Multivariate Procedures with R

Prof. Shalabh

Department of Mathematics and Statistics

IIT Kanpur

Week – 06

Lecture – 28

Maximum Likelihood Estimation

Hello friends, welcome to the course Multivariate Procedure with R. So, you can recall that in the last lecture I have given you an introduction about estimation of parameters. And we had discussed about point estimation and interval estimation. Now, there are various methods for finding out the point estimate as well as interval estimates for any parameter. And among them I would like to give you here some idea that how these parameters are estimated and what type of values do they get and how to compute them in a real life data situation. So, as I said in the past also couple of time, the objective of this course is not to teach you the complete statistical inference.

All the topics in the statistical inference like as point estimation, confidence interval estimation and another related topic. But my idea is that I want to prepare here a background and I want to give you a sufficient background so that you can understand the topics in the forthcoming lectures. So, keeping this in mind, in this lecture I am going to talk about maximum likelihood estimation technique. So, this maximum likelihood estimation technique gives us the values of the unknown parameter on the basis of given sample of data.

So, as the name of this topic suggests maximum likelihood. What is likelihood? Likelihood is essentially the probability or the chances of occurrence of an event. For example, if you try to see the black clouds in the sky, then you say that okay there is a very good likelihood that it may rain. And in case if you see a clear sky, then you always say that okay the chances of rains are very remote, they are very less or in some way I say that okay the likelihood that it will rain is very very low. So, if you try to understand this sentence what are you trying to do? You are trying to take some decision about some unknown process unknown data based on the value of the probability.

And you are trying to take the decision which is corresponding to the maximum probability of occurrence. So, now if you consider a sample, the sample is coming from a population whose population parameters are unknown to us. So, when the population parameters are unknown to us, I want to know what can be the value of the parameter such that the sample is coming or originating from that population has the maximum probability. So, basically we try to write down the joint probability of all the observation and we try to maximize it. And based on that the value of the parameter which is maximizing the joint probability is obtained and that is called as the maximum likelihood estimator.

So, now when we are talking about the maximization, there are different ways by which we can do the maximization. One of the popular method which you have studied is the principle of maxima and minima. Besides those things there are several method, there are several numerical procedure, there are several optimization method which can give you the value of the parameter for which the function is maximum. So, I will try to give you here some simple example in univariate and I will try to take it to the multivariate scut up also. I will try to consider here those example which have a closed form which are giving you a very clear cut solution.

But in many situation the value of the estimate based on the maximizing the likelihood function will be only based on the numerical data and the obtained numerical procedure that I am not covering here. Because my objective is to make you familiar with the maximum likelihood estimation technique that we are going to use in some forthcoming lecture. So, with this objective let us begin our lecture. Now, in this lecture we are going to talk about the maximum likelihood estimation of the parameters, right. So, there are different methods in statistical inference to obtain the values of the parameter on the basis of given sample of data, method of moments, method of maximum likelihood, method of minimum price square etcetera.

And one among them the method of the method of maximum likelihood estimation is a popular method. And as the name implies method of maximum likelihood, maximum likelihood means the likelihood is the maximum. So, the estimator will be the value of the parameter that maximizes the likelihood function. So, for example, let us say that let x1, x2,… xn be a random sample from a probability density function or equivalently a probability mass function say f(x) theta which is here like given like this. Because now we are trying to look at f as a function of x1, x2,… xn for a given value of theta which for a fixed theta, the theta belongs to this parametric space capital theta, right.

Then what we try to do here? We try to find out the joint probability density function or probability mass function of x1, x2,… xn on the basis of given sample. So, we try to find out here this f of x1, x2,… xn given theta. Now, because we are assuming that they are independent because they are the random sample. So, I can write down this joint pdf as the product of marginal pdf or equivalently the probability mass functions. And this joint distribution of x1, x2,… xn is indicated by here function L.

L is a function of theta and x1, x2, … xn, right. So, L is called as the likelihood function and this is a function of the parameter theta given the observed and known sample values x 1, x 2, x n. Similarly, if you for example, if you are trying to consider the discrete random variable where you are trying to consider the probability mass function. So, then instead of here joint probability density function we can also find out the joint probability mass function like here is like here this probability that X1 is equal to x1, X2 is equal to x2, up to Xn is equal to xn, right. And then based on this likelihood function we try to find out the value of the parameter theta which is likely to maximize the probability in the sample, right.

