Multivariate Procedures with R

Prof. Shalabh

Department of Mathematics and Statistics

IIT Kanpur

Week – 09

Lecture – 41

Analysis of Variance and Implementation in R Software

Hello friends, welcome to the course multivariate procedure with R. So, you can recall that in the last lecture we had considered that test of hypothesis on the individual regression coefficients in the setup of multiple linear regression model. So, if you try to recall earlier we had conducted the test of hypothesis in the case of normal distribution when we had one sample data or one sample test then we consider the two sample test and after that we had considered the analysis of variance. So, the analysis of variance was a tool when we wanted to test the equality of more than two means from the normal population. Exactly on the same concept we have that test of hypothesis in the case of multiple linear regression also. So, we have conducted that test of hypothesis for the one sample test, H0: beta j = beta j0 which we have considered as 0.

Similarly, if you want to compare two beta j's also that also you can do. But anyway I have not considered here because beta j = 0 is more popular because we want choose the means relevant independent variable which are going to affect the outcome. And now I want to extend it to a case where we want to test the equality of more than two regression coefficients. For example, if I have a hypothesis like H0: beta1 = beta2 = beta3.

So that means I want to test whether the effect of all the three variables on Y is the same or different. And surely if this hypothesis is not accepted that means there is some variable whose effect is different than others and then we try to go for the multiple comparison procedures. So as I said in the during the introduction to analysis of variance that we have different types of multiple comparison test also. But here we have not considered it but it does not mean that they do not exist. The same thing will continue in the case of multiple linear regression analysis also.

So in this lecture we are going to consider the analysis of variance in the setup of multiple linear regression model. We will try to see that whatever expression we had obtained earlier about sum of square due to total, sum of square due to error etc. How they are going to change but remember the concept is the same. The concept will remain the same. Whatever concept I had given you to understand the analysis of variance earlier that will remain the same.

So I am not going to give you once again the introduction and the concept to the analysis of variance but I will try to show you that how those concept can be implemented in a regression framework in this lecture. So I will try to show you that how to conduct the analysis of variance in the multiple linear regression model and how to implement them in the R software and how to interpret the outcome of the R software. So let us begin our lecture and try to understand the ANOVA in multiple linear regression model. So now in this lecture we are going to talk about analysis of variance and its implementation in the R software. So suppose we want to answer a question what is the overall adequacy of the model.

So this can be achieved through the test of hypothesis concerning the regression coefficient because when you took the hypothesis like H0: beta j = 0 then it is trying to indicate that okay means any particular variable is important or not that can be concluded through the test of hypothesis. So now we want to extend this concept to the overall adequacy of the model. So we consider here the null hypothesis about the equality of the regression coefficient and yeah remember one thing we are trying to take here no intercept term. So we are trying to consider here the same model Y = X beta + epsilon but which is like beta0 + beta1 X1 + beta k Xk + epsilon but you have to just be I would like to address it actually. Yes I had told you earlier that in what if I try to consider here the number of independent variables here it is here k + 1.

But the number of explanatory variables which are not taking uniformly the values 1 it is here k and they are associated with beta1, beta2, … beta k and beta1, beta2, … beta k they are the slope parameters. So the analysis of variance is basically concerned with the testing of equality of the slope parameter. So ideally if you go with the earlier setup then this number of independent variables will be actually k - 1 but you will see that in just to handle it practically it is more easier to handle k rather than k - 1. So that is why I am taking it here k and but it does not changes anything you simply have to be careful when you are trying to choose the correct degrees of freedom in the analysis of variance and that test of hypothesis. So remember one thing that I am considering here H0: beta2 =

beta3 = beta k = 0 not from beta1 and my H1 is that at least one beta j $\neq 0$ for j goes from 2, 3, … k.

So this hypothesis essentially determines if there is a linear relationship between y and any set of the explanatory variables x2, x3, … xk, no x1 remember. So this is an overall or global test of model adequacy and if H0 is rejected then it indicate that at least one of the explanatory variables among x2, x3, … xk contribute significantly to the model and this is called as analysis of variance. And because you will see that we are trying to analyze the variance which is obtained through different quantities within group between group and similar type of concept what we have used earlier will be extended to variability due to regression, variability due to random errors and so on. So as I discussed earlier the analysis of variance is based on partitioning the total variation in the values of response variable in two orthogonal components. And these orthogonal components reflect the variation in the data which is explained by the fitted model and the unexplained variation which is due to the random disturbances.

