Multivariate Procedures with R

Prof. Shalabh

Department of Mathematics and Statistics

IIT Kanpur

Week – 08

Lecture – 36

Scaling of Data: Centering, Scaling and Z-Scores

Hello friend, welcome to the course Multivariate Procedure with R. In this lecture we are going to talk about a new topic that is Scaling of Data. What is scaling of data? When we are trying to deal with multivariate procedures, this means we are going to have more than one random variables and we are going to collect the observations on those random variables. These random variables may have the same units in which the observation have been measured or they may have different units. For example, in case if I take suppose I take three variables weight of human being, weight of vegetable and say weight of some students right. Now all the three units they will they can be measured in the same units kilogram.

Similarly if I say the length of a electric wire, length of road and say length of a room they all can be measured in meters. So, they will have the same units, but suppose if I take three different variables like as height, age and weight. So, height is going to be measured in say for example, centimeters, weight in kilograms and hemoglobin level has its own units in terms of levels. So, now in first cases sometime it is needed that all the data should have the same units, but now once that once the random variable have been decided which are going to affect the process you cannot change them.

So, the next option is given that data set can we make them unit free. For example, height, weight and age they will be transformed in such a way such that height, weight and age will have no units. When they have no units that means, their values can be compared and this is the requirement in several statistical procedures. I agree that if you try to transform the original data set the conclusions the statistical conclusion based on any statistical procedure they may or they may not change, but that is what you have to see when you try to use such a scaling of the data or you employ any scaling procedure to

normalize your data. So, now another thing is that sometime for example, you have seen that in many probability distributions we try to assume that their mean is 0.

For example, many times we say that that the random sample has been observed from a normal population with mean 0 and variance sigma square. Now, suppose in a real life the mean is not 0 then how are you going to make sure that the mean of those random observations is 0 in the sense that they have been obtained from a population whose mean is 0. So, these type of scaling procedures and normalizing procedures are used in statistics and in a different multivariate procedures sometimes they are needed. For example, if you remember when we did the turn off phases then there was a condition that the data has to be normalized. Similarly, in other multivariate analysis procedure with that we are going to handle either you will have this option that you can use them by standardizing the data or scaling the data or use the original observations and their outcomes may or may change.

But this is the need of the experimental which will decide also that whether it makes sense to scale the data or not. But my role here is only one to make you explain or to make you understand that how are we going to do the scaling and normalization of the data. So, that in case if you need you can do it, if you do not need you need not to do it. So, let us try to begin this topic and let us try to understand how we can scale or normalize the data in different situation in different setups. So, now we are going to talk about three topics centering, scaling and studies course in this lecture well studies course is normalizing right.

So, why this scaling is needed? So, many times the data is collected on certain random variables and these data are collected on some random variable which have some measurement units. For example, if the random variable is high then it is measured in centimeters, if the random variable is weight this is measured in kilograms etcetera. And suppose we want to make these observations unit free. So, for that we use the scaling of the data. This concept will make the observation unit free.

So, scaling is that a technique for comparing data that is not measured in the same way, but different variable have been measured in different way different units right. So, the scaling procedure or the standardization procedure they transform the data. So, that the new data will have a mean 0 and the standard deviation of 1 right. And this type of standardization when we are trying to mean 0 and standard deviation 1 this is also called as z score right. If you try to see you had done z test in which you had taken x bar minus mu upon sigma by root n with statistics.

So, in this case what we have done actually we have the data x1, x2, … xn and we had transformed the data is say x i minus mu by sigma right. So, this is called actually here as a credit score. And if you try to do it then x1, x2, … xn may if they have normal mu sigma square then z1, z2, … zn they will have normal 0, 1 distribution right. So, this is z scores are obtained by here like this means x minus x bar upon standard deviation right. Some other procedure where we want to make only the mean to be 0 then they are obtained by x minus x bar of this observation the mean is going to be 0 right.

