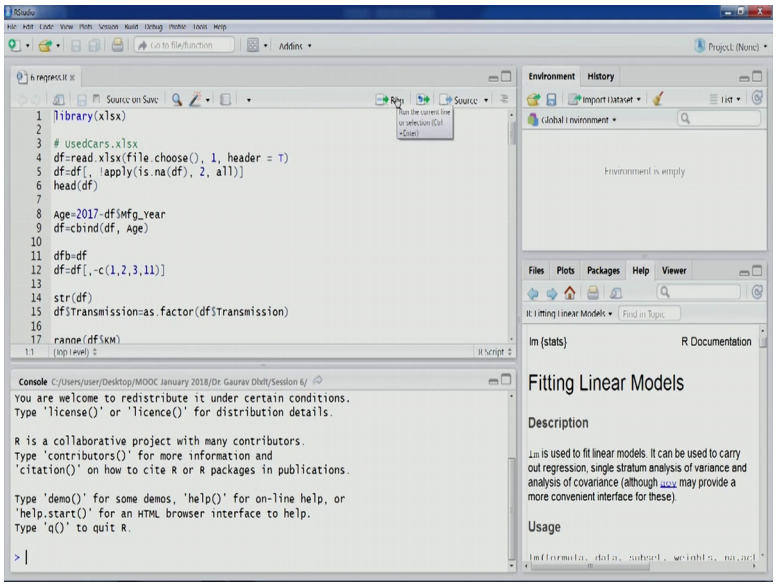
Business Analytics & Data Mining Modeling Using R

Dr. Gaurav Dixit
Department of Management Studies
Indian Institute of Technology, Roorkee

Lecture Number – 24
Multiple Linear Regression – Part III

Welcome to the course business analytics and data mining modeling using R, in the previous lecture we started our discussion on multiple linear regression, and we discussed the theoretical most of the theoretical part of it, and we were doing an exercise using r studio. So, let us start from there. So, let us open r studio.

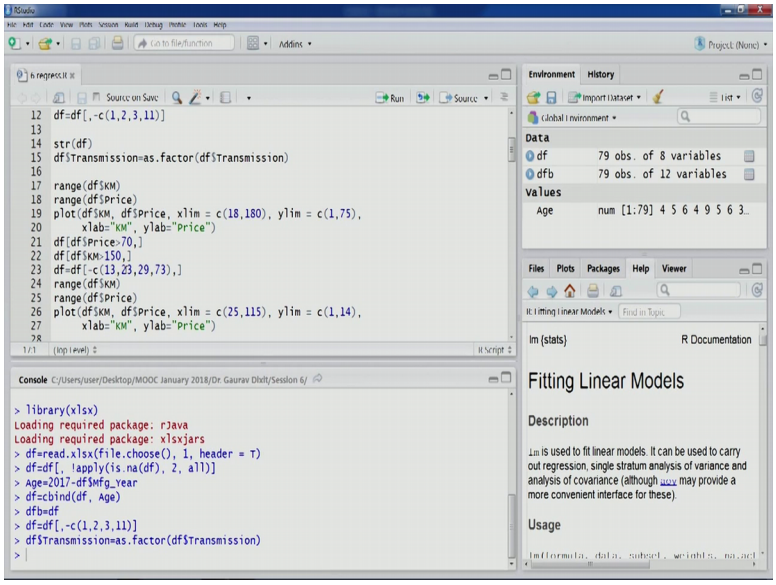
(Refer Slide Time: 00:44)



So, last time you are able to produce the results of the regression analysis. So, will going to reproduce the same, and then we are going to pick up our discussion from the same point.

