**Multivariate Procedures with R**

**Prof. Shalabh**

**Department of Mathematics and Statistics**

**IIT Kanpur**

**Week – 12**

**Lecture – 56**

## Canonical Variables Analysis in R

Hello friend, welcome to the course multivariate procedure with R. Now you can recall that in the last lecture we had talked about the canonical correlation and mainly we had discussed what is the statistical setup in which we try to find out this canonical correlation and canonical variables. Now in this lecture we are going to implement the same thing in the R software. So, in this lecture I am going to demonstrate here two commands for doing the canonical correlation analysis in R. In the first command I will try to show you and I will try to explain you each and everything and when I try to use another command cc then you will see that the same things are repeated in a different way. So, there I will not be explaining you each and everything but I will try to explain you the difference what is coming between the uses of two different commands and definitely in R software as this is open-source software so different people are trying to develop different types of packages different types of commands for doing the same thing.

So, that is why the different packages can do the same job and yeah, they may give different types of information in different formats. So, we simply have to understand these aspects but our main objective is that whatsoever is the outcome of the R software how it is related to our setup. For example, we had the concept of canonical correlation so what is here the canonical correlation what is here the canonical variable and how to construct the first second third etc. canonical pair of canonical variables. So, that is what we are going to discuss in this last lecture.

So, let us begin the journey of our last lecture and finish the course. So, now we are going to talk about the canonical variables analysis in the R software. So, now I believe that you all have understood what is canonical correlation what are the canonical variates.

So, now in order to compute the canonical correlation between two data matrices there is a command here cancor in the all in lowercase alphabets and because of the short form say canonical correlation cancor. So, then we have to give here x and y which are the numeric matrix of say n cross p order and n cross q order you can recall that we had two data vectors x1 and x2 in which we had assumed that first data vector has p variables and say another data vector has q variables on the second data vector.

So, whatever we have taken as say here x1 and here x2 this is indicated here by here x and here y respectively really in R. Then we have option here x center which is a logical or numerical vector of length and it describes whether centering is to be done on the x values be before the analysis or not and if also then it do not adjust the columns etc. So, and then similarly here for the y center also it is it also tells it is a logical or numeric vector of length and it describes whether the centering has to be done in the y values or not. So, I will say more details you can find out in the help menu where you will get more thing and some of the values which are coming here they are here with the cor we will get a list of the correlation coefficient then with xcoef you will getting the estimated coefficient for the x variable that is your here the eigen vector ak in some way in your terminology or say eigen vector of this a corresponding to a and similarly ycoef will give us the value of here say here b or say here bk and x center and y center will indicate that the values the values used to adjust the x and y variables. So, now in order to demonstrate it in the last lecture I have taken a summarize data and I had shown you manually how you can find out this canonical correlation and canonical variates but now here I am using here a built-in data set which is about the inter country life cycle saving.

This is the data that was collected during 1960 and 70 and it is available in the R software as LifeCycleSavings, life where L is upper case, cycle where C is upper case, and savings where S is upper case. So, this data has 50 observations on 5 variables like as here sr which is the numeric aggregate personal saving pop15 which is the numeric percentage of population under 15, pop75 which is the numeric percentage of population over 75 years, dpi is the numeric real per capita disposable income and ddpi is numeric percentage growth rate of dpi. Well, I am not interested in the definitions of the data but my main thing is this we have got here 5 variables and we want to create here canonical variables. So, and then these are here the details above of this data set and this was obtained from this book on Belseley, Kuh, and Welsch. So, anyway we come to our business.

So, you will see here this data looks like here this one. Here we have sr, pop15, pop75, dpi, ddpi for different countries, Australia, Belgium etc. And this is here the screenshot of the data set. Right, because I will try to show you on the our software but anyway our job

is that I want to create here a canonical variable. So, what I try to do here that I try to take here I try to divide this data into two parts.

So, in one set of vectors I try to take here pop15 and pop75. So, this will be nice something like here my x1 and then the remaining variables which are here sr and here dpi and ddpi they will be coming here as say here x2. So, this I can do with this one that I try to create here a vector say pop that is population which is the say this like as here and then I try to see here oec second data vector containing the sr, dpi and ddpi columns which is here say here by this command. And then I try to conduct my this canonical correlation analysis using the command can corr this pop oec. See here this is your here this pop data.

There is more data but I am trying to show you here that how it will look like there are only two variables pop15 and pop75 and then in the oec there are sr, dpi and ddpi which is obtained here like this. Now once you try to execute this can corr command then you get here this type of outcome. So, now we have to understand that what it is trying to tell you. So, you can see here that this is here the screenshot and before I try to move forward you let me try to show you this first on the R console and then I will try to take up the interpretation of different outcomes. So, firstly let me try to show you this dataset.

