Multivariate Procedures with R

Prof. Shalabh

Department of Mathematics and Statistics

IIT Kanpur

Week – 03

Lecture – 10

Bivariate Data

Hello friends, welcome to the course, Multivariate Procedure with R. So, you can recall that in the last lecture, we initiated a discussion on various topics in descriptive statistics, and we had considered basically the univariate data, in which we discussed the concept of mean and variability, that is, the measures of central tendency and measures of variation

Now, in this lecture, we are going to consider the bivariate data. That means the data is available on two variables for each individual. So, as soon as we have more than one variable, then their interrelationship concept also comes into the picture

So, the question becomes: how are you going to measure these interrelationships? And when you try to go with the multivariate data, meaning more than two dimensions or more than two variables, then all this interaction, variability, mean, etc., come together, and we have to understand how we can handle such a situation. That is what we are going to understand in this lecture. So, we begin this lecture, and we try to understand the implementation of these tools in the R software.

Now, in this lecture, we consider this bivariate data and different tools. You can recall that we discussed in the last lecture that whenever data comes to us, the first-hand tools that give us the first-hand information are about the central tendency of the data, variation in the data, and the relationship study.

The central tendency of the data gives you an idea about where the observations are concentrated, around what value. The variation gives you an idea about the scatteredness and the spread of the data around the central value. Relationship studies give you an idea about the nature and behavior of the relationships existing within the data. For these things, we had advised that both graphical and analytical tools can be used, but here we are going to talk about the analytical tool.

Whenever we have more than one variable, for example, two variables, the data obtained on them is bivariate data. For example, if I say I have two random variables, X and Y, suppose X indicates the height of people, and Y indicates the weight of people. Suppose we have person number 1, person number 2, person number 3, and so on. Now, we try to measure the height of the person. Suppose the first person's height is 150 centimeters, and their weight is 50 kilograms.

The second person's height is 100 centimeters, and their weight is 60 kilograms. The third person's height is 170 centimeters, and their weight is 70 kilograms. So, this is the data on two variables. One set of data will have two values like this, and this is called bivariate data. This quantitative measure provides the quantitative measures of relationships.

What does this mean? For example, if I say what is the relationship between height and weight usually we see it in the human being that as the height increases the weight also increases in general right. And in order to understand what is the trend in the data whether as the one variable increases, then the other variable decreases or vice versa that means if one variable increases other decreases then the graphical plots also provide a first-hand visual observation about the nature and degree or the relationship between these two variables. For example, if I say here x and y here are like this and if you try to plot this paired data and it is showing you here a trend that is indicating that as x is increasing y is also increasing, but if you have a data like this. So, it is indicating that as x is increasing the y is decreasing right. So, this type of relationships exist in different types of data and such relationships can be linear as well as non-linear.

For example, you can have here the relationship between x and y here as like this right. As you can see here that it is non-linear relationship. So, there are different types of statistical tools which are used to handle such situation when the relationships are linear or non-linear, but here we are focusing basically on the linear relationships right. So, this is our objective in this lecture right. So, when we have the observations on two variables suppose we try to indicate their data values as $x_1, x_2, \dots x_n$ and which are on the values on the random variable x and another variable y on which we are trying to obtain the values $y_1, y_2, \dots y_n$ right.

So, for example, x can be height and y can be weight. So, we have now organized the numerical values on these two random variables in two data vectors x and y. Now, first concept is covariance. Now, if you try to understand the meaning when we studied the variability then we had used the concept of variance and we had said that the variance is trying to measure the scatteredness of the observation around the central value. Similarly, when we have two variables then how one variable behaves with respect to other variable around the mean values this can be measured by the covariance and it is defined as say here cov that is covariance between x and y is $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$.

The covariance between x and y can be measured in R by the function cov and inside the parenthesis we have to give the data vectors like x comma y and if you try to see if you substitute here $x_i = y_i$ then this becomes nothing but your $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ which is nothing but your variance. So, this is what I am writing here that the variance can be computed by the command here v-a-r and inside the parenthesis the data vector. Now the thing is this covariance can lie between minus infinity and plus infinity and if the sign of the covariance is positive then it indicates that the relationship is negative and if the sign of the covariance is positive then it is indicating that the relationship between the two random variable is positive. That means, if relationship is negative that means, when one variable increases the other decreases and if the covariance is positive that means, if one variable is increases then the other variable is also increases, but now what the problem here is that it is lying between minus infinity and plus infinity. So, it is very difficult to measure the degree of dependence whether the rate of increase or decrease is very high or small.

