Multivariate Procedures with R

Prof. Shalabh

Department of Mathematics and Statistics

IIT Kanpur

Week – 08

Lecture – 37

Multiple Linear Regression Analysis: Introduction and Basic Concepts

Hello friends, welcome to the course Multivariate Procedure with R. From this lecture we are going to begin with a new topic which is about Multiple Linear Regression Modelling. So, the topic of multiple linear regression modelling is not a small one. In fact, usually we have a one semester course on the regression analysis. And the objective of the multiple linear regression analysis is to find out the statistical model. Yeah, I agree modelling is a very fancy word everybody wants to find out model on the basis of given set of data.

So, there are in fact different ways to find out the statistical model in statistics and among them this multiple linear regression modelling is a very popular technique which helps us in finding out the relationship between an output variable and a group of independent variable which are going to affect the outcome. For example, the yield of a crop, the yield of a crop depends on several factor like as quantity of fertilizer, temperature, rainfall, humidity etcetera. But our interest is only in the quantity of yield. So, this quantity of yield is now depending on more than one factor.

So, we would like to find out a relationship or statistical relationship between the yield and these variables. So, in regression analysis what we try to do that we try to observe the data by conducting an experiment, we try to take different values of the variables that are going to affect the outcome and then we try to come in the backward direction that using the available data we try to find out the relationship that is existing between the input and output variables. So, that is why it is called as regression analysis. The dictionary meaning or the literal meaning of regression is to move in the backward direction. So, my objective in this lecture is to give you an overview of the regression analysis.

Well, I will not be covering all the topics of multiple linear regression analysis in this course, but my objective is that because it is a multivariate procedure and it is a very important procedure for the regression modeling or for obtaining a statistical model. So, that is why I am taking it here, but definitely I will try to give you a fair idea, but surely I would request and I expect from you also that you will try to look into the books and you will try to understand this topic in more detail. And my advice is that there are many topics when you are trying to deal with regression analysis and they all are really useful in the practice because in practice the data is not going to follow my rules. The data will be coming in its own way following the real or the natural process. So, many complications will be happening.

For example, we will have the problems which you possibly are aware by the names multi collinearity, heteroscedasticity, autocorrelation, dummy variable etcetera etcetera. But we are not going to discuss all of them here. We are going to take here some representative topics of this multiple linear regression analysis so that I can make you confident so that you can learn yourself also. But definitely in this lecture I will try to give you an overview that what are we going to do and what are the different steps which are involved in a complete multiple linear regression analysis. So, let us begin our lecture and try to have an overview of this multiple linear regression analysis and its different aspects.

So, now in this lecture we are going to just talk about the introductory topics and basic concept in the context of multiple linear regression analysis. So, we know that linear models play a very important role in the modern statistical methods. These models are able to approximate a large amount of metric data structure in their entire range of definition or at least piecewise. Sometimes the data is too huge, there are many complications then we try to divide them into segments and then we try to solve the problem. So, let me try to take a very simple example to explain you the concept behind this multiple linear regression analysis.

We know that the marks of the student depend upon several factors. For example, number of hours of study in a week, number of assignment submitted in a month or number of hours of play in a week etcetera. They can be many more, but I am going to consider here only three. And suppose we want to find out the relationship between the marks of a student and these three variable number of hours per week of study, number of assignments submitted in a month and number of hours of play in a week. In terms of a mathematical equation or a statistical model.

So, suppose we try to conduct an experiment and we try to collect the data from say 20 students on these three input variables and one output variable. Why 20? Because I do not have here more space, so and my objective is to give you an idea so that I can show you the data so that you can understand what is really happening. Once you understand with this smaller set of data and a smaller number of variables, then it is not difficult for you extend it to any big data set or with the large number of variables. So, I write here, I dictate here small y the number the marks of the students obtained in the examination and x1 is the number of hours in a week of study, x2 is the number of assignments submitted in a month and x3 is the number of hours of play in a week. So, now you can see here I have obtained here the data.

So, this is student number 1, 2 up to here 20. Now student, so data is obtained here like this. Student number 1 has studied 34 hours in a week, student number 1 has submitted 3 assignments in a month and student number 1 has played for 15 hours in a week and the student number 1 has got 180 marks out of 200 or say 250. Now this is my first set of observation. Now I try to repeat the same thing for the second student.