You can see here this is the dataset which is available in the R software anyway. So, this is built in so I am not worried. Now I try to create here this one pop oec and I try to conduct my this canonical correlation analysis. So, you can see here these are here different values and these values exactly are the same which I have shown you here if you try to see. So, now my objective is that I try to take here these aspects one by one.

So, suppose I take first for this thing then I go for this thing then I go for this thing and so on and then I will try to explain you what is happening here. But in a what shall I can tell you here very clearly this is the value of first canonical correlation, this is the value of here second canonical correlation. You can write down here first canonical correlation and these are the values of here for this a1 and a2 and these are here the values of here b1 and b2 and say here b3 and then these are the values here for this x center and y center which have been used for centering the data. So, let us try to understand these values one by one. So, now if you try to see here first I try to take here this aspect what is this dollar cor.

So, this is the value the first value the value of the first canonical correlation and second value is the value of the second canonical correlation, this is the first canonical correlation. So, now you can see here that first canonical correlation is 0.825 is much

greater than the second canonical correlation which is 0.365. So, obviously it indicates that we can consider only the first canonical correlation and the first canonical variation.

Now how to get it done? If you try to see here this was your here this part. So, this I have copied here and I am trying to now explain you how can you do it here. So, this is here like this so the so if you try to see here this is here the value of here a1 which is corresponding to see here the canonical correlation 0.824 that is the first canonical correlation and this here is the value of here a2 which is corresponding to here second canonical correlation 0.36. So, now if you try to see because we are going to consider here only the first but as a fundamental, I will try to show you that how you can construct both the canonical variates. So, you can see here this U1 and U2 this will be coming from here now if you try to see this will be a linear combination of pop15 and pop75 so the first canonical variable here is this value - 0.009 * pop15 + 0.049 * pop75 and the second canonical variate will be obtained from here second column where this - 0.036 which is here like this and then here -0.26 * pop75. So, this is how you can see that I have obtained the U1 and U2. Similarly, I can obtain here V1 and V2 but now because the third set of variable X2 has three observations so there will be three values but we are going to consider only first one because our canonical correlation is er in the which is the first canonical correlation. So, if you try to see here under this ycoef which is here so I can show you here this is the part.

So, if you try to see here so now this is giving you here the values of V1 this is the value of here V2 and this is here the value of here V3. So, now the first canonical variable V1 will be considered from here that 0.008 * sr + 0.001 * dpi + 0.0041 * ddpi and similarly this from the second canonical variable is obtained here like from this one. So, this 3.33 * 10 power of - 0.02 it is here 0.003 then this value here it is here and this - 1.22 * 10 power of - 0.02 which is here like this. And similarly, the values in the third column here they are used to construct this third canonical variable V1 = - 0.0005 * sr + * dpi + 0.005 ddpi. So, these values are here like this first value, second value and here third value here like this. So, basically but if you try to see that the minimum value between 2 and 3 is here 2 so ideally, we need to consider only the first two canonical correlation.

Well, the software is giving you everything and it is only your decision that how you want to do it. So, similarly we have here the values of here X center which is here for population 15 this is 35.08 and for population 75 which is 2.29. So, it gives the value used to adjust the X variables.

So, it describes if any centering has to be done to the X values or not. And similarly, this

Y center is also giving you here 3 values for SR, for dpi and for ddpi so they also give the values used to adjust the Y variable. So, anyway so this is pretty straight forward actually. Now I would like to use another package and would try to find this canonical correlation and canonical variables from the same data. So, for that there is another package here CCA.

So, we need to install it and then we have to upload it and then the usage here is say ccxy earlier it was say cancor now it is CCA. So, this X and Y they are the numeric matrix of order n by p and n by q containing the X and Y coordinates which are equivalent to your data on X1 and X2. So, now if you and then after that there are different types of things what happened earlier like cor will give you canonical correlation, names will give you a list of containing the names to be used for individual and variable for graphical output, xcoef which is the estimated coefficient for X something like here a1, ycoef will give you the value of here the coefficient for Y which is equivalent to here b and scores is a list of return by the internal function compute which contains the individual and variables coordinate on the canonical variance basis. So, let us try to implement this command on the same data set. So, I try to use here the same data sets here pop and oec which I but now my command here is CC with pop, oec.

Now I will try to show you it has many components which it is a long output. So, and it has the same thing but I will try to show you that what it is trying to explain right. But firstly, let me try to show you it on the R console so that you are confident that we whatever I am doing here it is correct right. So, first I need to I already have installed this package on my computer so I do not need to install it but I will upload it certainly and then I will try to use here the command here this CC on this pop and oec. You can see here this is a you can see here this is a pretty long outcome like in my case here right.