For example if you try to see the example which I have considered that the random variation in the values of y which are the marks of a student this was a function of x1, x2, x3 number of hours of study, number of assignment and number of hours of play and there will be some random variation here. So definitely your model is going to be good if you fit a good model that means the amount of epsilon should be as minimum as possible. So whatever is the total variation in y that is being now contributed due to the model what we have fitted. So it is due to the fitted model and this is due to the random variation. So that is what we try to obtain through the analysis of variance that I try to partition the total variation into two orthogonal components such that the components of variability due to the fitted model and variability due to the random errors they are orthogonal to each other, right.

So I am not going into the whole detail that how are you going to obtain it mathematically and statistically, but I will try to give you all possible steps and the interpretation and its implementation in the R software. So these variations are measured as sum of squares and these are sum of squares due to total which is indicated by here SS total, sum of squares due to regression which is denoted as SS regression and sum of squares due to residuals which is indicated by SSres and this is an orthogonal partition so I can write down SS total = SSregression + SSres, right. So and these sum of squares they are obtained using the ordinary least square estimator V that is more important. So this sum of squares due to total is obtained here like this thing, right. These are the observed values.

So you can see here summation yi square, i goes from 1 to n can be written here as say y1, y2, … yn and then another vector here y1, y2, … yn. So this is written here as say y transpose y, right. You can see here and similarly you can see here this is only a function of here y. Now the sum of squares due to regression is obtained here like this b transpose x transpose y minus this summation yi whole square divided by n, right. So you can see here this sum of squares due to regression depends on the value of b which has been obtained on the basis of given set of data, right.

So this is, yeah, mean sum of squares due to regression if you want to wish you can also partition it into different component that is what is the contribution of the variable x1 in the variable t, what is the contribution of x2, x3, etc. in the total variable t. So it is possible easily in the software actually. So in R the sum of squares due to regression is further partitioned into various sum of squares due to explanatory variables with each sum of squares having 1 degrees of freedom, right. So this SSregression will be extended to SS due to variable x1 + sum of squares due to variable x2 and so on.

Yeah, it is not possible in all the software but R has this characteristic. So that is why I am trying to explain you here. Then the sum of squares due to residuals, this is obtained by this expression y - Xb transpose into y - Xb. So you can see here once again it is dependent on the value of B. So if your b is good then the modeling is going to be good and if you try to expand it you can obtain it as y transpose - b transpose x transpose y and this is obtained here say SStotal that is sum of squares due to total minus sum of squares due to regression.

And yeah, I would like to address here that when we try to write down the sum of squares due to total = sum of squares due to regression + sum of squares due to residual when we try to orthogonal partition there is a theorem which is called as Fisher-Cochran theorem. This actually works behind all these mathematical and statistical details. So sum of squares due to regression and residuals they are independently distributed and the degrees of freedom associated with sum of squares due to total are n - 1 and sum of squares due to regression are k - 1. And if you try to further partition the sum of squares due to regression to each of the k - 1 explanatory variables then each of the sum of squares will have a degree of freedom 1 and the degrees of freedom due to the sum of squares due to the residuals it is n - k. And if you recall then in the case of analysis of variance after obtaining the sum of squares we had obtained the mean square.

So the same thing is happening here also and by the same fundamental definition we are

going to obtain here the mean square for each of this sum of squares. So the mean square if you recall it is defined as say sum of squares divided by the corresponding degrees of freedom. So the mean square due to regression that can be defined as MS regression = SS regression divided by k-1 and if you try to further partition it into a particular explanatory variable so the mean square due to any jth explanatory variable xj is defined here as sum of squares due to xj divided by degrees of freedom which is here 1. So this is here MS(xj) that is mean square due to xj and then mean square due to residual is defined here as sum of squares due to residual divided by the degrees of freedom n - k and it is indicated by MSres. And now the test statistics to test H0 beta2 = beta3 = up to beta k = 0 is this mean square due to regression divided by mean square due to residual and this follows F distribution with degrees of freedom k-1 and n-k under H0.