I try to ask the student that how many hours of study, number of assignment and the number of hours of play and I record this data and this thing I try to repeat for all the 20 observation and we obtain this type of data set. Well, I am going to use this data set in the further lecture also. So, you please try to have a look and try to understand the structure of the data set also. So, now the first question comes how to find out the functional form, how to find out this here f. For example, there is a going to be relationship between y and x1, x2, x3.

So, there is some function here f, something like y is equal to f of x1, x2, x3 where your x1, x2, x3 are given as number of hours of study in a week, number of assignments submitted in a month and number of hours of play in a week and y is the marks of a student. So, form of f is actually not known to us in practice. We do not know, we can simply observe the data and this form can be linear as well as non-linear. So, this relationship f is known to us only when the form of the function f is known to us. So, how to handle this situation? So, now in this setup you have seen that we have two types of variables.

One are input variables and say another are output variable. This input variable they are also termed as independent variables and the output variables they are also called as dependent variables. So, independent variables here are x1, x2, x3 and dependent variable

here is y. So, y can be affected by x1, x2, x3, but x1, x2, x3 cannot be affected by y. So, that is why or that is how this definition of independent and dependent variable has come.

And what is our objective? Our objective is to find out a relationship between dependent and independent variables which describes the phenomena or the process in the best possible way. And this is called as a model. It is a very popular name. But before you try to obtain any model there are certain condition which people actually do not understand usually, but I would like to emphasize that whenever you are trying to do the statistical modeling your role as a statistician has several challenges and constraints. You have no right to change or alter the process.

The way it is happening it will happen in the same way. The statistician has to work only on the basis of a small sample which is a small fraction of the population. It has to be a representative sample and so the sample is expected to have all features of the population. These are the constraints. You cannot ask anyone that you go back and try to conduct the experiment again.

Now let me try to quickly give you the idea about the linear models. Suppose the outcome of any process is indicated by a random variable which is say here Y. We call it as a dependent variable or another popular name is a study variable also. And this dependent variable depends upon say here k independent variables and this there are different terminologies. So, this independent variables are also called as explanatory variables and they are denoted by X1, X2, … Xk.

And now we assume that there is going to be some functional relationship between Y and X1, X2, … Xk and this can be explained by a parametric model which has some parameters say this beta1, beta2, … beta k. And now it is a real data process. So, there will be some uncertainty in the data set and this uncertainty is indicated by adding here a term epsilon. It is something like this if you try to see the room temperature. We always say the room temperature is supposed 25 degree Celsius.

But do you think that at every place in the room the temperature is going to be the same? The temperature near the bulbs will be more than the temperature in the corner of a room. So, there is going to be small variation, but if I try to take the all the actual data it will be very difficult for me to handle the analysis. So, we try to assume there is the room temperature is suppose 23 degree Celsius plus minus something. So, this random error which is due to the factors which we cannot control and that is indicated by here this term

epsilon. So, now we assume here that here in this one this f is some here well defined function.

This beta1, beta2, beta k they are the parameters and they are the parameter because they characterize the role and contribution of X1, X2, … Xk respectively. So, beta1 explains the role of X1, beta2 explain the role of X2 and so on. The term epsilon is the random error and it reflect the stochastic nature of the relationship between Y and X1, X2, … Xk. And this also indicates that the relationship between Y and this X1, X2, … Xk cannot be exact in nature, but we are using here the sign equality sign here. So, this epsilon is trying to balance it to maintain this equality sign mathematically.

$$y = f(X_1, X_2, \ldots, X_k, \beta_1, \beta_2, \ldots, \beta_k) + \varepsilon$$

So, obviously we have heard about two models, one is mathematical model and another is statistical model. So, what is the difference between the two? So, in mathematical model there is no randomness. So, when I say that epsilon is equal to 0, then the same relationship is called as a mathematical model and in statistical model we try to take care of the random variation in the data. So, when epsilon is not equal to 0, then the relationship is called as statistical model. So, the term model is actually broadly used to represent any phenomena in a mathematical framework.