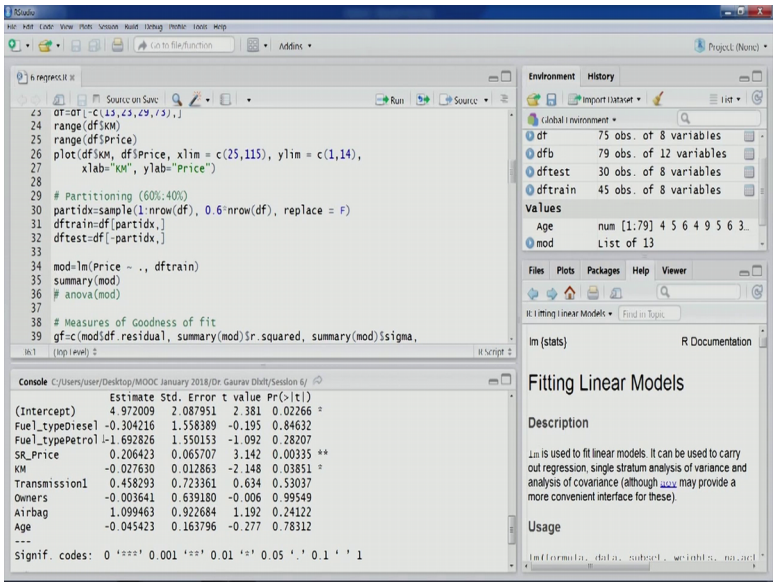
So, let us load this library the file the data set that we have been using. Let us move na columns let us create this variable age then we do not want let us take a backup and then we do not want few columns.

(Refer Slide Time: 01:28)



So, let us eliminate them then let us also change this variable type. Now in the previous lecture we also eliminated few points. So, let us again do the same thing here now let us also do the partition.

(Refer Slide Time: 01:57)

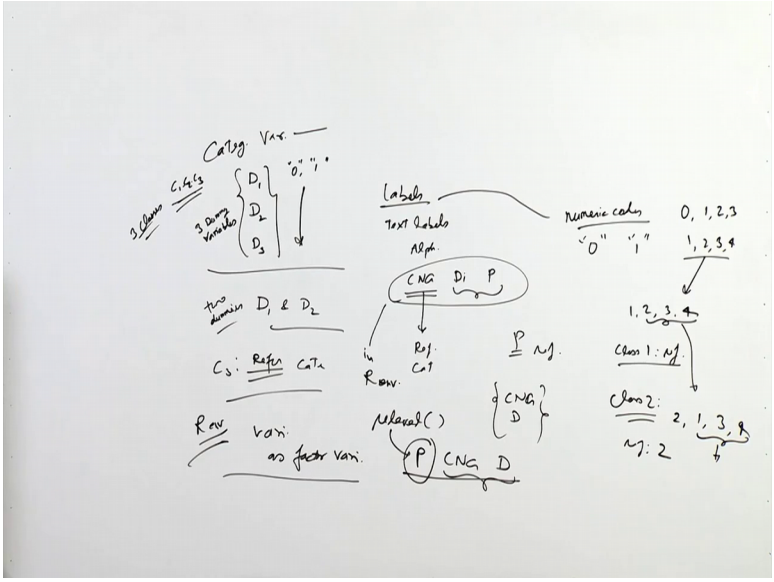


So, we were discussing the results of this regression analysis. So, let us come to that part. So, here at this point we were discussing the results of this particular regression this particular regression analysis, and we talked about the fuel type variable being the categorical variable the factor variable in r environment and how in the results we see these 2 category fuel type diesel and fuel type patrol, similarly for transmission which was converted into a factor variable.

So, this we can see as transmission 1 is there instead of transmission. So, how this is happening and why this is happening right. So, as discussed in the previous lecture as well, that any variable any categorical variable which is having data in textual format string format, it is going to be automatically treated as a factor variable in r environment. Any variable which is in the numeric which is having labels in 0 or once or some numeric forms in form the numeric code form. So, that will be treated as a numeric code.

So, therefore, we would have to convert that particular variable into factor variables. So, which we did for the transmission, now once you have converted your categorical variables as i's factor variables in in r environment right.

(Refer Slide Time: 03:41)



So, then another thing that we were discussing in the last lesson that for a particular categorical variable, having many categories many labels; let us say the many classes that it has c 1, 2 let us say c 1, c 2 and c 3 alright.