And so the maximum likelihood estimator or we shortly call as the MLE, M for maximum, L for likelihood and E for estimator. The MLE of the parameter theta is the value of theta that maximizes the likelihood function L which is the function of theta and x1, x2, … xn. So, now let me try to take here a simple example to explain you that what do we try to do. For example, I will try to take here two example one for discrete case and one for say continuous case. So, in the case of discrete random variable we consider here the Bernoulli distribution, right.

That the sample X1, X2, … Xn is coming from a Bernoulli distribution which is indicated by here binomial(1,p). And if you have got the only one parameter p and the probability mass function of Bernoulli distribution is given by probability X equal to x which takes value the probability p if X is equal to n and 1 minus p if x is equal to 0, right. So, all this x1, x2,… xn they are identically and independently distributed that is iid. So, the likelihood function of the parameter theta which is here p and x1, x2, … xn is found here like this L is equal to probability of joint probability function of X1 equal to x1, X2 equal to x2 and Xn equal to xn. So, you can see here this is only the product of probability of Xi equal to xi because X1, X2, … Xn are independent.

So, this is obtained here like this p is power of summation xi into 1 minus p is power of and minus summation xi. Now, this is the likelihood function and we want here the value of p for which this whole function is maxima. So, the maximum likelihood estimator of p is the value of p that maximizes the likelihood function L in this case. So, now the main issue here is that how to maximize it. There can be different ways to maximize it, right, but we are going to consider here one very commonly used method which is the principle of maxima and minima.

So, in order to maximize this likelihood we consider its log or we or more especially the natural log which is given by the function L n. So, we try to minimize the natural log function of L using the principle of maxima minima. So, the log of L is obtained here by here with this thing and its log comes out to be here like this. That is a very simple calculation. Now, in the principle of maxima minima we try to find out its first derivative and put it equal to 0.

So, we try to find out the first derivative of log L with respect to p which comes out to be here like this and we equate it to 0. This is called as the likelihood equation or the normal equation, right. And in case if you try to substitute it here this will solve to p is equal to x bar. So, I can write down that the maximum likelihood estimator of p which is indicated by here p hat MLE is equal to x bar. And if you want to see whether this value of p is maximizing or minimizing the likelihood function, so we can verify it with the second order condition.

Well, I am not going to give you the details that how principle of maxima minima work, but I am assuming that you all are aware and you know it, right. So, thus I can say here that if x bar is the maximum likelihood estimator of p. So, now you can see here that in this function, this p was unknown to us, p was unknown. And what we have done? We have got here a sample x1, x2, … xn and based on that we have found that if I try to take p is equal to here the capital X bar that means the estimator of p which is indicated by p hat is equal to capital X bar. So, that in case if you get here a sample x1, x2 here xn then you simply have to find out here the sample mean x bar like this 1 upon n summation n, xi and it will give you the good value of the parameter p, right.

Now, after this let me try to take here one more example from the univariate normal distribution. So, what you try to do? That first you try to understand this example and I will try to extend it to a multivariate case that is for the case of multivariate normal distribution, right. So, suppose X1, X2, … Xn is a random sample from normal distribution with mu, mu and variance sigma square whose pdf is here being like this, right. Where x lies between minus infinity and infinity mu lies between minus infinity and infinity and sigma square is greater than 0 and this X1, X2, … Xn they are iid. So, the joint probability function which is here the likelihood function can be written as the product of individual pdf of xi which will come out to be here like this.

It is simply I am just trying to multiply this effect. Now, it is easier for us to take the log of this function means log of this L and then we try to maximize it as a respect to mu and sigma square. Well, we are assuming here that mu and sigma square both are unknown to us, right. That is more interesting case otherwise if sigma square is unknown then we need to differentiate the likelihood function with respect to only here mu and if mu is known sigma square is unknown then you need to differentiate the likelihood function with respect to only sigma square, right. And if both mu and sigma square are unknown then we need to differentiate the likelihood function with respect to mu as well as sigma square.

