Multivariate Procedures with R

Prof. Shalabh

Department of Mathematics and Statistics

IIT Kanpur

Week – 04

Lecture – 18

Scatter Plots, Smooth Scatter Plots and Matrix Plots

Hello friends, welcome to the course Multivariate Procedure with R. So, you can recall that in the last lecture we had considered the histograms and we had extended the concept of univariate histogram to a multiple histogram setup. So, now we are moving towards that how we are going to consider more than one variable in a single plot. Now, from it today we are going to increase the number of variables in our graphs. So, from one variable so now we are moving to bivariate graphs. Bivariate graphs are used when there is an observation which depends on two variables right.

In order to plot it we simply use x y axis on x axis one values on y axis another values. And based on that we try to plot them in on a two-dimensional plot. Scattered diagram, two-dimensional scattered diagram etcetera they are very popular graphics which are used. And these graphics are very helpful in determining the trend in the data.

Why trend in the data? One of the very important objectives in statistics is to find out a model. Model means the relationship between the input and output data sets. So, how to know whether the relationship between the two variables is linear or not? So, for that we always try to collect some observation and we try to plot them in a two-dimensional setup and we try to see what is the trend of the data. If the trend is like this, I can say that okay there is an increasing trend when if you try to increase one value the other value is increasing or vice versa that if you try to increase one value then the other value is decreasing. Or if there is any other non-linear relationship that can also be found from here.

And based on that we try to decide that what type of statistical modelling has to be done, what type of model can be fitted. So, this lecture is going to introduce the concept of scattered plots in two dimension and we will try to extend it to a multivariate setup when we have more than two variables. And this is going to be used when we are trying to do the statistical modelling using the regression analysis. So, let us begin our lecture and try to understand this scattered plot and its different variants. Okay, so now in this lecture we are going to cover scatter plot, smooth scatter plots as well as matrix plot.

So, these bivariate plots they provide the first-hand visual information about the nature and degree of relationship between two variables. This relationship can be linear as well as non-linear. Now, we will try to discuss here different types of such plots through various example. So, one command to create the scattered plot is plot, P-L-O-T. What is a scatter plot? That if you have data on say in two dimension you can make here this type of plot.

So, this is called here as a scatter plot. So, if I try to contain the data in two data vectors X and Y, I mean the data which has to be plotted on the X axis, this is indicated by X and the data which has to be plotted on the Y axis, this will be here, right. So, now the command to make this two-dimensional scatter plot is P-L-O-T and inside the parenthesis you would like to express this two data vectors X and Y, right. After this there are different ways in which this scatter plot can be represented and those options can be given here by a type. And the different choices for types are say P, if I write P is equal to points say P then it is indicating the points that there will only be dots.

And if I write here type is equal to l, then it is going to be here for lines. Then if I try to write type equal to b, then points and line actually both. And if I try to write here type is equal to c, then it will give us only the lines, line part alone which was created here in b. If I say here type is equal to 0, that means both points, line they will over plot it. If I use here type equal to s, then there will be stair steps.

If I use here h, then this is going to be a histogram like graphic which is also the high-density vertical lines, right. So, let me try to show you here some more options which are available in the plot command, but my request to you all is that please try to look into the help menu. They will give you more information in detail, right. Suppose if I want to put the title of the graph, then I can use here main, M A I N. If I want to write down the subtitle of the plot, this will be your here suba.

If you want to write down the title on the x axis, then it will be xlaba. And if you want to write down the title on the y axis, then it will be ylaba. And if you want to control the aspect ratio that is y, y is 2 x, then the command here or the option here is aspect. With this and then there are many more actually. If you try to look into the help of this type and other plot command, you will get more information.

So, let me try to give you this illustration using a very simple example. And we had considered this data set earlier also in one of the lectures, but now here I will use it in more detail. We know that the demand of water in a city, it depends upon the weather temperature. We know that during summer when the temperature is high, then people use more water in comparison to during winters when the temperature is low, right. So, from our experience we know that the water consumption increases as the weather temperature increases.