So, it will have. So, we talked about dummy coding, we talked about that we will have to compute 3 dummy variables. If there are 3 labels and then we also talked about that we will have to include only D 1 let us say only D 1 and into only 2 of the dummy variables in this case we have 3, three classes we have 3 classes we talked about that we would have to create 3 dummy variables this is the part that we were discussing in the last lecture and then 2 of 2 dummies will have to incorporate in our model the remaining one would be the reference category.

So, the remaining class that is there that class 3 would be the reference correctly, and we talked about that any results and interpretation of the results that we get for these dummy variables for this categorical variable, it has to be the interpretation has to be with respect to this reference category right. So, we talk why. So, talked about that in an r environment, we also displayed this that generally if you the labels the way they are created, labels if it is textual labels, text labels right then they would be in alphabetical order. So, the ordering of those labels would be alphabetically.

So, if you do not change the labels. So, the default; for example, in this case we had CNG and diesel and then petrol. So, therefore, this would be the ordering of the labels. So, for the categorical variable, and by default CNG would be treated as the reference category and the remaining 2 would be part of the model.

So, this is what has happened in the results you can see that fuel type diesel and fuel type patrol, they are part of the model they have the regression coefficient has being displayed in the output and the CNG has become the reference category. But if you wanted patrol to be your reference category and you wanted CNG and diesel to be incorporated in your model, then you would have to change this ordering in r you would have to change this ordering in r environment right.

So, there is different functions that are available to perform this ordering. So, this particular change in ordering we will cover and in coming lectures, where we would be requiring such a where will be in such a situation, that we would be requiring the change in labels. So, relabel is the function that can be used in r environment, to actually perform this change in. We will do this using an exercise in coming lectures you can see, in the help section variable reorder labels of factor.

So, this particular function can actually be used to change the labels, change the ordering of these labels and therefore, whatever levels comes first that would actually we after changing let us say you wanted patrol to be your reference category, you can have an ordering like this and now once you have changed this ordering, then patrol is going to be your reference category and CNG and diesel would be part of your model.

So, depending on the task at your hand depending on the kind of interpretation, that require you require depending on your requirements of reference category, which particular class you want to treat as reference category, and which particular classes you want to have in the model. So, that you will have to decide and accordingly you will have to change it. Similarly if your labels are numeric codes, which is generally the case. So, they would be 0 or once for example, if the it is we are talking about if its 2 class case, but is more than 2 class case then the labels could be anything. Depending on the way data set is presented to you labels could be in this in these numeric codes. So, they could be in any ordering which numeric codes that if they are being used they could be in any ordering right.

So, there also if you want to change the labeling they are in r environment put again be in the increasing value order right. So, for example, if you have 1 2 3 and 4 levels, the labeling would be 1 2 3 and 4 and therefore, the class 1 would become the reference category and the others would be included in the model. So, therefore, if your task requires you to have a class 2 as reference category then you will need to change this order all right. So, again same thing you can do using the relabel command. So, you will have to change these orders, and we will be doing in the coming lectures we would be doing such an exercise and then your ordering would change and your reference category will become 2 right and other categories would be part of the model part of the results that we produce right.

So, this is about this is mainly about the categorical variable and the labeling and the reference category that we need to take care. Now another difference between r and other commercial softwares other statistical software is, that in other statistical software you might have to perform your dummy coding explicitly, and then bring that those variables dummy coded variables into your model. In r if you are having text labels or once you have declared a variable in R environment once you declare a variable as factor variable that is nothing, but the categorical variable then most of the packages most of the functions that are available they take care of the dummy coding.

So, you do not have to explicitly convert your categorical variable and do these dummy variables zeros and ones as we talked about in the previous session presence or absence of that thing. So, this part we have discussed in the previous lecture. So, you do not have to do this dummy coding 0 and 1 in R environment, if you have converted your variable into a factor variable. In other commercial statistical software, you will have to perform this dummy coding on your own there are many software provide specific utilities for the same. So, therefore, it is quite easier given for in those softwares also to perform the system according.