So, in order to avoid such issues, we have a concept of correlation coefficient. This correlation coefficient measures the degree of linear relationship between the two variables and it is defined as the ratio of $\frac{Cov(x,y)}{se(x)se(y)}$ or this $\frac{Cov(x,y)}{\sqrt{var(x)var(y)}}$ which is here like this and it is usually defined as covariance here say r, r(x,y) that is the correlation coefficient between x and y. So, you can see here this correlation coefficient is a function of covariance and in the denominator variance of x and variance of y they are always going to be positive. So, the sign of the covariance is reflected in the sign of correlation coefficient. So, if covariance is positive then correlation coefficient is also going to be positive and covariance is negative then correlation coefficient is also going to be negative.

So, that is how we can overcome this issue by looking at the ranges of this correlation coefficient. You can see here that this correlation coefficient lies between -1 and 1. So, now if you try to see the value of r(x,y) is bounded between -1 and 1. So, it is very easy for us to understand to decide about the nature of degree of the dependence and the sign of the covariance is indicating in r that if r is positive then the relationship is positive and if the relationship is negative then the sign of r is negative. So, this correlation coefficient in the R software can be computed by the command c o r.

So, if I write c o r inside the parenthesis x, y it can very easily give us the value of the correlation coefficient between x and y. Now, let me try to show you how it interprets what is the meaning of various values of r because r is lying between minus 1 and plus so, it is pretty easy for us to use it. So, if you try to see here first you try to see at this one. I am talking here r equal to plus 1.

You can see here all the values are lying in a straight line that means as x increases y also increases and that is indicated by the plus sign of this correlation coefficient. Now, if you try to look at other value here r equal to 0.9. So, this is here actually plus 0.9. So, if you try to compare it with this r equal to 1 you can see here this the line which is passing through most of the point is like this, but some of the points they are away from the line.

So, this is indicating the degree of departure from the central line and that is what is the difference which is between r equal to 1 and r equal to 0.9. And similarly, if you try to look into this picture here the line will be somewhere here, but the difference between the lines and this observed value as compared to r equal to 0.9 this is more and this is reflected in the value r equal to 0.5 which is plus 0.5. So, that will be indicating the difference between the magnitude of r equal to 0.9 and r equal to 0.5. This is graphically you can understand it very clearly. In both the cases as x is increasing y is also increasing.

Now if you try to look into this first picture r equal to 0.90, but it is with minus sign you can see here. So, minus sign is indicating that as x is increasing then y is decreasing. So, if you try to plot here a line it will go like this that is a straight line, but most of the points are lying very close to the line, but not on the line. If all the points are lying exactly on the line then possibly in such a case the r is going to take the value minus 1.

Similar is the story in the next figure where you can see here r here is minus 0.5 that means as x is increasing y is decreasing and the points are quite far away from the rest of the values. And the difference or the deviation is reflected from this picture to this picture which are the numerical values of r equal to minus 0.9 and r equal to minus 0.5. And if there is no linear relationship between x and y that if it is not clear what happens when x increases or y increases then this correlation coefficient is 0 and you can see here there is no clear-cut linear relationship.

But then we have to be very careful while interpreting and understanding this concept. This type of relationship indicates that the occurrence of x and y are not affecting each other and in statistical knowledge that is interpreted as if x and y are independent random variables. But I would like to make it clear that in case then correlation coefficient x between x and y will be 0. But if r x y the correlation coefficient between x and y is 0 then it is not necessarily x and y are independent.

What does this mean? This means that in case if the value of correlation coefficient is close to 0 please do not always infer that x and y are not affecting each other. Why? Because as I said that correlation coefficient is a measure of the degree of linear relationship between the two variable but there may exist a non-linear relationship. Some data can be here like this and if you try to put here a linear trend then definitely the value of correlation coefficient is going to be pretty close to 0 but it does not mean that x and y are independent but x and y have got some relationship which is non-linear. So this is how you have to be very careful when you try to interpret these concepts. Now in case if you try to see for example if I try to take here two data vectors x and y suppose I am taking here x here as say 1 2 3 4 and y here as say 1 2 3 4 both are the same then in that case the co-variance between them is coming out to be positive.