$\frac{x-\bar{x}}{sd}$, $x-\bar{x}$ or $\frac{x_i}{\frac{1}{n}\sqrt{\sum_{i=1}^{n} x_i^2}}$

Means if you try to take x1, x2, … xn and try to transform them x1 minus x bar, x2 minus x bar and here xn minus x bar. Then these observation they may have some non 0 mean, but they may have mean 0 right. And similarly the third procedure which is xi upon 1 upon n square root of summation of sum of squares of the observations. This is another technique for the standardization of the data. So, these are different types of procedure what we are going to discuss in the next couple of slides.

And we will also demonstrate that how they can be used in the R software right. So, if you want to transform the data in any way or you want to standardize the data or you want to center the data or you want to scale the data in R software then we have a command here s c a l e, scale. And scale has different types of options by using them you can achieve what you want like as only scaling only centering or both right. And this can be used on different types of object or data vectors, matrix, data frame etcetera. So, if you want to use this scale function then the command here is s c a l e and after that you have to give here the x which is a numeric matrix or a data vector or a data frame also.

I will try to show you with the different examples. Then we have the option here center c e n t e r and this is a logical variable which takes value TRUE or FALSE. So, this is as the name suggests centering this is used for centering of the data centering means x minus mean. So, that all the data is center and the mean becomes 0. So, this center is used to decide whether to subtract the mean values and scaling.

And the default here is TRUE and then the centering is done using the root mean square. Root mean square is what I have explained you 1 upon n summation xi square this quantity. And if this is FALSE then no scaling is done. Similarly there is another option here is scale this is also a logical variable here TRUE and it decides that whether to divide the observation by the standard deviation while doing the scaling and the default is

the TRUE right. And if TRUE is given then the scaling is done by dividing the center columns of x by their standard deviations.

Well I will try to show you and write and I will explain you how what procedures are being followed inside the software right. So, the value of center actually determines how column centering is to be done right. If a center is a numeric like vector with the length equal to the number of columns of x then what happen the each column of x has the corresponding value from center subtracted from it. And if center is given the value here TRUE then the centering is done by subtracting the columns mean of x from the corresponding columns. And in case if center is equal to here FALSE then no centering is done right.

And then yeah means here this a unavailable observation and is they are omitted. Similarly the value of a scale determines how column scaling is done after centering. And if a scale is a numeric vector whose length is equal to the number of columns of x then each column of x divided by the corresponding value of the scale. And if the scale takes value TRUE then the scaling is done by dividing the centered column of x by their standard deviation. And if center is TRUE then it happens and in case if center is FALSE then no scale is done and when center is TRUE then we use the root mean square to divide the values.

So, how it is going to be do how it is going to be done let me try to take very simple example and I try to explain you. First way I try to show you on the slide then I will try to show you on the R software also. So, let me take a very simple example in a data vector I just take 4 values 1, 2, 3, which are stored in a data vector x. Suppose you try to find out its mean it will come out to be 2.5 if you want to find out standard deviation it will come out to be a 1.29 which is obtained by square root of variance of x. Now, if you try to compute this quantity x minus mean of x and square root of variance of x. So, this I am writing here x minus mean of x divided by square root of variance of x then this value comes out to be here like this. So, this is a generalization of the data or the scaling of the data, but now let me try to see the same thing if I want to obtain in the R software how it is going to be obtained. So, I try to take here the same data vector 1, 2, 3, 4 and I try to use the scale x.

So, this is going to give me the outcome here see here x minus x bar divided by standard deviation and you can see here this is how the outcome will look like you can see here minus 1.16 minus 0.38 plus 0.38 and 1.16. So, you have to just remember minus 1.16

0.38 and plus there will be the positive sign and what you have obtained here earlier 1.16 0.38 with minus sign and the same quantity with the plus sign.

So, you can see here this is the same result what we had obtained by manual computation, but now it has been obtained directly by the by this scale function right and then it is it is trying to give you here the scaled that means, it is trying to inform you that it is centered. So, the arithmetic mean here is 2.5 which is which you had obtained here earlier and then it is giving you the scale value which is here 1.29 which is the standard deviation which we had obtained here you can see here 1.29 and here also 1.29.