But now this is only our observation. Now we want to choose data set and we want to know about these aspects. So, what we have done that data on the daily water demand and the day temperature that is collected for 27 days, right. And this data on the daily water demand which is measured in million litres, it is stored in a data vector here water and the temperature it is stored in the T E M P, temp, right. And this is like this that if you try to read here the first value. So, this is indicating the first value 33710 and the first value in temperature that is 23. That means, when the temperature was 23, then the consumption was 33710 million litres. So, this is how it will go like. Now we try to create these plots, right. So, this is about the plot command.

So, if you try to see I am writing here plot water temp and you get here this type of graphics. You can see here this is here water, this is the temp and now these are the paired observation which are plotted here as like this one, this one and so on. By looking at this graphic, do you think that is it a linear trend in the relationship? If you try to plot an average relationship, do not you think it is going like this or this can be here like this also, whatever you want that is a reco equivalent thing, right. So, but at least if I try to remove this line, you can see here that there is a linear trend in the data. So, this type of information can be obtained from such graphics.

And now this will help you in deciding whether you want to create a linear model or a non-linear model etcetera. Now, simply if you try to change here the type, now I try to take here type is equal to l that means l for lines. So, all these points they are going to be joined here like this. If you try to see here in this one, if you try to join here these points like this one, then it will become a line diagram. Now, in case if you try to use here both, then type will become actually here B and then I can see here line and both, both will be plotted.

So, you can see here these are the points and then they are joined by this line, right. And similarly, if you try to use here type is equal to o that means over plotted, then these lines are over plotting those dots, those points. For example, if you try to see in the earlier plot, there is no overwriting that means this line is ending here and these points are here separated. But in this case, when the type here is o, then this line is over plotting the points. Now, so similarly if you try to use here the plot here H, then all these points will behave like this one and this high-density lines will be created here.

So, it is, it looks like as a histogram, it says a very small width and these are essentially the high-density vertical lines. So, this is how the, it is going to be created. In case if you use here the option type is equal to s, then all these observations here they will be like this. It looks like stairs, stairs we know that on which people actually climb like this. So, now depending on these different types of observation or say requirement, we try to make suitable choice of the type and we try to make suitable plots.

Now, in the same plot if you want to add more information, for example, if you want to add here this command or the title of the chart daily water consumption versus day temperature, then this can be given as say here like this one say here, which is the same command which you have used earlier also if you recall. And then there is here if you want to put it here day temperature, then it is the value on the y, this is given here by ylab. This is the same command that you use earlier also. If you want to give the statement on the or writing on the x-axis daily water consumption, then it can be given by xlab, right. So, this is how you can create or you can modify or you can add more information to the same graph.

And in case if you want to make a smooth line also or smooth curve also that to understand the trend of the data, then we have an option here that we can use here a command scatter. smooth, and inside the parenthesis you have to give the data vectors. For example, scatter dot smooth of water and temp, this will give you here this type of line, right. So, now depending on your choice and need means you can choose quotable graphs and whatever you want. And in this smooth scatter plot also there are various options are available like a span, degree, family, etcetera, etcetera. So, I would my request to you will be that you please try to look into the help menu and try to see what these different options can do for you, right.

And in the same scatter smooth plot, if you want to for example, if you want to make different type of line something like you can see here this is a line which has the broken line like this one. Then for example, if you simply use here the command l pass is equal to list where color is equal to red, lwd is equal to 3 and lty is equal to 3. This lwd and lty values for different values you get different type of this graphics. So, my request will be either you look into the help or you try to change this value and try to see that how this pattern is changing when you try to change these values, right. So, let me try to first give you the illustration of these commands on the R console.

So, first let me try to copy this data over here. And similarly, I try to copy the data here for for here this temperature, right. So, you can see here this is your here data on water and this is your here data on temperature. Now, if you see here plot water comma temp you get here this type of plot, right. And if you try to interchange this role for example, now if I try to write down here temp at the first-place temp comma water then you will see that this temp will become here like here here water, right.

