Multivariate Procedures with R

Prof. Shalabh

Department of Mathematics and Statistics

IIT Kanpur

Week – 01

Lecture – 01

R Software and its Installation

Hello friend, welcome to the course multivariate procedure with R. So, as the name suggests, we are going to learn the topics of multivariate analysis and their implementation in the R software. Now, the question comes that what are we expected to know in the R software and in statistics. So, the way I am going to handle this course is that in this lecture and in the next couple of lectures, I will try to give you a brief background about the topics in the R software which are going to be helpful in this course. Well, I am not saying that you have to learn only the selected topics in the R software. You have to be well versed with the R software So, that you can learn the course in a better way.

But still there is a minimum requirement to begin the course. So, to help you in this regard, I have compiled couple of common elementary topics from the R software So, that me and you both are at the same platform when we formally start the topics in the statistics. Now before we try to learn these topics in the R software or in the statistics, let me try to give you some idea that what are we going to do. You see as the name suggests multivariate procedures.

Multivariate means there is also a univariate procedure. So, in statistics whenever we are trying to do any data analysis, the data analysis is done over some outcomes. And you will agree with me that in real life the outcome is affected by more than one variable. For example, if I ask you a question that can you judge that somebody is healthy or not, So, how are you going to judge it? You would like to look at the blood pressure level, blood sugar level, body temperature, weight, weight and other body parameters. And if all of them are in control within the given range, then you would classify the person as healthy.

So, now in this example if I ask you that if somebody has got a very high blood pressure but his blood sugar level is in control, will you call him or her as healthy? And I mean to say that the outcome of a single variable may not be able to give you the correct judgment. The judgment is given by when you try to consider all the variables together. Usually when we try to understand the yield of a crop that how it is going to affected. The yield of the crop is going to be affected by the quantity of fertilizer, temperature, rainfall and many other factors. Now in case if you try to say that only the quantity of fertilizer is going to affect the yield of a crop, do you think that is it a valid statement? Not really, but the yield of a crop is affected by more than one factors, more than one variable.

So, obviously there can be many such a situation in the real life where you would like to consider more than one variable which are jointly affecting the outcome. Now when I use the word jointly, So, that means this increases the number of parameters to be considered. For example, if I am considering only one variable, then for example mean and variance may help us in describing the population. But if I have two variables, then their joint variation in terms of covariance or correlation coefficient they are also required. So, when we try to extend the univariate analysis to a multivariate analysis, then whatever was required to understand the univariate analysis more than that is required to understand the multivariate analysis.

Now when I use the word multivariate analysis, those who are coming from the statistics background, they will possibly understand that I am talking of the course multivariate analysis. There is a standard course in the MSc programs in statistics about multivariate analysis. All those procedures they are covered in the multivariate analysis which I am going to do here. But when I try to think about students from other sciences like as management, economics, psychology, etc., then they are using different types of multivariate procedures which are taught in different courses in the BSc and MSc Statistic Program in statistics.

So, my request to all the students will be that please try to take my own multivariate analysis in this way. So, that is why the name of the course is kept as multivariate procedures. For example, the multiple linear regression analysis, this is also a multivariate procedure. This is taught in the course of regression analysis, cluster analysis, discriminant analysis, classification problem, factor analysis, etc., they are taught in the multivariate analysis course.

So, in this course, we are going to compile some popular topics which our students are using in general in various type of data analysis and my objective in this course is to keep the level of mathematics to be as minimum as low as possible. Well, you know that in statistics whatever tools we are going to develop, they are based on mathematics. Without mathematics, we cannot develop a statistical tool. And when we try to develop a tool, there is a step-by-step hardcore mathematics involved and whatever is the outcome of that analysis, this is used by different people for the statistical analysis of the data set. So, I want to make it clear here that I will try to strike a balance between the details of the tools and application along with its implementation in the R software.

I believe and hopefully you will agree with me that unless and until I give you the basic concept about the statistical tool which you want to use, you cannot use in different applications because I personally believe that one statistical tool can be used in different situations and it depends on the capability of the experimenter or the person who is doing the analysis because they have only a question. Nobody tells them that which statistical tool is going to use or going to help them in getting the answer of the query. This is only the judgment and knowledge of the statistician or the analysis who is analyzing the data that which of the tool is going to help them. So, in order to understand this thing, unless and until you understand the basic fundamentals of this tool that what it is doing, how it is doing, you may not be able to use it at an approved place. You cannot take a decision whether you are using the tool at the right position in the correct way.