Now in case if you try to use here the complete command which is advisable actually then you must use the scale and center both the option which can be written here by here scale of x scale is equal to TRUE and center is equal to TRUE. Yes, you have to be be watchful and careful here that scale is also the R command and scale is also a parameter to be given inside the parenthesis right. And similarly if you use only here scale x scale is equal to 2 then both this option both this commands will give you the same outcome like x minus x bar divided by s t right. So, now you can see here this is the outcome I can show you here. So, that this is my here data whose mean is 2.5 and variance is 1.29 and when I try to scale them by x minus x bar upon s d this comes out to be here like this and when I try to use here this is scale function.

So, you can see here this function is matching with this function this is matching with this this is matching with this and this is matching with this right. And in case if you try to use here scale is equal to TRUE center is equal to TRUE or only scale is equal to TRUE you can see here that these values are the same right. But my advice to you all is that to use this one both scale and center options are clearly mentioned here right. So, let me try to first show you these things on the R software also.

So, let me try to have this data set here. So, now let me come to the R software and I try to give you here the data here x equal to like this. So, x here is like this and if you try to find out this command here that where you are trying to write down here the square root of mean and variance of this thing. So, let me try to execute it here you can see here it is here like this. And on the other hand if you try to use it as the function here scale you are getting here the same value right that you can see this value, this value, this value, this value and remaining the two values.

And similarly if you try to obtain here these two commands also here then also you will get here the same value. For example, if I try to execute you can see here this when you are trying to say scale is equal to TRUE and center is equal to TRUE it is giving you here

these values. And if you are using only scale is equal to TRUE then it is giving you the same values right. So, that is how you can see that it is not a very difficult thing and when you will be using different types of multivariate procedure then there will be some option which is given to you that how are you going to whether you want to scale the data or not. Now, I try to use here one more option of this scale function that scale is equal to FALSE.

So, when you try to use here scale is equal to FALSE then it will only do the same thing that means, x minus x bar. So, now, what will happen? The mean of all the observation will become 0. So, if you try to see what I am doing I take here the same data 1, 2, 3, 4 its mean is 2.5. So, if I try to subtract each of the data by 2.5, 1 minus 2.5 this will be here minus 1.5 right 2 minus 2.5 this will be your here minus 0.5. Similarly, here 3 minus 2.5 this is here 0.5 and 4 minus 2.5 this is here 1.5. So, every observation has been subtracted by its mean and the same operation can be done in the R software if I try to write down here scale x with an option that scale is equal to FALSE.

So, you can see here this is my here data set. So, this value if you try to match this is coming here this value it is coming here third value it is coming here and the fourth value it is coming here. And what is here this 2.5 this is here the mean which we have obtained here right. And if you try to hit on the R software also it will here look like this is your here data vector and with this I am trying to obtain it manually x minus mean of x this values are obtained. And then I try to use here scale function we can see here this is matching with this, this is matching with this, third value is matching with third value and the fourth value is matching with the fourth value right.

So, let me try to show you these things on the R software also. So, that you can be confident right. So, you try to see x is here like this and if you want to find out here x minus mean of x this is here like this. And if you want to use the scale function it is giving you here the same value right. So, now, you can be confident that it is working in the same way as I explained you on the slide.

So, let me take here this same data vector x equal to c(1, 2, 3, 4) and I would like to use here another option center is equal to FALSE in the same data vector x using the same function here same command scale. So, if you try to use this scale on a data vector here x with the center is equal to f that is FALSE then it gives you here this type of value every observation is divided by 1 over n minus 1 square root of sum of i goes from 1 to n x i square. That means, you have to take the positive square root of the sum of squares of all the observation and divide them by the sample size minus 1. So, if you try to see here in

this case I try to take here x equal to like this then square root of sum of x into x divided by 3 which is here 4 minus 1 it comes out to be here 3.16 and if I try to divide x divided by this quantity then it will come out to be here like this these are the 4 values which we are going to obtain.

