, this is here beta1 hat, this is here beta2 hat and this is here beta3 hat. So you are simply trying to substitute these values and you are trying to write down here. Sometime people get confused that why I am not writing here epsilon, but if you try to see these values have been obtained in such a way so that summation epsilon square is minimum and because this is the sum of squares so its minimum value is 0. So it should not create any confusion in your mind. So and in case if you want to find out only the coefficient from here you are not interested in all these things because in many times in simulation you want to use only the coefficients or particular information.

So for that we have a command here coefficient Coefficeints and inside the parenthesis you have to write this command for linear models lm inside the parenthesis y tilde x1 + x2 + x3 and you will get here this value. So this is the same outcome which you have, you can see here you have obtained here. So this is only a subpart or a part of this outcome and this is I am trying to take here, this is the value of here like this expression. The ordinary least square estimator or it can also, I can also say this is my here MLE, this is the same thing. So this is the 4 cross 1 vector of the value beta b0, b1, b2, b3, this is the value of b0, this is the value of b1, this is the value of b2 and this is the value of b3.

And this is here the screenshot. So let me try to first show you these things on the R console and then I will try to move ahead. So first let me try to copy this data here and then I will try to come here. So you can see here this I have copied the data and first I try to create here this matrix scatter plot. And now in this plot if you want to use here different types of this information like a color, what is to be written on the y axis etc. you can do it. So now you can see here this is my scatter diagram, yeah it is you can increase it on the screen and it will give you a much clearer information. So this is the same which we have just analyzed. So now you should be confident that these things are working. Now if I try to write down here this command for obtaining the linear regression model, you can see here this is the same thing which I have obtained here. And you can see here this is formula is equal to here like this.

So this is the only thing that there is a special way in which you have to write this formula and that is precisely I told you that it is easier for me to explain you when I execute it. Now if I want to find out only the coefficients here you can see here this is the outcome here and you simply have to just write coefficients of lm y tilde x1 + x2 + x3 and you get here this information. You can see here this information and this information that is here the same, right. So this is how you can obtain the model. Okay now I try to come to my next topic which I had explained you during the last lecture.

This is about the fitted values. So you can recall that once you have this model y equal to x beta + epsilon you had obtained the ordinary least square estimator like as b is equal to x transpose x whole inverse x transpose y then you try to obtain the fitted value of y for x b, right. So it is like this if I try to say here suppose my fitted model here is 2 + 3 x1, right. And suppose you have obtained here some value of here y1 and y2 and for a given value of here x let me try to take care only 3x, right. So now this is the value which you have observed. Now you try to obtain the y1 fitted value here as say 2 + 3 x1, y2 hat is equal to 2 + 3 x2.

So you try to substitute the value of x equal to x1 in the model and try to obtain y1 hat, try to put x equal to x2 second value and try to obtain the second value of y. So if you try to see here this thing what is this indicating? The value of y which was observed as y1 is now obtained as y1 hat from the fitted model. So the difference between these two will be called as error or residual and this y1 hat y2 hat they are the fitted values, right. So in order to obtain this fitted values and residual values we have a special command in the R software and we try to find out here both. So in order to obtain the fitted value we have the command here f i t t e d dot values v a l u e s that is fitted.values() and inside the parenthesis you have to give the write down the same command for obtaining the linear model that is l m inside parenthesis y tilde x1 + x2 + x3 and you can obtain here this thing.

So you can see here there is 1, 2 up to here 20 observation, right. So this is the value of here y when your x1 was the first set of observations x1, x2 and x3. So this is your here y1 hat. This is your here the value of y when you use the second set of observation this is indicating here as a y2 hat. Now before I move forward do you think that for example if you try to see here this value is 180 and 160.

Now if I try to come to my original example here this value is here given here as say y. So this value here is say here y1 hat now you have obtained and this value you have obtained here as say y2 hat. Now if the difference between y1 hat and say here y1 and if y2 hat and y2 is small then what do we expect that your model is good. But if difference is large so then you feel that the model is not bad. So if you try to see here your observed values are 180 and 116 whereas what you have obtained them here they are 180.0045 and 116.0273. So the difference is only 0.0045 and 0.0273 respectively in the first two values. It is not much, right.

So I can see that okay the values which you have obtained from the model they are almost similar to what we had obtained in the real data. So now you can imagine that when this is happening for the known values on the other hand if you try to take some unknown values of x1, x2 and x3 and then try to find out the value of y you can expect that when the difference between y and y hat is smaller in this 20 values the same may continue in the next set of values where x1, x2, x3 are known but y is not known and this is called as forecasting.

So your forecasting depends on the estimated model. If your estimated model is good

then your forecasting if the similar type of values continues in the future that will also be good. So this is how we try to obtain the fitted values and you can see here this is here the screenshot of the same observation but let me try to first also give you here this concept which I just explained you. You can recall that we had obtained the residual vector as a difference between y and y hat that is the difference between fitted and observed values. So this was obtained here as a y - xB. So if you want to obtain only the residuals then you have the command here residuals() inside parentheses you have to write down the expression lm() inside the parentheses y tilde x1 + x2 + x3 and you get here these values.