So, you can see here this will get changed and the pattern of the curve will change. So, it depends on your need what exactly do you want, right. Similarly, if you try to see here in this water temp if you want to add here the type here type is equal to here and say here l.

We can see here this is here now the lines will be added. And if you want to change this type is equal to here over plotted now you get here this type of figure where points and line they are over plotting each other.

And if you try to make it a type is equal to here b then you can see here now this points and lines they are separated, they are not over plotting. And if you try to use it here the option here h then this is for high density line and we will get here this type of graphics. And if you try to use here type is equal to s then you will get here a stair type graphic like this. Now, this is up to you that which graphic you feel is representing the data in a better way in a given situation. And then based on that you can choose a suitable command, right.

Now, similarly if you try to use here the command here scatter dot smooth and here water comma temp, right. You can see here now you are getting the same plot, but now there is a here a line, right. Similarly, if you want to make it here more beautiful you can see here if I want to write down here. See a plot with different types of options you can see here. Now, you are getting here this title daily water consumption versus day temperature then x axis y axis they are daily water consumption and day temperature they are here, right.

Similarly, if you want to make this scatter.smooth command if you want to use with different options then you can see here if you want to make it here like this then the this line is going to be changed like this one. So, now you can experiment with the different types of this values and try to see how you can make the use of this plot command to give you the proper information. Now, I try to extend this plot command to a multivariate setup, right. Well, I will be taking here only two variables because of the limitation of the space on my slide, but you can create it for more than two variables, but you will see that if I try to create the such a plot for more than two variables 4, 5 then the picture will become very clumsy and it will be and it will be very difficult for me to explain. So, that is why I am trying to restrict myself only to the means two variables only, right.

The basic concept here is like this. Suppose, I have here three variables x, y and z and suppose they are interdependent on each other. Now, in order to understand the interrelationship between two variables at a time, then what we can do that we can make here a plot of say x and y, x and z and y and z. So, now if you try to make three different plots, then it might be difficult for you to make a comparison, but similar to the concept what we had used in the multiple histogram that we try to create the histogram of different variables in the same plot. Similarly, we can create this scatter plot in a same graphic. For example, if I say here suppose if I try to make here like this and I try to take here the variable here x, y, z horizontally and x, y, z vertically.

Now, in this one if you try to see here x versus x, this will be a plot between x and x which has practically no meaning actually because that is the same variable. Now, if you try to come to this cell where here this x come here, so I can make here a plot of x and y and if you come to here the third cell then x comes from here and z comes from here, so I can make here a command of here x and z. And similarly, in the second row this will be your here y versus y which has no meaning and then you will have here y from horizontal and z from vertical which is here plot between the variables y and z. And similarly, in the third row this is going to be between z and z which has no meaning actually. And then yeah, the plot between x and y and the plot between y and x they are going to be the same, the only thing will be there x's are changed but the pattern is not really going to change.

And similarly, this x and z will be replicated here with z and x and y and z is replicated here with z and y. So, basically if you try to look into this plot then you are getting here when you have three variables you are getting here say all possible plots between x, y, x, z, y, z and there if you try to interchange their axis is also that you can also get. So, this type of plot is actually called as matrix scatter plot as the name suggests that you are trying to create here a matrix of the different scatter plots. So, in order to create such scatter plots you have a command here pairs, p-a-i-r-s. But now we have to first combine the variables for which we want to create the plot.

So, for that we have a command here c bind. For example, if I want to create a matrix scatter plot for the same data water and temperature I can write down here c bind say water and temp, water comma temp and then I use here a command pairs, p-a-i-r-s. So, you will see here now this will be your here water and here temperature. And similarly on the horizontal part also this will be here water and temperature. So, now in case if you try to look into this part here. So, this is between water and water which is not required and this second diagonal element will be temperature on x axis and temperature on y axis also which has no utility.

But if you come here if you try to see this is between water and temp. So, it is equivalent to my command plot water and temp. And if you try to look into this lower diagonal. So, this temperature is coming from here and water is coming from here. So, this is equivalent to my command here plot between temperature and water.

