Multivariate Procedures with R

Prof. Shalabh

Department of Mathematics and Statistics

IIT Kanpur

Week – 03

Lecture – 11

Missing Data Handling

Hello friends, welcome to the course multivariate procedure with R. Now, in this lecture, I am going to introduce a new concept. This is about how to handle the missing data. What is missing data? Whenever we are trying to conduct any experiment in which we are trying to record some data, then it is possible that some of the data may get missed that may happen during the experimentation or in some cases after the experimentation also. For example, if an experiment is conducting a clinical trial in which suppose there are 20 patients which are given some medicine and they have to report to the clinic every day. In several days, they will be given a dose and next day their recordings will be ah recorded in the clinic.

Now, suppose there is one patient which starts taking the medicine on first day, second day, third day, but on fourth day the patient does not report in the clinic and the patient comes again to the clinic on the next day. In this case what will happen? The data on that particular day will not be available and this may happen on multiple days and with multiple number of patients. In another example, suppose we are trying to conduct a crop cutting experiment means in some agricultural field, some crop is there and we want to experiment in that field for that we have to cut some of the crop or we want to estimate that how much crop is going to be produced after the season. And suppose one day some cattle some cow comes and the cow it eats some of the part of the field.

Now, what will happen? There is no crop in that part of the field, but the remaining part is available. So, in these types of example what is happening that some part of the data sets gets missed. Now, you have two options either you try to ignore all sets of data and you try to conduct the whole experiment once again. In some cases it may be possible, in some cases it may not be possible, but it will always be an expensive affair which will take more cost, more time, more energy, more labour. So, now what to do? The next option what we have that whatever data we have we should use it and try to ignore the missing part.

Well, in statistics we have got a concept of missing data models in which we try to estimate the missing value on the basis of given set of data and we try to impute it back. That means whatever the value we are trying to know on the basis of given set of data that is replaced back in the sample. Now sample looks complete as if there was no missing value and then we try to conduct over statistical analysis, but definitely we are not going into that aspect in this lecture. But our main objective in this lecture is that how to handle the missing values in the data and how we can handle them, how we can compute our statistical quantities when there is missing data. So, what I will do? I will try to take different types of example.

In this lecture I am going to explain you how are you going to handle the missing data. Then in the next two lecture I am going to repeat the same concept what we have discussed earlier on mean and variance or the measures of central tendency and measures of variation that I will try to show you how you can compute them when we have the missing data. Now I will try to show you here in the case of univariate setup, but without any problem these things can be extended to a multivariate. So, my objective here is to give you the idea that how are you going to handle it, what are the commands and the same commands can be replicated when you have more than one data vectors or matrices or data frames or data in any other format and wherever we are trying to use the multivariate procedure. So, now let us begin our lecture and try to understand how we can handle the missing data in the R software.

So, now about this missing data handling. So, in R software the data which is missing that is indicated by capital and capital A, different software may have different types of symbols and notations indicated, but in R we use the notation here capital and capital A and A. The first question comes whenever data switch comes to B. Usually as a statistician you may not always be compiling or collecting the data, but there will be somebody who will be conducting the experiment and they try to bring the data to us in some format may be in some CSV file, Excel file etcetera. So, now our first step is how should we know if there is any missing data because if there is no missing data and if there is no missing data then our R command for the statistical procedure we have to be modified.

So, in order to know, in order to detect whether there are any missing values or not we have a command here is.na, all in lowercase alphabets. So, whatever is my data set that we try to write inside the parenthesis and we try to write is.na and inside the parenthesis the data vector and its outcome will be a logical vector that means the outcome will come in terms of two or fours. And in all the places where the data is missing that means there is value na this is represented by true.

For example, just to give you an example here suppose if I try to take here one value here x which I am taking it as na and if I try to use my command is.na x then it will come out to be here true. So, now, whatever is the data vector wherever I am getting the true that means the data is there. So, you can see here that when I am trying to use here x equal to na this is assigning na to a variable x and when I am trying to operate this is.na it is giving me here true.