Now the tool is mathematical and it has to be exposed over a dataset. So, now if I try to do the manual calculations, yes it can be done but it will take a very long time. So, that is why we need to take the help of some software. So, there are various software which can be used for the statistical analysis of the data and among them R software is an open-source software. This is freely available.

Anybody can download, anybody can use it. So, that is why I have chosen the R software to demonstrate the application of these statistical tools. So, this is how I am going to proceed in the lecture that I will try to introduce you with the basic concept. I will explain you in a reasonable depth So, that you are convinced and you understand that where I am going to use it. Then I will try to give you the basic fundamental that how are you going to use in the dataset and that is and after this the implementation of the same tool in the R software will be demonstrated and for that I will try to use the R software directly inside this lecture.

I will try to use the screenshots from the outcome of the R software and I will try my best to make you understand. So, this is how we are going to follow this course and this is what is our objective in the beginning we want to achieve at the end of the course. So, now in this lecture I am going to give you a brief introduction about the R software that how are you going to get it and why R is important, why R can be used for statistical analysis. So, with this small very elementary storytelling lecture I would like to begin this course. So, let us begin our lecture now.

In this lecture I am going to talk about the R software and its installation. So, now the first question comes here that what is R. So, R is basically an environment for the data manipulation, statistical computing, graphic display and data analysis. It is just like any other good software. There are many software which are available for mathematical and statistical analysis they can and similar is the R software also.

Effective data handling and storage of the output is possible just like any other software. Simple as well as complicated calculations be it simple calculation or simulation Monte Carlo simulation etc. all are possible and they can be done in the R software. Just like any software will also give us different types of graphics then R is also capable of doing the same thing. R can display the graphic on the screen as well as their hard copies are also possible that they can be stored in different formats JPEG, PDF, PS etc.

Any software has also a programming language So, R also has a programming language which is quite good, quite effective and it has several advantages over other type of languages and it includes all sorts of possibilities. Whatever you can do with any other programming language the same thing can be done in the R language also. I would just like to give you an idea that this R language is very similar in appearance to the S language. Really earlier there was a software S plus this was a very good software or rather this is a very good software but it is a paid software that means you have to pay if you want to use it you have to purchase it. So, then people started developing a similar software to this S plus software and this came in the form of R software.

So, the programming language that was used in the S plus software was called as S language. So, similarly in the R software this programming language is called as R language. Now the next question comes here why should you switch to R? I am sure that most of you must be using some statistical software, some programming language then why should you come to learn this R and why should you do your analysis in the R software right? So, just to give you a brief background and some information about this

software I would like to inform you that many people researchers, design officers, analytics firms they already have started using R and there are reasons. The first reason is that it is a free software, it is an open-source software, you do not have to pay anything and if you modify any programming according to your need this can be done and hence this is not a black box. How do you mean by this black box? Suppose if you have a software and you suppose you want to modify the programming then you cannot do it.

The company which has developed the software only they can make the changes. You can only use it in the way they are asking you but if you want to make any modification according to your need usually you cannot do it but in R software it is possible because it is an open-source software or its codes are freely available So, you can go inside the codes and you can modify the programming according to your need. And now as of now more than or say nearly 20,000 statistical packages are freely available to work in the R software. What does this mean? That every package is trying to facilitate some type of statistical analysis. Suppose you want to do a time series analysis then there will be a package you want to do clustered analysis; there will be another package etc.

So, different people from all over the world and different researchers they are trying to develop these packages as and when they try to develop some new tools. And they can upload it on the website of R software and so, this is how this number is continuously growing. So, at the moment to the best of my knowledge there are about 20,000 freely packages which are available. And in R software when you try to download the basic version some of these packages are built in and some of the packages which are contributed packages. Contributed packages means if I am doing my research and I have developed something new then I can do the programming and create a package and I can upload it, get it uploaded on the R software website which other people can directly download it and just by giving the data as input they can use the tool.

So, different people around the world, different researchers they are contributing in this repository and that is how these more than 20,000 packages are freely available, right. And you can also contribute your own packages also. And most over R has a statistical computing environment that means you can do different types of mathematical operation, different types of statistical operations inside it and as I said it is a free open-source software but it is not a black box. This is the biggest advantage. And the language in R that is what we call as R language this is not difficult.

It is just like means any other programming language in the computers and which is very convenient to use in statistics and graphical applications. You will see that there are

packages and then you can do your programming and it is not difficult to understand and learn that how to get it done. There are beautiful graphics. They can be created in this software and graphics can be directly saved in post script format, PDF format, JPEG format and they can directly be copied from the R console and can be pasted in other applications, right. And you are going to use different types of syntax commands to execute your analysis.