And if I try to do the same thing in the R software also using the command here scale then it is coming out to be here like scale x center is equal to FALSE and you are getting here the same value. So, if you try to see here and try to match here what are you going to get the first value is matching with the first value here the second value is matching with the here the second value then third value here is matching with the here third value and the fourth value here is matching with the fourth value here. And this quantity here scale this is here matching with the square root of sum of x into x divided by x3 right. So, you can see here that the same thing has been done here without any problem. So, this is another type of scaling and yeah I would try to show you this on the R software also.

So, let me try to take here right my x is here already here like this and if I try to say here like this the scale function it is giving me this thing and if you try to use here the same command which you have written manually yourself it will also give you the same value right. So, you can see here it is not difficult to do this ok. So, now I have explained you these things over a data vector. Now I try to repeat the same operation over the data frame because data frame will also play an important role when you try to do the multivariate statistics. So, if I try to take here two data vectors one is here x is c(1, 2, 3, 4) and another I am taking here y is c(10, 20, 30, 40).

So, that it is easy to remember you that which data is corresponding to which variable. So, the first variable is 1 2 3 4 and second is 10 20 30 40. Now I try to create the data frame using x and y using the command data.frame like this and we store it in a variable d f x y. So, now you can see here this is the data frame that we have obtained right. Now I try to use the same command which I have used earlier and try to see what happens.

You can you will see the similar outcome will come, but now in a different way all the scaling is done for each and every variable. So, if you try to use here similar to scale of x if you try to use this command here scale of d f x y then it will give you the value of x minus x bar divided by s day that is standard deviation for every observation in each of the variable. So, you can see here this is the value for the first variable x these are the values for the second variable and they are now normalized. So, these are actually the z scores and you can see here this is here giving you the values of the means of the x and y which have been used in the squinting or say scaling right that is x minus x bar.

So, for x the mean is 2.5 and for y the mean is 25.0 and similarly the standard deviations have for x and y they have been obtained here like that. Now I have done it for using the command scale, but if you want you can also do it using your writing your own command the way I have shown you earlier. Similarly if you try to use the second command that scale earlier we had used x and scale is equal to FALSE. So, now I am using it over the data frame. So, this is going to give us the centered value for each of the column that means every observation for each of the variable will be centered by x minus its mean.

So, for the column it will be every observation minus its column mean right. So, you can see here this is the center observation for x and these are the center observation for y and the sample means are obtained here 2.5 and 25. So, for x every observation is centered by x minus 2.5 and for y every observation is centered by y minus 25. So, you can see here this option is also working in the data frame also right. Similarly if you try to use here in the earlier command scale df xy with option that center is equal to FALSE. So, you can see here that it will give you the similar value that xi divided by 1 upon n minus 1 square root of sum of squares of all the observation for each column. So, you can see here this is the standardized value which are obtained using this function and these are the standardized value of y which are also obtained using this function column wise. And these are the scale value which have been used in this data frame which are essentially the values of 1 over n minus 1 summation xi square and this is square root and 1 over n minus 1 summation yi square and this is square root right.

So, these are these values. So, now you can see here this is the screen shot of the same observation I explain you. So, here I am trying to create the data frame then I am using only scale df xy or you can also use the scale is equal to TRUE and center is equal to TRUE which I always recommend and then I am using here the scale df xy with the scale equal to FALSE and the next command is this center is equal to FALSE. So, these are its outcome, but I will try to show you on the R console also. So, let me try to first create this data frame.

So, this is my here data frame you can see. Now I am trying to use here say scale of df xy you can see here these are the values which are the same values which you can see here in the slides I have obtained here. Now I try to use here the next command scale with scale is equal to FALSE option. So, if you try to see here it is here coming out to be here like this where every column value have been subtracted by their column mean and now if you try to use here another type of this scaling then if you try to execute them it will it is giving you here this value. So, these are the same value which I have shown here on the screen shot also and in this slide also.