You can see here if you try to plot it this will look like this and in the R software it is cov and you can get the value here 1.66. Now if I try to take here x as 1 2 3 4 and y here is exactly the negative values the same negative values then you can see here that the covariance between these two sets of values is minus 1.66. So if you try to compare here with this value and this value the difference here is with the negative sign only.

So you can see here in this case when the value of x's are increasing the values of y's are decreasing. So, this negative sign here this is indicating that the relationship between x and y is negative and the same will be reflected when you try to compute the correlation coefficient. So, before I try to move forward let me try to show you these things on the R console so that you can get confident about these things and then I will try to show you something more. So if I try to say here this covariance between say here C 1 comma 2 comma 3 comma 4 and C 1 comma 2 comma 3 comma 4 you can see here this is 1.66 but if I try to make it here minus 1 minus 2 minus 3 minus 4 the numerical values are going to be the same only the minus sign is here which is indicating the nature of relationship and the magnitude of the degree of dependence between x and y this is the same 1.66.

And if you try to see here suppose if I want to make only two points negative this value will remain the same but now this sign has become positive. So this is how we try to understand this thing. And similarly, if I try to show you here that what about the correlation coefficient, correlation coefficient between c(1, 2, 3, 4) and c(1, 2, 3, 4) how do you expect from here what should be the value of this correlation coefficient it should be 1 you can see here. And if you try to take the same values but there is a negative relationship like this one here this will be here minus 1. And if you try to make here a small deviation here suppose I can make it here 1.1 and here see here 2.1 and instead of 3, I can write 2.7 you can see here this is 0.99. So with this type of practice you can understand that how this numerical values are going to give you this different types of interpretations. Now let me try to come back to our slides and try to understand.

Now you see I want to explain you here a basic concept. Now you have understood that if I have a random variable here X then it has got variance of X. If I have one more variable here Y then this is got here variance of Y. Now if I have a bivariate distribution then also, we have covariance between X and Y. And similarly if you try to think suppose if I add here one more variable here Z then we will add here variance of here Z this will be here and simultaneously then we will have here covariance between X and Z and covariance between Y and Z.

So, now the question is this when you are trying to increase this label this order means more number of variables are getting added and you are going from bivariate to tri-variate to multivariate then how would you like to get this information about the variability of the data set. In order to do this thing, we have a concept of covariance matrix. So in the covariance matrix what we try to do here is the following that suppose I have here three random variables X, Y and Z then we try to write down the covariance matrix of this like this. This is a matrix on the diagonal elements there will be variances and on the off diagonal elements there will be covariances. For example, if you want to write down the covariance matrix of say X and Y of Y variate random variable then it will be variance of X and variance of Y on the diagonal and covariance between X and Y and covariance between X and Y on the off-diagonal elements.

Well the covariance between X and Y and the covariance between Y and X they are the same. So, it does not make any difference whether you want to write covariance X, Y and covariance of Y, X. Similarly if you want to write down the covariance matrix of X, Y and Z. So, this will be now be on the diagonal elements, this will be variance of X, variance of Y and here variance of Z. And on the off-diagonal elements this will be covariance between X and Y, covariance between X and Z and covariance between Y and Z.

And here this will be the same thing, covariance between X and Y and covariance between X and Z and covariance between say Y and Z. So, this is a matrix. So, now you can see here as soon as you go from univariate to bivariate case, all your parameters which you have defined earlier or the tools which you have defined earlier they have to be translated to a higher dimension. And for that you need to understand that what is happening in the univariate case, right. You can also ask me I can take this opportunity to explain you here if you observe here in this side.

What will happen to mean? In the univariate case if you have a random variable x then it is mean of x only. But if you have two variables x and y then this will have a mean of x and mean of here y and if you have three variables x, y and z then it will be a vector where you will have the mean of x, mean of y and mean of z. And this is called as mean vector. So, this is how you can see that why it was important for us to understand the univariate case so that you can understand this multivariate case and now you can see here this is the direct extension with some modification in such a case, right. So, this is how we want to progress further gradually and this is my objective that I want to build up the concept gradually so that you can pick up the things and after that you should be able to develop them yourself without my help.