Those commands can be saved, they can be run, they can be stored in a script file. Script files means something like what we call in the common language as programs, right. And say different types of software they are available in our different platform like as Windows, Unix, Linux, Macintosh. So, similar is the R software also. This is also available in all such popular platforms.

And you can download it freely depending on your platform from the website of the R software. So, now you may consider that you should also start using R and you should assist to the R software. Now once you are convinced that yes, I want to use it, the next question comes how are you going to obtain the software? So, this software is freely available at this website www.

r-project.org. And if you try to open this website on your computer, it will look like this. That is the screenshot I have taken here to explain you. And here if you come here, you can see here this is here the address of the website as soon as you open it, there will be a section here, this is a line here where it is written download R, right. So, if you just click on this download R here, right, then you will come to next screen. I will show you and beside these things there are different information which is given on this webpage that is if you want you can have a look.

But as soon as you click on the download R, you will come to the next screen which will look like this. What is this screen? If you try to see, there are here different names of the countries Argentina, Australia, Austria, Belgium and So, on and this list will continue. And in every country, there are certain addresses which are given here like this. So, actually this is called as CRAN.

This means Comprehensive R Archive Network. So, what really happened then that when this R software was created, then it has to be uploaded on a website. So, if it is uploaded only on website and if there are large number of people who are trying to access it, there is a possibility of the server getting crashed. So, people decided to distribute the

load and different institutions agreed to host this software on their websites. So, that is why different institutions in different countries, they are hosting it and their addresses are given here.

You can choose any one of them. All the softwares are going to be the same. So, whichever address you want, whichever country you want, you can just click on this address, any one of them and then you will come to here this site. As soon as you come here, you will see here like this. This download R for Linux, download R for Macintosh, download R for Windows. So, whatever is your platform on your laptop or computer, you can choose the appropriate version from this site.

You just have to click on this and after that you will download a file and you simply have to say click, click, click over this file and it will install the R software on your computer right. So, now this R software is available on your computer. After this, you have two options to work. You can work correctly into the R software inside the R console or you can use some other software which are trying to provide you a helping hand to work in the R software. So, there are several such softwares which makes your life easier to work in the R software.

For example, this R studio is a popular software nowadays which is helping you and it makes your life easier when you try to execute the program or you try to do any analysis in the R software. And similarly, there are other software also like another popular software is TinR. So, this is available on this website right. Remember one thing, I am not trying to propagate or advertise any particular software. My objective here is only to inform you that these are different types of software which are available on different platform depending on your convenience, depending on your choice, depending on your interest, you can use them.

And yeah, means both these softwares which I have told you R studio and Tin, they are the free software. Well R studio has a paid version also, but usually the free version is enough for us to use in this course right. So, this R studio is actually written in the C++ programming and it is a free and open-source integrated development environment IDE for R and you can download this R studio software from this website. So, as soon as you click on this website, you will get a site like this one and this is about R studio IDE and here you can see here there is a point here download R studio, here you can click and you can see here or here also you can click it, that is the same thing. Now you can choose here whether you want to download R studio or you want to download R studio for server.

So, depending on your need means you can download it and after this you will get file, you can save it on your computer or laptop and you just double click on it and it will install the R studio software in your computer right. So, now after this thing I think in one of the next lecture, I will try to demonstrate that how you can use the R software. But as I said in the beginning in this lecture, I wanted to give you some idea that why should you use the R software and how can you get the software and its related components. So, that in the forthcoming lectures you can move along with me. So, you try to install this R software, R studio software on your computer and laptop.

With that I would like to clarify one thing here. In this course I will be working only on the R console. The reason being today R studio is popular, tomorrow 10 R will be popular and after that some other software may come. So, I would not like you to be dependent on the software because these are the supplementary software from time to time you will get better software also. But the R software and its console will remain the same. So, if you try to learn the topic by using the R console only then your learning will become better.

Whatever basic commands we are going to type in the R console they are for example available just by a click in the R studio software. But when you are trying to write your own simulation program then in that case you will require to know that what is the command which I have to type inside my programming file So, that the execution can be done automatically. At that stage instead of using the button on the software to click and execute your things what you want you will require the basic command So, that you can complete your program and the program can be executed in an automated way. So, that is why I would like to emphasize here that during the course you also try to use the R console, it is not difficult. And parallelly you can also experiment the same thing in the R studio or any other software whichever you want whichever is your choice, no condition from my side.

And if you try to learn all the things by executing them in the R console as well as into other software there is no doubt that this is going to be a better option. So, you try to install this software on your computer laptop, try to brush up your knowledge about your R and R studio working and I will see you in the next lecture where I will try to give you some more information about the R software till then goodbye. Thank you.