So this statistics is popularly called as F statistic. So you can recall that the same thing we had done in the case of analysis of variance also where we had partitioned the total variation into within group variability, between group variability and then we have defined a similar F statistics. So the decision rule remains here the same that reject H0 against H1 at alpha level of significance if p value is less than alpha or if you want to do it through the tables of F distribution then reject H0 at alpha level of significance when the calculated value of F which you have obtained here this is greater than the tabulated value of F which is obtained from the tables of F distribution. So now all these calculations they are presented in the form of an analysis of variance table or popularly called as ANOVA table. So this also has the similar structure which we have considered in the case of analysis of variance earlier.

So there will be first column is about source of variation. So in this case the source of variation is due to regression, due to residual and then the total which is the sum of the both. The second column here is about sum of squares, sum of squares due to regression, sum of squares due to residual and sum of squares due to total. In software it will try to divide it into SS due to x1, SS due to x2 and so on right that what I will try to show you in the software. Then the third column is degrees of freedom.

So this sum of squares due to regression has degrees of freedom k - 1, residual has n - k and sum of squares due to total has n – 1 degrees of freedom. Then we have the next column which is about the mean square. So mean square due to regression which is obtained here as a sum of squares due to regression divided by its degrees of freedom. And similarly here we have the mean square due to residual that is sum of squares due to residual divided by the degrees of freedom right. And then you try to take the ratio of these two and we try to define here the, the F statistics right.

You would give here affix statistics which is here the mean square due to regression divided by mean square due to residual. And if H0 is rejected then it indicates that it is likely that at least one of the beta j $\neq 0$ right. So this is how we try to do it and now I will try to implement in the R software using the same example which we have used up to now. So we have a data set of 20 observation on the marks of 20 students which are here.

These are 20 students. Their marks are here indicated by y and the three variable x1, x2, x3 they are indicating here the number of hours of study in a week, number of assignments submitted in a month and number of hours of play in a week respectively. So the interpretation here goes like this as I told you earlier. Therefore the student number 1, the student number 1 has studied for 34 hours in a week and submitted 3 assignments in a month and played 15 hours in a week and has received 180 marks out of 250. And same is true for other students right. So now we try to fit here a model with n = here 20 and my model here is yi = beta0 + beta1 xi1 + beta2 xi2 + beta3 xi3 + epsilon i and I try to store this data on x1, x2, x3 and y in this data vector.

Now you know that how to find out a linear model which was obtained by using the command lm() inside the parenthesis y tilde x1 + x2 + x3. Now in order to find out the analysis of variance of this regression modelling we use the command here ANOVA, anova() and we write down here anova() and inside the parenthesis we write the command which was used to find out the linear model. And you will get here this type of outcome. Now we can see here in this outcome yeah it is a similar what we did in the last term that here there is a response like as here it is here y and the first column is here 3 variable x1, x2, x3 which are actually contributing the regression right. Because regression is depending on 3 variable x1, x2, x3 and then you have here residuals right.

So there are here means if you try to see n = here 20 and k here = 4 and this k - 1 = 3 are the x1, x2, x3 which are the variables associated with the slope parameter no intercept term is required here. So that is why if you try to see here this next column is about the degrees of freedom. So you can see here this is indicating that x1 has got 1 degrees of freedom yeah it is a degrees of freedom about the sum of squares which is given in the next column. X2 has 1 degrees of freedom, x3 has 1 degrees of freedom and residual has say this n - k that is 20 - 4 that is here 16 degrees of freedom right. And similarly if you try to come here to the next column which is here sum of squares.

So the sum of squares due to x1 is here like this sum of squares due to x2 here is 188 and sum of squares due to x3 here is this thing and if you try to sum all of them 21036.8

+ 188.0 + 6056.7 that means these 3 values then you will get here sum of squares due to regression right ok. And then you have here the value 13.6 which is the sum of squares due to residuals. Now after this in the next column mean s cube we have obtained the mean squares. So mean squares have been obtained by dividing the sum of squares by their respective degrees of freedom. So you can see here the degrees of freedom are only 1 1 1. So that is why this first 3 values they are here the same as sum of squares.