So, now although I have already shown you on the R software but still, I would try to give you more detail. So, if you want to see here the correlation coefficient between C1234 and 1234, first try to observe on this graph. First point is there, 22 point is there, 33 point is there and 44 point is there and there is an exact 100% relationship and that is indicated by the value plus 1, right. And similarly, if you try to take the negative that first data vector here is 1234 and second one is just negative of this, you can see here these are the 4 points and this relationship is exactly negative and x is increasing, y is decreasing and the value of the correlation coefficient here is minus 1 and the same was shown to you on the R console also, right. So now let me try to give you here one example so that you can understand what is the use of such a tool actually.

So, I am trying to take here one very simple example. We know that in those areas where the temperature is high for example during summer, right, then the consumption of the water increases, right. So, suppose in a city the demand of water is recorded and the day temperature on those days was also recorded and it is our experience which tells us that water consumption increases as the weather temperature increases, but this is our only assumption, we do not know. So, now we would like to see what happens when we try to give any idea from the data set. So, the data on the 27 days is collected like this that data on the water is stored in a data vector water and the data on the temperature is stored in a data vector temp, right.

And you can see here it is like this, it is a paired data that means the first value here in the data vector water and the first value in the data vector temperature that is temp, they are indicating that this is the first set of observation from the first day. Similarly, from the second day the paired observation is at the second position in the data vector. Now we would like to see that what is their relationship. Although as of now we have not done the graphical part, but still you can see suppose if I want to plot data between temperature and water, it will come out to be here like this on the x axis temperature, y axis water and it will see here this is here like this. So, one can see that here there is a linear trend, right.

Well, there is some variation because the observations are not lying exactly on the line, but at least it is apparent that there is an increasing trend, right. And even if you want to see how much the line deviate or how much the observation deviate from the line, then we have a command here scatter.smooth which will give us also a line. So, between temperature and water data, if I try to plot this smooth scatter plot that is by using the command scattr.smoth and inside the parenthesis the two data vector temp and water we get here this type of graphics.

So, this is a temperature on x axis, water on y axis and this is here the line of this data set. Still you can see there okay, the line is not 100% linear. Well, this will not really happen in real life, but it is more or less linear and we are convinced that there is an increasing trend. But now how to measure it? So, if you try to find out the covariance between water and temp, this is coming out to be like this. But here we are interested in knowing that the sign is positive and this was also indicated in this figure and in this figure also.

So, this is how you can see that whatever the data is trying to inform us that can be confirmed with the numerical as well as graphical devices. And ideally, I always suggest that when you get the data then you should use the graphical devices as well as the analytical tools together to understand this thing. Now if you want to find out the correlation between water and temperature then this value is coming out to be here 0.95. And you can see here that this is the, it is apparent that the data is not that much far away from the line.

So, let me try to show you these things on the R console so that you get here more convinced with these things. So let me try to store this data water on the R console and this data on the temperature. And you can see here that if I try to make here, I think we can see here this is the water data and this is here temp data. And if I try to make a plot between here temp and water it will have come out in here like this that you can see here this is the plot. And if you want to make here a scattered smooth plot let me try to copy this command here to avoid any typing mistake.

You can see here like this. You can see here this is here the plot. And similarly, if you try to find out here the covariance between these two data vectors it is like the 39099.5 and if you try to find out the correlation by using the command COR it is 0.9567. And if you try to extend this concept new way for example, this concept of correlation coefficient can also be extended to a multivariate case.

For example, the way you have extended the case of this mean and various to mean vector and covariance matrix. Similarly we have a concept of correlation matrix. For example, if you have here two random variable X and Y then we know that the correlation coefficient between X and X will be equal to 1 and correlation coefficient between Y and Y is going to be 1. But their covariance will have some other value usually not 1. So, if I try to create thus a matrix exactly on the same concept that if I try to see here there are two variables X and Y and I try to write down here 1 1.

What are this 1, 1? This is correlation coefficient between X and X and this is correlation coefficient between Y and Y. And here this will be your here correlation coefficient between X and Y and correlation coefficient between X and Y. The correlation coefficient between X and Y and Y and X they are the same. So symbolically I can write down here 1, 1 R X Y and R X Y. And similarly, if you have here three variables say X Y and here Z then this matrix can be made bigger.