So, we try to differentiate log of L and equate it equal to 0 and once you just differentiate it put equal to 0 solve it and then you will get here the value of mu as the x bar and that is going to be your the maximum likelihood estimator of mu and similarly you will get the value of sigma square as 1 upon n summation i goes from 1 to n x i minus x bar whole square which is the maximum that you estimated of sigma square. Now, the question here is that whether these two values mu equal to mu hat mle and sigma square is equal to sigma square hat mle whether are they going to maximize or minimize the likelihood function that is what we have to check. So, what we try to do here that we try to consider here the Baudel-Hessian matrix here like this and we find that it is a negative definite matrix at mu equal to mu hat mle and sigma square is equal to sigma square hat MLE right. So, now, these are the maximum likelihood symmetries. So, now, you can see here once again I would try to explain you here that we have got here this probability density function where mu and sigma square unknown we try to take here the sample of size x1, x2 … xn and from there we try to find out here the value of mu as x bar and the value of sigma square as 1 upon n summation i goes from 1 to n, xi minus x bar whole square.

So, you can see here that both this estimated they are function of sample values only. So, they are actually statistic right. So, now, if you try to see if you want to compute the mle in the R software. So, you can compute it by here mean function and here variance function VAR, but now you can see here the divisor here is 1 upon n whereas, variance in R gives you here the values with the divisor n minus 1 like this one. So, we can make a small transformation and by writing n minus 1 into VAR divided by n we will get here the maximum likelihood is symmetric of sigma square right.

So, this is how you can do it. Now, our basic objective in the course is to consider a multivariate setup. So, now, I have given you the idea that how are we going to implement the maximum likelihood estimator in a univariate case in a univariate normal population and just to make you understand better I will take here the example of multivariate normal. And I will try to show you that how the same concept of maximum likelihood estimation is extended to a multivariate case right. So, suppose if I say here like this X1, X2, Xn is a random sample from a normal population right, which has got here mean vector here mu and covariance matrix here is sigma which is a positive definite matrix. So, we have discussed earlier that the probability density function of multivariate normal distribution is given here by this thing right which I am writing here.

Now, if you try to see, now we are trying to say that we need to find out the joint probability density function of this X1, X2, Xn. So, since all X1, X2, Xn vector they have been obtained independently. So, we try to find out here the product of this exercise and if you try to see here this will come here like this. Now, if we try to take here the natural log of this here likelihood function like as here log of n and if you try to just write down this log of L function and now I can have here three possible cases right, one of the parameters is known or both parameters are known. So, if I assume that suppose that mu is unknown and sigma is known then in that case whatever you are trying to do you just try to differentiate the log of L with respect to here the mean vector mu and put it equal to 0 and solve it and you will get here that mu hat is equal to X bar vector.

So, X bar vector means here because you are saying that you have here a vector X1, X2, Xp. So, now you are trying to take here a sample like as X1 equal to this and say here Xn equal to here say here like X1, X2, Xp. So, you try to take here all X1 and then find out their sample mean this will be indicated here as X bar 1. Similarly, if you try to take all the observations on the pth head variable and you try to find out here X bar which is indicated by here X bar p. So, this is indicated here by this mean vector and after that you can also go for the second order derivative to make sure that it is the value of the mu which is maximizing the log of likelihood function.

The second case is that if a mean vector is known and we do not know the covariance matrix. So, in that case you simply have to somehow find out the maximum value of the sigma which is going to maximize here the likelihood function right. Here I would try to say that you have to be little bit careful, washful that when we are trying to differentiate the likelihood function with respect to a vector or a matrices then we have to follow some rules and regulations of the matrix theory. And I am not giving you here the details of those things, but the maximum likelihood estimator of sigma will come out to be here like this 1 upon n summation i goes from 1 to n, x alpha minus mu and transpose x alpha minus mu right. So, you can see here because now here is mu is known.

So, that is why this mu is coming here right. So, now, in this case let me assume that both mu and sigma they are unknown to us. So, we try to write down here the likelihood function and we try to maximize it with respect to a mean vector mu and covariance matrix sigma. So, obviously, the type of algebra the type of results what are needed to maximize the log of L they are different than the univariate case and yeah I am not giving you here those intermediate step, but after solving them we can obtain here the maximum likelihood estimator of mu as a sample mean vector. So, in such a way that x bar this vector is has p components.