So, is it, it is actually trying to know is any data missing is it missing. So, now, let me try to take here data vector and try to show it. Suppose if I take a data vector of 4 values like 11, 13 and 2 values here na and na they are missing. I am trying to increase the intensity so that I can show you the outcome so that you can understand it easily. Now when I try to operate here is.na inside the parenthesis x you can see here this is my hair outcome false, true, false, true and wherever it is here true you can communicate it to the na value available in the data vector. And wherever we have here you can see here 11 is indicated by here false and this 13 is indicated by here and this is here that is been shot. So, you can believe at the moment that it is going to be correct. So, this is how we try to take the data vector. Now let me try to show you what are the consequences if you have missing data that will give you an idea and this concept can be extended to any statistical measure.

I am just trying to illustrate it through a very simple command mean mean we understand that arithmetic mean. So, if I try to take this and if I try to find out the arithmetic mean. So, basically we are trying to find out 11 plus na plus 13 plus na divided by 4 and this will come out to be here the value na when we try to operate it in the R software. But do you think that this is what you want certainly because if there is a data vector which has got missing values and if you do not know and if you try to compute their mean it is going to give you the value and then it will be used in further in the programming and you will not come to know that what is really happening. But if you try to understand what do you want.

Suppose you want that whatever are the missing values removed from the data vector for example this and this I am removing it and whatever are the values available in the data set we would like to find out the arithmetic mean of those values. So, in order to execute it we try to use there an option na.rm is equal to true. If you say that na for missing value and rm is true you would be wrong. That means do you want to remove the missing values I am trying to say true yes please remove. Now if you try to compute the mean of x with using the command na.rm is equal to true then the outcome will come out to be here 12 which is obtained here 11 plus 13 divided by 2 which is equal to here 2.

So, now you can see that how it is being operated. As I explained in the beginning in the statistics, we have some procedures which are under the missing data models in which what we try to do suppose this is my data vector c 11 and a 13 and a then we try to use some imputation methods and then we try to estimate missing value based on available data then we replace them. So, what will happen here now the new x star will become here something like a c 11 and there will be some value here not na and then 13 and then there will be another value here in case of na and these are going to be some numerical values obtained through imputation procedures. One common imputation procedure is to find out the arithmetic mean and we try to replace all the missing values by the arithmetic mean. So, for example, in this case you can see that the missing value is coming and the value of the remaining values here is 12.

So, one option is that I can use here in place of this missing value I can use here 12. Now this new data vector x becomes here c(11, 12, 13, 12). But I can use my e over this thing but then definitely as I said earlier, I explained you this thing because in case if you want to use such missing data models here it will be in the R software and you can use it also. But here I am not going to use it and my objective is simply to explain you how you can diagnose if there are missing values and how you can find out their position, their details and how you can modify your here computation in the R software. So, look at me try to this illustration if I try to take here x equal to na and if I say here x dot na x you can see that it is coming out to be here true.

And similarly if I try to take here another vector here c(11, 13, 14) you can see that the mean of x comes out to be here 12.5. On the other hand, in case if I try to give here two values and if I make it here NA and NA one value also it does not make any difference then if you try to find out the mean of x this will come out to be here NA. But now in the same command if you try to use here na dot rm is equal to true one typography mistake here like this. You can see here that the mean of x is coming out to be here 13 which is here 12 plus 14 this is 26 divided by 2 which is here 13.

You can see here if you try to you can slip or mistakes then R gives you an error that cannot move forward. On the other hand, if there is a warning message that mean it will move forward. Anyway so if you try to see this is how you can handle this missing value and then yeah if you do that command you have this option which you have to give in the parenthesis and my advice to you all will be that whenever you are trying to use any statistical procedure please try to look into the help and try to see that how the missing values are going to be handled. When you are working in the R software then we have two concept one is here NA and another here is null right. And they are somehow in some common sense I can say that they are trying to indicate that the values are not available.

Values are not available means whether they are missing or not available there is some difference right. So, in R you will see that both this NA and NULL they have a different value and sometimes you will get outcome as NULL. So, let us try to understand what is the difference between NA and NULL right. Actually, this outcome NULL is written by some of the functions and expression when you try to execute it. But if NA and NULL they are not the same.