So, X variable, Y variable, Z variable like this and now you will have the correlation coefficient between X and X which is 1, correlation coefficient between Y and Y which is here 1, correlation coefficient between Z and Z which is here 1. And these values will be on the diagonal elements and off diagonal elements will have the correlation coefficient between X and Y, correlation coefficient between X and Z and correlation coefficient between Y and Z like this. Yeah, this will be the symmetric matrix but still I am trying to write down just for the sake of completion. So, practically it will look like this 1 1 1 R X Y, R X Z and R Y Z that is all.

This is a symmetric matrix. So, this is called this is the correlation matrix of the random variable vector X Y Z. So now we come to an end to this lecture. You can see that was a very introductory lecture but it was important because you can see for the first time I have extended the univariate concept into a bivariate and multivariate concepts. The concept of mean from the earlier lecture is extended to mean vector. The concept of variance that we discussed in last lecture is now extended to covariance matrix and the concept of correlation coefficient that is used for bivariate data but now this we have extended to a correlation matrix.

So, what is happening in the correlation matrix? We are getting the pair wise information. If you have three variables then what is the relationship between X and Y, X and Z and Y and Z? You are getting inside the same matrix. Now looking at this matrix you can have some information that what is the pair wise behavior of the data and all these things they can be done graphically also and when we are talking about the concept of correlation coefficient in a multivariate setup then there are some more concepts like as suppose you have three variables X Y and Z and they are interdependent on each other. Note that only they are dependent pair wise. For example, if I say height, weight and age they all are interrelated up to certain age the height and weight increases after that this becomes less and so on.

Similarly, if I add one more variable say here blood pressure that also depends on other factor but in this correlation matrix, we are trying to consider two values or two random variables at a time. But there can be a situation where you want to consider more than two variables under different types of constraint. For example, what is the correlation coefficient between X and Y when the values of Z's are fixed or there can be two groups of variables and every group has more than two variables or there can be one thing that there is one variable which is dependent on several variable like as we have considered the example of say health. Health depends on height, weight, age, blood pressure, blood sugar. Similarly yield of a crop that depends on quantity of fertilizer, quantity and the temperature, rainfall etc.

So, now this yield one single value is depending on more than one variable. If you want to find out the correlation coefficient between a single variable and a group of variables then how to get it done. This can also be done by using the concept of correlation coefficient but that concept is multiple correlation coefficient. And similarly, if you want to fix the behavior of one random variable and you want to find out the correlation coefficient among other variable we have a concept of partial correlation. So, as we grow further in this course as we try to cover more topics, I will try to introduce this concept when the time comes but this is how we try to think about the multivariate setup. Multivariate setup does not always mean that the outcome is always going to be in a vectors and matrices format.

Yes, the output can be in the format of vectors and matrices but sometime the output is also in terms of a scalar. For example, you have seen this correlation coefficient as well as the multiple correlation coefficient although we have not covered it but it is a scalar value. But it is trying to give you a multivariate idea. So, you try to practice these things, try to understand, try to grasp and try to create a vision inside your mind that how these things will look like and if I try to look on these values what I can infer.

For example, in the correlation matrix I can get the behavior of x, y and x, z. Now y, z is also there. So, it is giving you a pair wise idea but now what is the total idea? These types of things very much depend on your capability and this capability can be enhanced by practice by dealing more real datasets. There is no magic that this can be done in a day or a two. The more you practice, the more you deal with the real dataset, the more you try to observe the process better you will learn. It is just like as a very good doctor, very experienced doctor as soon as the patient goes to the doctor just by asking couple of things medical doctor can diagnose what is happening inside the stomach.

The doctor never opens the stomach in the first shot, first step. So similar is the role for the data analyst also. By looking at the symptoms of the data, by looking at the various characteristics of the data, the data analyst should have a capability to understand what the data is trying to tell us. Why? Because data cannot speak or I would say the data speaks but we do not understand its language and in statistics we are trying to learn the language of the data. The better you learn the language of the data, better we will understand.

So, you try to take some dataset and practice and I will see you in the next lecture. Till then goodbye. Thank you.