So, you can be confident that these things are working. Now I try to repeat the same thing for a matrix because you will see in multivariate analysis many times you have to give the data in the form of a matrix. So, you have to be confident that these procedures are working over the matrix also. So, let me try to create here a 3 by 2 matrix of the values 1 2 3 4 5 6 using the command matrix and row equal to 3 and col equal to 2 and data is equal to 1 to 6 and I store this data into this matrix xm. Now I try to repeat the same thing what I have done earlier for the matrix data also.

So, first I try to use here scale xm. So, it will give you here these values which is here that x minus mean divided by standard deviation for each of the column. So, this is for here the first column this is here for the second column and these are here the mean values column means and these are the standard variation for each of the column. Similarly, if you try to use the earlier command for centering scale then this matrix data x m then scale is equal to FALSE. That means every column will be subtracted by its by the arithmetic mean of the column means. So, you will see here this type of outcome will come here that this is the output for the first value and this is the output for the second column and these are the arithmetic mean of or the column means of the data in the two columns.

And if you try to use here the earlier command that scale say xm center is equal to FALSE then using this command here you can get here this data set. So, these are the values for the values in the first column and these are the values for the values in the second column. And value of the denominator in the two columns is obtained here like this. Well, it has these are very straight forward things and you can see here this is how I have obtained this is my here data this is my here scale function. Then I try to use here scale is equal to FALSE then I try to use here scale is equal to center is equal to FALSE.

So, if I try to show you these things on the R console also then you will feel more confident. So, let me try to create here this matrix here first. So, you can see here this is my here matrix and if I try to say here scale is equal to here x m you can see here like this. And if I try to give here with another command.

So, using this only scale function will essentially give you the credit score. So, many kind people will ask you please standardize your data using the normal score or credit score that means you have to simply operate the scale function. Now if you use here scale is equal to FALSE then it is going only the centering and if you try to use here center is equal to FALSE. Center is FALSE means only it has to be means no centering is required only scaling is required. So, if you see here it will come here like this. So, you can see

here it is not a very difficult thing to do and now with this we come to an end to this lecture and you can see here it was not a very difficult thing to do, but the main idea is that you are trying to transform the data in a particular way and the need whether you want to do only centering, only scaling or only this divided dividing by the standard deviation or this root mini square deviation.

It depends on the statistical function which you want to use and what is the requirement of the data set. And when you are trying to do it now the data is going to change its nature. So, it is possible that the statistical conclusions may also change. So, in my personal experience the way I do is that whatever I am trying to do with this statistical procedure I try to do it twice. It takes only a couple of seconds or a couple of minutes only and I try to get both the results with standardizing and without standardizing.

Usually they are different, but then I try to interpret them that with which of the case the results are matching with the TRUE value that means what is happening in the real experiment. And then I try to take a logical discussion in consultation with the person who has collected the data and who has conducted the experiment because they are the people who are going to tell me that if I try to do the centering, scaling etcetera is it really going to make any sense or not. Well, I am not going into this discussion that whether you should use or not, but anyway this is the requirement of the statistical procedure. So, that is why I just covered it here so that you are well aware of this scaling and normalization of the data including the centering. So, I would say that whatever we have done in the past, we have done now lots of statistical procedures.

Why do not you try to repeat all the things with the after scaling the data, you try to scale them, you try to standardize them, you try to splinter them and then try to see what is happening. Definitely it is going to take some time, but definitely without practice you cannot learn. So, my advice to you in this lecture towards the end is that do not do anything new, just try to take the earlier tools whatever you have done either that is me whether that is standard deviation or t test or anything and try to repeat those things whatever you have done with the scale data. In some statistical function in the R software, there can be an option like scaling is equal to or scale is equal to TRUE or FALSE, but if that is not there you can simply store the new data which is obtained from this scale function directly into a data vector or a matrix or a data frame and then you can use them directly.

So, it is not difficult at all. So, my request to you all is that please try to have a look, please try to exercise these concepts and make yourself learned and well aware with the

statistical tool and I will see you in the next lecture with a new topic on regression analysis till then goodbye. Thank you.