NA is a placeholder for something that exists but it is missing. Here NULL stands for something that does not exist that never existed at all. For example, if I try to take an example and then we understand it clearly. Suppose there are two persons say person number 1 and person number 2 and both of them appear in the addition test in a screen. Now person number 1 gets selected and person number 2 is not selected.

Now if you try to look in the attendance the person number 2 will be there. Person number 1 or person number 2. The attendance register in the school will have only the name of person number 1 and person number 2 will not be there because that person number 2 is not admitted in the screen. So, now on a particular day if during the attendance this person P1 is absent then the attendance will be marked as P2. And when we try to take the attendance of person number 2, P2 this will be marked here as say null because person number 2 does not belong to that school or that college.

So, this is what I meant that null stands for something that never existed at all. So, person number 2 never existed in the school. So, the name of person number 2 will not be present in the attendance register. So, it will be marked as null. So, this is the basic difference between NA and NULL.

Now let me try to address another aspect of missing data. Whenever we have some missing data then we would like to know what is the location of the missing data or we want to identify the location of NAs. For that we can use the command here WHICHS and inside the parentheses we have to write is dot NA and then inside the parentheses data vector. For example, if I try to take the data vector X like this c(11, NA, 13, NA). So, these two values they are here missing.

So, now if I try to operate this command here WHICHS inside parentheses is dot NA inside parentheses X. So, now you can see here what are the location on which we have a missing data. If you try to look into this data which I have here X in if you try to write down the position of different values this is the position number 1, this is position number 2, this is position number 3 and this is position number 4. So, N A is the position number 2, 13 is the position number 3 and NA here is at position number 4. Now if you try to execute this command it is giving you here 24.

What is this 2 and 4 that is what we have to understand. If this 2 is indicating and this 4 is indicating, so the command whichis.na is trying to indicate that the missing values are at setting and four places. This is how you can understand and this type of information can be used further. Now suppose if you want to operate any procedure or any command over this thing suppose if you want to find out the sum and we want to know that how many missing values are there for example. So, in order to count the number of missing values that is number of N A's we use the command here sum inside the parentheses using the command is dot NA.

So, if I try to write down here sum and inside the parentheses is dot NA then you can see here what we have to do. Let me try to take the same data which is 11 NA 13 NA. So, this is the this thing. So, now if you try to see how many positions the missing values are happening. First position and this is the second position.

So, total number of missing values are of missing values they are here 2. Now if you try to see here sum and inside the parentheses is dot NA inside the parentheses x then it will give you here the 2. So, this is indicating these two that there are two missing values in this data set. So, let me try to show you these commands on the R console here.

So, that you can see here. So, let me type the particular command. So, that you can see here. So, now you see here let me and this is your here like this. If you try to see here. So, now if you see here that the missing values are occurring at first and third position.

If you try to enter you will get here. And so similarly if you want to obtain here the sum of the these values that means how many values are here missing then if you try to see here this will is giving you here. Now similarly if you try to take here means if you try to increase here some more than here NA comma NA if you try to see here and if you again try to find out here this sum if you come out with here 4 because there are 1, 2, 3 and 4. There are 4 values in the data vector which are missing. So, this is how we try to find the sum. Now, because this means there is another objective which you would like to achieve.

Suppose you want to know the complete cases that means what are the values where the data is available. It is up to now we were trying to find out the places or the data which are missing. Now I am trying to concentrate on the data which is available. So, for this I can use the command here complete dot complete.cases and inside the parenthesis you have to write down the data vector. So, this will return a logical vector and it will try to identify the rows which have the complete cases and whenever we have the complete cases if the outcome will be a logical vector which is something like 2 and if the value is missing then it will come out to be false.

So, what we try to take here the same example c(11, NA, 13, NA) and if I try to operate here complete cases inside the parenthesis x then the outcome comes out to be here like this. So, if you try to understand this thing in the data vector x there are two places where the data is available and then you try to see the outcome of complete dot c 11 is indicated by here true and this 13 is indicated by here true. And similarly if you try to see here this n a value. So, this n a is indicated by here false and this n a is indicated by here false. So, by looking at this outcome one can see that the values at the first and they are available they are not.