The first component here is x1 bar that is the mean of the first n observation on the variable x1. Similarly, x bar 2 is the mean of the n observation on the n variable x2 and similarly here x bar p is the mean of n observation on the last variable xp right. And similarly the maximum likelihood estimator of covariance matrix sigma is obtained here like this 1 upon n summation alpha goes from 1 to n, x alpha minus sample mean vector x bar transpose into x bar alpha mean vector minus sample mean vector right. And now in case if you want to compute these values in the R software for a given set of data. So, in order to compute the sample mean vector the R command is col means.

So, if you try to compile your data in the data vector x, then you have to write if you col in lower case m in upper case and then means in the lower case this is essentially the column means right. And in order to compute the covariance matrix then this thing you have to simply use here the command VAR, but as you know that this VAR that is the variance command computes the covariance matrix or the variance with divisor n minus 1. So, if you want to have it by this divisor n for example, here like this 1 upon n then you have to just modify the VAR command by here this n minus 1 into VAR of x divided by n right. So, this is how you can compute the maximum likelihood estimator of the mean vector and covariance matrix right. Just to give you some idea here that how do we handle the multivariate normal distribution.

So, trying to generate here first the observation from a multivariate normal distribution right. So, in order to do it we require here a library mvtnorm() right. This is short form of multivariate normal mvtnorm() all in lower case alphabet and for that the command to generate the random number, but here the numbers are going to be actually a vector of a random numbers. The command here is rmvnorm() right and then you have to specify here n is equal to 5. That means, you want here 5 data vector and you need to specify the mean vector by here command here mean is equal to C 10, 20, 30.

For example, it is something like here mu and you need to specify here sigma by writing s i g m a in lower case alphabets and suppose for the sake of simplicity I am trying to take it here as a diagonal matrix of 2, 3, 4 like as here 2, 3 and 4 and all other elements are in the off diagonal they are 0. So, if you try to generate here suppose this is theta x then you can see here it really here comes in like this. So, this first row is going to indicate the first data vector, the second row is going to indicate the second data vector, the third row is going to indicate the third data vector and so on. So, now, in case if you try to operate here the command here col means, colMeans() in lower case alphabet x then it will here come out to be here like this right. So, what it is trying to do here if I try to show you here that it is trying to find out the mean of the observation in the first column and it is reproducing here in at the first value.

Similarly, it is trying to compute the mean of the observation in the second column and it is computing it in the second place. And thirdly, if you take the mean of the observation in the third column that is reproduced at the third place. So, this is how it tries to compute the sample mean vector on the basis of data given in x right. So, you can see that it is not difficult. Similarly, if you try to compute here the covariance matrix.

So, you can see here that we have got here 3 variables. So, this is your here the variance of first variable sigma 11, this is second diagonal element is the sigma 22 which is the variance of third variable. Similarly the third diagonal element this is the variance of third variable x 3 right. Similarly, if you try to come here this variable in writing this is your here sigma 1 2 which is here the covariance between x 1 and x 2 right which is something like 1 over n minus 1 summation x i minus x bar 1 i minus x 1 bar and x 2 i minus x bar 2 like this. And similarly, if you try to see this value in the second row first column this is they are the same because covariance between x and y is the same as covariance between y and x.

Similarly, this value here this is here the sigma 13 which is the covariance between say here x1 and say x3 and this is here the same value here if you try to see here as here like this. And similarly, if you try to see here this value here is covariance between here x2 and x3 right and this is the same value here they are going to be the same. So, this is your actually here the covariance matrix right. So, similarly if you try to find out the covariance matrix which has the divisor n then you have to operate it here by this here n minus 1 into variance of x divided by n and here is the number of columns that I can find out here n col command which is equal to here 3 and you get here is the same outcome, but it the divisor n in place of n minus 1 right. And similarly, if you want to find out the correlation matrix right.