It does not mean that when I am trying to obtain a statistical model, I am not going to use the mathematics, I am going to use the tools of pure mathematics to obtain the statistical model, but due to the random nature we are calling it as a statistical model. So, now what are the different steps in when you want to conduct any regression analysis? I am just trying to give you the steps here, but I will not be giving you here all the detailed procedures how to get it done. But broadly I will try to explain you in a which is sufficient for us to understand. So, the first step is you have to give the statement of the problem under consideration, what exactly you want to do right, what you really want to find. Then based on that you have to make a choice of the relevant variables that if you try to look at the nature of the problem, then you try to see that the outcome is going to be affected by which of the variables.

There will be large number of variable, some variable may contribute more, some variable may contribute less, if you try to increase the number of variable that will increase the complexity in the mathematical computation. So, you have to strike a balance by choosing those variables which are contributing more to the outcome. And then we try to collect the data on those relevant variables and we have statistical

technique to choose that which of the variables are relevant. Then after that we have to specify the model. Once you have specified the model, then now you have then you also have collected the data on the input and output variables.

So, now you have to choose a proper statistical method by which you can fit the data to the model and find out the equation. So, then we try to go for the fitting of the model. There are various techniques by which you can do it, but then you have to choose the appropriate technique depending on the nature of the data and the prevailing condition in the experiment. Once you have obtained the data, then you have to actually check whether the data is telling you the same thing which is happening inside the experiment or not. So, for that we have the model validation technique, then we try to criticize our own data, that means we try to find out what are the issues which are remaining and then we try to redo the modeling and we try to update our this model.

And then whatsoever model we have chosen based on that we try to obtain the fitted model and the chosen model is used in solving different types of problems and particularly in the forecasting. So, now there are two option that when we are trying to find out here or consider a model then when the dependent variable is affected only by one independent variable that we try to consider only one independent variable or one explanatory variable, then the model is generally called as simple linear regression model right. And when we are trying to consider the regression modeling between the dependent variable and having more than one independent variables, then the linear model is called as multiple linear regression model. So, we are going to talk here about the multi linear regression model which is which has a multivariate nature right. So, when we are trying to consider the multiple linear regression model, then this model generalizes the simple linear regression model in two ways.

It allows the mean function that is the average value of y to depend on more than one explanatory variables and to have the shapes other than the straight line although it does not allow for any arbitrary shape right. So, you will see later on I will try to show you that how are you going to define the linear model right. So now, we will try to decide over this multiple linear regression model. So, let small y just to denote the dependent or the steady variable that is linearly related. Now I am posing a condition here linearly related to k independent variables x1, x2, … xk through the parameter beta1, beta2 to beta k.

So, basically I have specified here the function f which is now here like this y is equal to x 1 beta1 plus x 2 beta2 plus up to here xk beta k plus epsilon and this is called multiple

linear regression model. So, this linear word is coming because now we are assuming a linear relationship between y and x1, x2, … xk right. Here the parameters beta1, beta2, beta k are called as regression coefficients and they are associated with the explanatory variables x1, x2, … xk respectively. So, beta1 is associated with x1, beta2 is associated with x2 and so on and epsilon is the random error component which is reflecting the difference between the observed and fitted linear relationship. That means, whatever you are observing and corresponding to which you are trying to obtain the same value of y based on the inputted data set input variables and then you try to see why this difference is coming on.

So, epsilon is taking care of all such issues which we cannot handle ourselves right. For example, there will be many variables which we understand, but we cannot capture the data on them. There will be some variable which are contributing very actually a small amount in explaining the variation y. So, there can be various reason for such a differences which are captured in epsilon. So, we try to see here that the joint effect of all those variable which are not included in the model random factors which cannot be accounted in the model and so on.

There is a long list because of which one can justify the use of this epsilon in the model. For your understanding it is important to learn that the jth regression coefficient beta j, it represent the average change in the y when there is a unit change in the value of jth independent variable xj right. So, in case if I assume that expected value epsilon is equal to 0 that means the average of all those random errors is 0 that will be an assumption which I will explain you later on, but it means that some observation will have error which is positive, some observation will have errors in negative. So, in totality I assume I expect that this error is going to be 0 and so this beta j is the partial derivative of expected value of y with respect to xj right. So, now what is the linear model? You have to be very clear here by looking at the picture you cannot always say find out or decide that the model is linear right.