But the mean square due to residual that is obtained at 13.6 upon here 16. Then we have here f values which are obtained for x1, x2, x3 separately and then we have here this p value. So this is the structure of the outcome and now if we get me try to show you this outcome one by one. So you can see here this is what I have shown you here right. X1 has a sum of squares due to regression like this it has degrees of freedom 1. The mean square due to regression is here like this divided by 1. Similar is the case for x2 right what I explained you and similar is the case of here x3 right. And for this thing we have obtained here the f value also. So f value for x1 is here like this f value for x2 is like this and f value for x3 here is like this and the corresponding value p value I have written here in the form of an over table here right p(x1), p(x2), p(x3).

And then the residual sum of a square this is obtained here like this one degrees of freedom n - k total sum of a square that can be obtained by summing all the sum of a squares and which has degrees of freedom n - 1 that is 20-1 = 19. So you can see here these are the values of here f statistic which have been obtained. Now let me try to take them one by here one and I try to show you that how they have been obtained in a better way right. So if you remember the expression for the sum of square due to regression was here like this right. So this has been used to obtain these sum of squares.

The expression for the mean sum of a square due to regression was like this sum of a square due to regression divided by k - 1. So yeah this is for the total. So if you try to make the sum of these things and then try to divide it by here this k - 1 you will get the answer. And then you have here f values we have been obtained by m s regression divided by MSres. So we have three values here like this and these are here the p values.

If you try to see here this f value and about here means you have to compare it with the value of alpha right. So you can see here each of these values is much much smaller than alpha right. So and then yeah it is these three stars they are indicating the level of significance code right which are here like this one that that has been used in the software. And anyway means you can choose your own region by choosing different values of alpha.

So you can see here in each of the case alpha was 0.05 and the p values are much much bigger than p. So what you can see here that p values are less than alpha. So you are going to reject the hypothesis H0. And then s s residual that can be obtained by SStotal - SSregression here like this and MSresidual is obtained by SSresidual divided by n - k right. So let me try to show you these things on the R software also and right.

So if you try to see here that first I need to input my data. So I try to input my data here like this and then I try to take here this ANOVA command right. You can see here this is the same outcome which I have shown you on the slide also. And if you try to see here these are the sum of square this is mean is square F value p value and so on right. So you can see here these are the values of the alpha indicated by this star right.

So this is the three stars are indicating to the this value here 0. And now you can simply take the conclusion by comparing the p values with alpha right. So now we come to an end to this lecture. So you can see here now you have extended the concept of analysis of variance which you studied in the context of normal population to this regression analysis also. And regression analysis is giving you an idea about the overall adequacy of the model.

So here in this case for example, H0 is rejected that means beta1, beta2, beta3 they are unequal and they are $\neq 0$ and it is good that is what we expected that if you are trying to analyze the data on the marks of the student they are going to depend on the number of hours of study, number of hours they play, number of assignment they solve that is correct. And that is why we always say it could a student that study you can only you will get good marks. So if you try to see with this analysis of variance also you are getting a similar type of conclusion what you had obtained earlier. In a univariate case also you obtain the same outcome that H0: beta0 = 0, H0: beta1 = 0, H0: beta2 = 0, and H0: beta3 = 0 were also rejected that means they were important variables. But now my idea was not to do this analysis but to show you the application of this tool and to tell you how you can take the correct conclusion.

In case if this H0 is means if you want to go further to analyze the data then you can go for this multiple comparison test. For example in this case you can see the H0 is rejected that means the 3 factors or 3 variable which are going to affect the y they have got different effect but if you want to group them that which are the variables which have got the similar effect. If you want to group them then you have to go for multiple comparison test which I am not doing it here but you have to do it yourself if you need it and they are

available in the R software without any problem. So now my request as usual is that you try to take some example, try to take some data and then try to practice them. The more you practice better you will understand it, try to think about this analysis of variant concept that how it has been extended to different thing.

So analysis of variance is actually a tool that can be used in different situation. For example designing design of experiment we have one way analysis of variance, two way analysis of variance, three way analysis of variance etc. So try to think in now this direction and there can be many situations in data science when you are trying to handle a more complicated situation where the tool of analysis of variance can help you in different ways which even I cannot think at the moment. It depends on your capability that how you can apply the same tool in under different type of condition, different type of situation. So you try to think, you try to practice and I will see you next lecture in the next lecture till then good bye.