So if you try to see what is happening. What is first value? That is the difference between y1 - y1 hat. How you had obtained the y1 hat? That was obtained by fitted value, right. So observed value, difference, fitted value. So you can see here this is very close to 0. Similarly for the second observation also this difference is pretty close to 0 and both are in the negative direction.

But if you try to look at the third value this difference is more than the difference in the first two observations. And similarly somewhere you can see it is negative, somewhere it is positive and so on. But if you try to take the average of all these one, so it is will be 0 and so you can expect that in the population this random errors will have almost a 0 value. And that is what we assume or indicated by your assumption expected value epsilon is equal to 0. So now you can see here this is the screenshot of the same thing but let me try to show you these things first on the R console, right.

So you can see here if I try to write down here the fitted values here. So now you can see here these are the fitted values what you have obtained from the model. And secondly I try to obtain here this residuals. So for that I try to use the command here like this and you can see here these values have been obtained without any problem that is so straightforward, right. So now we come to an end to this lecture. I am not saying that I have finished all the things what you have, what is to be done in the multiple linear regression modelling.

There are many more things which I will try to take up in the next lectures. But if you try to see in this lecture what I did, although the earlier lecture was pretty long but whatsoever I explained you that was required to understand the output in this current lecture. When you get the values from the software these values are trying to indicate something they are trying to speak they want to give you certain information. And what are these information that is what we have to understand. For example, if I say what is the meaning of this regression coefficients that we have obtained.

Suppose I have obtained the regression model as say y is equal to say 2 + 3 x1 + 4 x2. So intercept term here is 2, right. So this means that if I try to take the value of x1 and x2 to be 0 then the average value of y that will be observed is equal to the value of the intercept term what we have obtained from the software, right. Similarly if you have the value of beta1 hat suppose equal to here 2. So now what will happen here? beta1 hat is indicating the change in the value of or the change in the average value of y when there is a unit change in the value of x1.

That is the partial derivative of expected value of y with respect to x1. So if this is 2 that means if you try to make a change in the value of x1 by 1 unit then the expected average change in the average value of y will be 2 units. So this is how we try to interpret it. Now in case if the sign of the estimated regression coefficient is positive. Suppose it is + 2 x1.

So this means the relationship between y and x1 is positive. If there is a 1 unit change in the value of x1 then the average change or average increase in the value of y will be 2 units. But on the other hand suppose my second variable has a coefficient minus 3 negative sign. This is indicating that the relationship between y and x3 is decreasing. That means if there is a 1 unit change in the value of x3 then the average expected change in the value of y will be 3 units and the values of y will decrease.

They will become lower. So that is the indication of this plus and minus signs of this regression coefficient. So now if you try to see in this example I have used the data without any scaling. This one was used in its units, x2 was used in its units so called. But now suppose there is a situation where I want to make this dataset independent of the units. Then in that case you can use the option of scaling and for that we already have done a command scale.

So if you operate the scale on these values then whatever scaling you want whether the centering or only the means as credit score or something like this you can choose it that we already have done. But try to see that if the data is scaled are you getting the same output or are you getting the same conclusion? It is not always possible that whatever the information which you are getting on the scale delta will be copied and pasted in the scale data also. It may or may not happen. So then you have to see how to interpret the data in the context of the real situation. What is happening in the process, which of them is going to demonstrate in a better way and then you try to decide.

So that is why I always say that this data analysis cannot be done in a single shot that data comes you can do click, click, click and you get the model. No, it is a recursive process. Many times you will see that some variables are not important. How to adjust that we will try to discuss in the forthcoming lecture. But if you feel that the variables are not important then you have to drop them and then you have to recalculate these regression coefficient.

For example, if you have taken here the variables x1, x2, x3 and then you have obtained the model. But if you try to consider only an x1, x2 and you drop x3 do you think that are you going to get the same model? Certainly not. So the value of the regression coefficient estimator like is b0, b1, b2 and the values of residuals, fitted value they all will change. So we try to keep these experiments going on unless and until we get a model which is as close as possible to the data set what we have obtained which is coming from the process. So the process is trying to say something through the data set and we have to understand the language of the process.

As soon as our statistical model speaks the same thing what our process is trying to tell we should be happy that we have got a good model. Now how to judge quantitatively? That we will try to discuss in the forthcoming lectures. But anyway this is all about the statistical modeling. So you will understand and you will agree with me that a good model can be obtained by good practice and experience. The more you try to spend your time, try to understand, try to think better statistical model you can obtained.

So you try to practice it and I will see you in the next lecture with some more topics on the multiple linear regression model. Till then goodbye. Thank you.