And now after this I want to go one step further then I will go out to handle the missing values. Suppose I want to ignore all the missing values and I want to consider only the values which are complete. So, after what we have considered suppose we have a data vector which has missing values. So, we use the command na.rm is equal to true and after that whenever we are trying to execute any statistical measure or any function of r it is trying to remove the n a value and it is trying to give the outcome. Now our objective is this we want to create a new vector new data vector after removing the missing values from the original data vector.

So, now how to do it. So, for that we have here the function n a dot omit n a dot o m i t and inside the parentheses you try to write down the data vector it returns the object with the least wise mention of missing values that is it will drop out any rows with the missing values anywhere in them and for this name forever. Let me try to explain this thing with this data set. I think that is the same data vector x =c(11, NA, 13, NA). I try to create here another data set y by writing na.omit(x). So, what will happen now this n a and n a will be removed from this x data vector and we will get here only 11 and 13 which is here like this.

After that we have here some it is trying to tell that this n a values are available at second and fourth position and its class is omit. But anyway we are not going into that part. But now in simple words we have removed the missing values from the data vector x and we have obtained here a new data vector y. Now if you want to make any operation on this data vector y this should not be a problem. For example now if I try to find out here the arithmetic mean.

So, we have here x which has missing value, we have here y which has new missing value. Now if you try to find out here mean of x and mean of y then you can see here that mean of x of this x comes out to be here n a and mean of this 11 and 13 comes out to be here 12 like this here. So, I think I can show you all these things in the R software also so that you can understand it very easily. Let me try to show you here this here x is equal to here c(11, NA, 12, NA).

So, now if you try to see here complete cases in x. We have complete cases 11 and 12 which are available at first and third position. So, in this outcome first and third position have true and second and fourth position have here false. So, that is indicated by this thing. Now in this data vector 11 and 12 are present they are available.

So, it is here. And now if you try to find out here mean of here x and if you try to find out here mean of y this will be 11 plus 12 divided by 2 which is 11.5. So, now I have given you these sufficient examples to explain it to you the concept of missing data that we how to handle the missing data in the R software. Well, I have taken here univariate case with only 3 or 4 values which I can illustrate. And whenever we are trying to think about the multivariate procedures, the multivariate procedures you will see that they are combined with the univariate procedures.

So, now whatever I have told you here for handling the missing data, this concept will continue to the multivariate procedures also. Whatsoever we are going to take there will always be an option that how are you going to handle the missing value in that context. Since context my advice is that if you want to know the best place is to look into the health of this of that function. Suppose if I want to know that how to handle the missing value in means, suppose I want to know how to handle the missing value in variance or quantiles etcetera, the best is to look into the help menu because in some cases it may work but in some cases it may work in a different way or it can be a combination of this. For example, in the next lecture I will try to show you that whenever we are trying to find out the length of a data vector and suppose the data vector has some missing values then na.rm equal to true will not work but there I have to work with na.omit.

So, these minor shuffling minor changes will be there in different concepts but the basic fundamental basic concept will remain the same. So, my advice to you all is that you please try to take now different examples, try to take simple data vector, try to combine them in matrix, data frame etcetera and try to give some missing value yourself and try to see whatever missing values you have given at a given place are they being diagnosed by the R software and the more important part is that you try to look at the outcome of the R software and try to look into backward what it is trying to indicate. For example, if the outcome is 2, 4 that means you have to understand that in your data vector the values are missing but what this 2 and 4 are indicating is it second place, fourth place or the values which are 2 and 4. So, this type of understanding if you develop with the R software that is surely going to help you whenever you are trying to do with the multivariate procedures. For multivariate procedure our thought process starts with the univariate and that is the reason I am trying to emphasize first on the univariate procedure and all these concepts are going to be used later on whenever we are trying to do the hardcore multivariate procedures like principal component, discriminant analysis, cluster analysis etcetera.

So, you try to practice it and I will see you in the next lecture where we are going to understand that how this missing values are being handled in the measures of central tendency. So, my request to you all is that please have a quick revision of the lecture where we had considered the measures of central tendency and then we will see how they can be modified when there is a missing data and how are we going to handle them, how are we going to modify those commands that will help you in better understanding. So, you revise and I will see you in the next lecture till then goodbye.