Well, I have taken here only two variables. So, you can see here that it looks very simple, but surely you can have 3, 4, 5 variables and then you will get pair wise illustration of the relationship among different variables. And now it will be your capability that how are you going to take a judgment for the joint variable. For example, if I have three variables suppose x and y have some trend y and z has some different trend and x and z has some different trend. Now, it is a challenge to decide what is the joint variation of x, y and z. But it is not difficult because if you practice then it is then you can always find out.

So, this is how this matrix command matrix scatter plot lots are created. And if you want to give here more information for example, I can put here labels like as here daily water demand, day temperature. So, instead of water and temperature you are getting here like daily water demand and day temperature. Similarly, you can write down here main, title, you can choose the values on x and y, x is etcetera. All those things can be done exactly in the same way as you used to do earlier.

So, let me try to first show you this thing on the R console and then I will try to conclude this lecture. So, right. So, let me try to remove this thing. If you try to see here as soon as I say pears, c bind water and temperature this matrix scatter plot is obtained which is of which has 2 by 2 matrices. And in case if you try to change the role of water and temperature.

So, I can make it c bind between water between temperature and water. So, you can see that as soon as I execute it this water and temperature places they will be changed you can see here. So, now means you can make different types of such options and you can create different type of say scatter plots with different types of combinations of different types of variables and they will give you a fair idea about the pair wise relationships existing in the data right. So, now we come to an end to this lecture and you can see that we have a demonstrated here the use of plot command. It is a very simple command, but believe me it is a very important tool.

Whenever we are trying to handle the data in a multivariate setup for example, if I say if an output say y is dependent on say 3 variables x1, x2, x3 for example, yield of the crop is dependent on quantity of fertilizer, temperature, rainfall and suppose we have this data. So, what are we interested in? We want to know what is the relationship between y with respect to a group of variable x1, x2, x3. It is not an easy job to exactly determine such a relationship, but this is the starting point to create any statistical model. In such a case this plot command helps us and if I try to plot y with respect to x1, y with respect to x2 and y with respect to x3 and suppose all the 3 relationships they are indicating that they are increasing and there is a linear relationship. So, what you can now judge? You can very easily judge that the joint relationship of y with respect to x1, x2, x3 is going to be linear.

Yes, this is a very simple option. The story will become more tricky or a complicated when the relationship between y and say x1 is linear, but y between x2 is say non-linear, y and x3 is linear, but suppose the relationship between y and x1 is linear, whether trends of y versus x1 and y versus x3 are reverse or there can be different combination if you try to increase the number of variables, but at least this will give you a starting point to think that what type of model has to be assumed so that we can move further and you will see this particular this matrix scatter plot helps a lot. So, my request to you here is that please try to take some examples and try to create such plots. I would say why do not you try to create a data set in which suppose there are 4 variables. Now, you choose that the relationship between y and say x1 is increasing, but linear, relationship between y and x2 is linear, but decreasing, etc. And then you try to create this type of data set from where you have some idea and you try to make a scatter plot, matrix scatter plot and try to see whatever you had created, now is it coming in the matrix scatter plot also? And then you try to make different types of permutations, different types of trends and try to see how you can judge the correct trend by looking at the matrix scatter plot and you will be able to judge it because you know the truth, you have a data which was created only by you.

So, that is how you can gain some experience and as I say always it is something like by looking at the x-ray the doctor comes to know what is happening inside the body because all the parameters in our body, they are multivariate. The blood pressure does not depend only on one variable, but depends on several variables, but doctor has to take a proper clinical information based on that photograph of x-ray that what is happening inside the body or looking at the ECG the doctors try to see what is happening inside the heart. So, that is why the interpretation of the graphics in the proper way, correct way is very important to learn and experience and working with these different types of examples will make you a better data scientist who can always take a correct decision from the graphics. So, you try to practice it and I will see you in the next lecture till then, goodbye.