So, what I have done here I have taken the screenshot and I will briefly explain you what are they trying to tell you right. Otherwise, it is difficult for me to take even the different components and explain you here. So, if you try to see this is the beginning part. So, the first value which you are going here this is the value of first canonical correlation this is the value of here second canonical correlation and then these are the names of on the in the data vector X1 you can see here pop15, pop75 then it is here the names in the say data vector here X2 which was sr, dpi, ddpi and it is trying to give the names of the values in the first data vector pop and similarly it is trying to give here the coefficient corresponding to this pop15 and pop75. So, these are say here a1 and this is here a2 like is here like this and similarly here the Y coefficient this is here say here b1 and this is here b2 and then it is trying to give you the scores here for each of the observation it is

trying to compute the scores right you can see here the score for X, the score for Y and so on right.

So, that is how and then it is trying to give you some more information on the on the correlation structure and so and so on here. But the main part if you try to see here in which we are more interested at the moment that it is trying to give you here the similar outcome of this X cough and Y cough from where you can easily write down the U1, V1 and U2, V2. But here you can see the difference here is that it is giving you only here U1, U2 and V1, V2 it is not giving you here V3 at that was happening in the case of cancor. Beside this it is giving you here some more information so if you need it, you can use it otherwise yeah mean the mean the outcome and the interpretation of the values is the same what we use in the case of cancor and this is here the values of the first and second correlation coefficient in which we are interested right. So, you can see here it is not a very difficult thing to to understand it and we can do it very easily in the R software right.

So, now we come to an end to this lecture and it is not only end to this lecture but the end to this course also. Well in this lecture you have seen that we have implemented the canonical correlation analysis and it was very simple. The only thing is that you should know the command and the correct package. You will get the outcome but in order to understand the outcome you have to understand first what is happening behind the curtain and for that you have to understand the statistical knowledge which is being used to develop this type of tool. The execution part is not is not so difficult it is very simple but understanding and making a familiarity with what is happening inside the data based on this outcome is more important.

But it is not difficult also it just depends on how much you practice. So, now the course is going to end but that is true only for me but for you the course is going to begin from today. When we met, we had no idea what is multivariate. So, I used the univariate concepts to build up in stairs this multivariate setup and you can see that in the last couple of weeks you are very much comfortable with the multivariate setup and if you try to think that when in the first lecture if I had told you directly about this multivariate analysis and this vectors and matrices etc. possibly it would had been very difficult for you to understand.

But I opted a way out that I started from a univariate and then stepwise I moved to multivariate. Yes, it takes some time for me because it is possible that some of the concepts you already have done in the univariate but my constraint is that I do not know my students, I do not know my population, I do not know my participants and whatever I

am doing it today to which population it is going to cater I do not know. So, that is why I have to take care of all the people in my population who are my students or who are my candidates. Some of them might have a good knowledge of statistics then they are doing this course and some people might have very little knowledge of statistics and they are also doing this course. In order to bring them on the same platform in the beginning I have to repeat couple of things.

Then I had taken some selected topics to explain you assuming that you have a reasonable knowledge of statistics and now you can see that all of them were used either it is maximum likelihood estimation or normal distribution or the matrix command in R. I always used to say that when we will do the further lectures then we will try to understand it. But now you can see that there was not even a single thing which we have not used it. But that will give you an idea that if you want to begin in the course then what are the different things which you should know out of many many things.

So, that was my approach. I tried my best to explain you about these multivariate topics but definitely there are many more topics but we have a time constraint also. So, I am not doing it here but I can assure you that they are not difficult at all. And if you have learned these many things learning those topics yourself is not an easy job. For example, if I say MANOVA, MANOVA is simply multivariate analysis of variance. You have done the means analysis of variance in different cases in the univariate case.

Now you simply have to extend it to a multivariate case that is all. And concept will remain the same. The utility will remain the same. Now instead of having H0: mu1 = mu2 = mu p we will have mu1 mu2 mu p as vectors that is all. And the job is going to be done by the software but for that you have to read couple of pages in the book to understand what it is trying to do.

It is not a difficult thing. And then you see there are many more topics which are very important for this multivariate analysis. So, in multivariate procedures we have considered different types of aspects, different types of setups and my sole motive was to make you understand what is multivariate analysis and above all to take out the fear of learning multivariate analysis from your heart. If I can do it in case if I can remove the fear from your heart then you are fearless and nobody can stop you in learning any topic, any concept in this world. So, with this hope I stop in this lecture and I request you that you please try to understand it, try to learn it, do good practice, take care of yourself and may God bless all of you. So, see you sometime somewhere till then goodbye. Thank you.