A model is said to be linear when it is linear in parameters right. So, if you want to judge whether a model is linear or not then what you can do you can find out the partial derivative of the expected value of y with respect to beta j and if this quantity is not depending on beta j then you can say that the model is linear. For example, if you try to take here this model say expected value of y is equal to beta1 plus beta2 x it is a linear model because it is linear in parameters beta1 and beta2 mean the power of this beta1 and beta2 here is 1. So, if you try to partially differentiate expected value of y with respect to here x this is going to be not x, but beta beta here 2 this is it is here simply here x which is independent of any beta1 beta2 right. Similarly, if you try to take here say here another

model that expected value of y is equal to beta1 into x is power of here beta2 this is non-linear in parameters right.

But if you try to make here certain transformations then it is possible to convert it into a linear model. For example, if you try to take here log on both the sides then it becomes a log of expected value of y is equal to log of beta1 plus beta2 log x and this can be written here as a log of expected value of y is written here as a y star log of beta1 is indicated by here beta1 star and log of x is now indicated by here x star. But now you have to be very careful when you try to interpret this model when you try to understand this model. The model this one is non-linear in parameters beta1 and beta2. So, it is non-linear, but when you are trying to consider this model y star is equal to beta1 star plus beta2 into x star then it is a linear model because it is linear in parameters beta1 star and beta2 right.

So, but now if you try to see you are trying to change your criteria. Now you are trying to judge the linearity of the model with respect to the parameters beta1 star and beta2 with variables log of y and log of x not with equal to x, y, beta1 and beta2. So, once you try to transform to this model after that you can apply the rules of multiple statistical linear regression model, but then you have to be very careful with the interpretation right. Similarly, if I try to take here another example here this is important for you to understand that expected value of y is equal to beta1 plus beta2x plus beta3x square right. So, you try to see you try to obtain the partial derivatives of this expected value of y.

As I said and you will find that it is linear in parameter, but it is non-linear in variable x, but it this is not my concern because our definition is saying that the model is linear when it is linear in parameters. So, hence this is also a linear model. On the other hand this model expected value of y is equal to beta1 plus beta2 upon x minus beta3 is non-linear in parameters y. Now, you can take the first derivative of expected value of y that is the partial derivative with respect to beta1, beta2, beta3 and can check yourself right.

So, this is also a non-linear model. So, now with this one I come to an end to this lecture. Well, this was a very say short, crisp and informative lecture where I have not done any mathematical analysis or I have not used the R software, but my basic idea was to give you an overview of the topics which I am going to cover in the multiple linear regression analysis and for that it is very important for you to understand how are we going to define the multiple linear regression models. Many time people try to look only into the scatter diagram and if the relationship is linear they try to consider it only as a linear model and if the relationship is somehow on the scatter diagram has a non-linear trend they assume that a multiple linear regression model cannot be fitted or a linear regression model cannot be fitted which may or may not be true. Because the linearity of

the model is defined with respect to the parameters not with respect to the variables and when we are trying to find out the scatter diagram where I have x on the x axis that means x values and y values on the y axis then this is a curve between y and x. So, the relationship between y and x may be non-linear, but still that may be a linear regression model and a linear regression model can be fitted.

So, that is why I wanted to give you all this intricacies that we are going to handle when we try to deal with the real data in real life. So, my request is that as I said that I am not going to cover all the things, but my of this regression analysis. So, it is important that you pick up a book and try to read the chapters corresponding chapters yourself from the book and as I am going further you please try to follow a book. Yes, I will be stopping at one place, but your journey should not stop with my stopping in the lecture, but that should continue further. So, for the next lecture I would request you that you try to pick up a book and try to read at least the simple linear regression model yourself because whatever are the results in the simple linear regression model they are extended to a multiple linear regression model also.

And in case if you want to understand it then I also have a course on this NPTEL Essentials of Data Science where I have explained about the regression analysis. So, you can also go through with those lectures also, but it is up to you whether you want to follow a book or go through with the lectures, but anyway without books you cannot become a good data scientist because they will give you the information in more detail and I have here a limited time. So, you try to practice it and I will see you in the next lecture with more details on the multiple linear regression model, till then goodbye. Thank you.