So, correlation matrix can also be found just by the command here see here c o r. If you remember c o r was also used in the case of univariate data to find out the correlation coefficient, but now when the data structure is in the format of a matrix then it is giving you the correlation coefficient. Then similar interpretation can be given to is for example, this the elements on the first diagonal element is the correlation coefficient between x 1 and x1 which is here 1. Similarly, the second diagonal element is the correlation coefficient between x2 and x2 and the third element on the diagonal element is the correlation between x3 and x3 and that is why they are 1 right. Then this element the element on the first row and second column this is the correlation coefficient between x1 and here x2 and the element in the second row and first column this is also the correlation coefficient between x1 and x2 which is the same value like here like this.

Similarly, if you try to take here this value here 0.3300610 this is the correlation coefficient between x1 and x3 and this is here the same as here like this which is the correlation coefficient between x3 and x1. And the remaining value here this one and this one they are the correlation coefficient between x2 and x3 and this is here correlation coefficient between x 3 and x 2 because correlation coefficient between x and y and y and x they are the same. So, that is why the off diagonal elements are same actually. So, now you can see here it is not difficult and here if you try to see this is the screenshot here, but let me try to show you this on the R console so that you get here more confidence. But definitely my this outcome is not going to match with you because I am I have not used here any set sheet.

So, if you try to see I am uploading first the package mvtnorm() and yeah means you have to install it because I already did it. So, if on my computer earlier. So, now if I try to get here the value of x you can see here this is here like this and if you want I can change it here suppose if I take it here suppose I want here say 6 observations. So, you can see here now I have here 6 observation 1, 2, 3, 4, 5 here and in the second case it is going to be 6 right and if I try to find out here the column means here like this. So, you can see here this comes out to be here like this and similarly if I try to find out here the variance of here x you can see here it will it will come like this and similarly if you want find the correlation of x you can here find out here like this right.

So, now, you can see here that it is not difficult and with this we come to an end to this lecture. So, you can see here that we started with univariate case one discrete one continuous and then we shifted to the multivariate case. My objective was to demonstrate you that how are we going to find such estimators. Well, when we are going to deal with more topics you will see that in most of the cases in real life data these parameters are unknown to us and unless and until we know the value of the parameters we cannot move ahead and we cannot draw the statistical inference. So, at many places I will simply say ok let us try to estimate these parameters by the method of maximum likelihood or some other method and whatever is the value or whatever is the value of those estimated parameters that will be replaced back.

So, that is why this was very important for me to explain you that in case if I say somewhere that ok we are going to replace this mu and sigma by their respective maximum likelihood estimators then you must understand what I am trying to say. And now I also have demonstrated that these methods these estimation methods work for univariate case as well as for multivariate case. The same thing will happen in the case of confidence interval estimation. So, later on I would try to take some techniques which I explained you in the last lecture that we try to create a statistic which whose distribution is not dependent on the parameter and then using that statistic we try to create the construct the confidence interval, but that I will try to take in the further lectures. And then in the case of multivariate also the concept of confidence interval can be extended.

So, obviously the confidence interval is in the form of some unidimensional space because we are considering only univariate case when we have a two variables then this confidence interval become a confidence region and when we have more than two variables then it will become a sort of some structure some ellipsoid confidence ellipsoid that is this is how we call it right. So, all these things will come in the lectures in future. So, but at the moment my basic objective is that I wanted to tell you that method of maximum likelihood in these cases we could obtain the values of parameter very easily by using the principle of maxima minima, but many times we had to observe the likelihood function and then we have to find out the value of parameter such that the likelihood function is maximized. And in many cases the function is so complicated that we cannot use such a clear cut methods and in those cases we try to optimize them using some algorithm or some numerical technique, but in all the cases the objective is the same that we are trying to maximize the likelihood function. So, please understand that we are simply trying to maximize the likelihood function using any of the approach whatever is convenient and possible to use.

And method of maximum likelihood is for univariate, bivariate as well as multivariate also. As we are going into more direction the number of parameters are going to be increased. And so, that is why we will be considering the estimation of the vectors and matrices also. So, you please try to look into book try to understand more about this method of maximum likelihood and other possible methods at different places I have given you the direct expression you please try to solve them algebraically and try to see that you are not facing any problem this will give you more confidence. So, you try to practice it and I will see you in the next lecture with more topics till then goodbye.