But you will have to perform dummy coding first and then you will have to pick your variables right and then you will have to include them in your model that is how it will happen in other softwares. In r you just need to convert your variable into a factor variable irrespective of whether the labeling is text or based or the numeric codes based. Once you convert them the most of the things would be taken care of, only thing that you would be required to change is the reference category.

Now, let us look at the results that we have to as be discussed the l m function the formula, that is being passed here and we also we can also see them the descriptive charts about residuals and then we come to the coefficient part. Here we can see the intercept, the then the fuel type diesel and fuel type patrol, but we can see that neither of this fuel type are significant significance how do we see in the results, you can see at the end of this particular results you would see a asterisk there.

Now, these if any of the variable any of the row has asterisk, star then that particular variable that particular pre predictor has a significant relationship with the outcome variable. So, now there are different labels of significance. So, you can see between 0 and point 0 0 1 if. So, if your p value which is the last column here this your p value. So, if this p value the concept of p value we have discussed in supplementary lectures. So, you are recommended to watch those videos. So, if your p value is lying between 0 and 0.001 then that is indicated using 3 star. If your p value is lying between 0.001 and 0.01 then that is indicated using 2 stars, if your p value is lying between 0.01 and 0.05 so, that is indicated by a single asterisk.

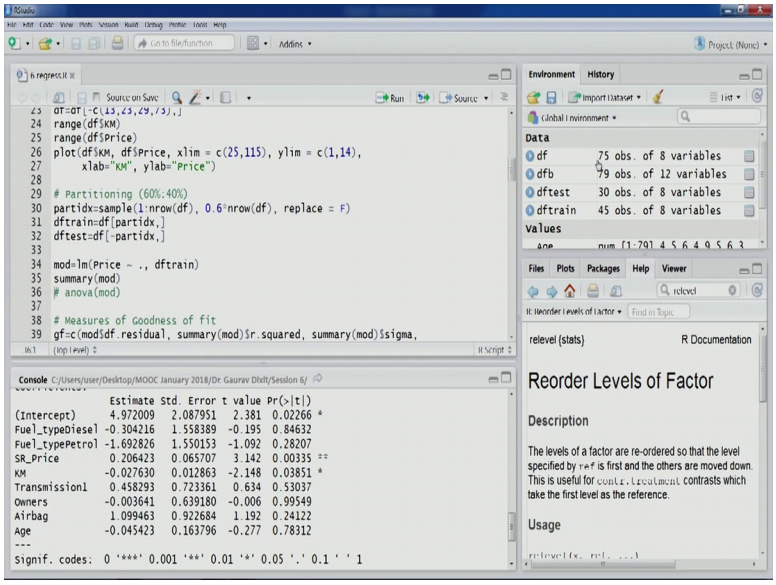
If your value is between 0.05 and 0.1. So, that is being indicated in this case is dot otherwise it would be a blank. So, we are generally you know. So, these ranges actually also signify the confidence interval for example, if we have single asterisk here; that means, the value is between point zero; that means, the value is less than 0.05, that actually also signifies that we have accepted the confidence we are working with the confidence interval of 95 percent. If the value is less than 0.01; that means, 2 asterisk would be used. So, therefore, we are looking for the relationship with 2 asterisk or more and therefore, the 99 percent confidence interval is being used if it is 3 star then it has to be 99.9 percent.

So, it put up the value is less than 0.001, then we are deal we are working with the 99.9 percent confidence interval. Sometimes people might also pick even the 90 percent confidence interval, in those situations the p value is going to be less than 0.1. So, depending on the acceptance of your the confidence interval that you want then the level of acceptance of your error, that you can select the significant relationship.

For example, the most typical confidence interval that is residual is 95 percent and generally we are looking for relationships which are having 3 asterisk 3 star significance levels. So, if we look at the results any particular rows belonging to different variables that we have in the model, if there are no such asterisks as discussed no no such codes so; that means, those are in significant relationship.

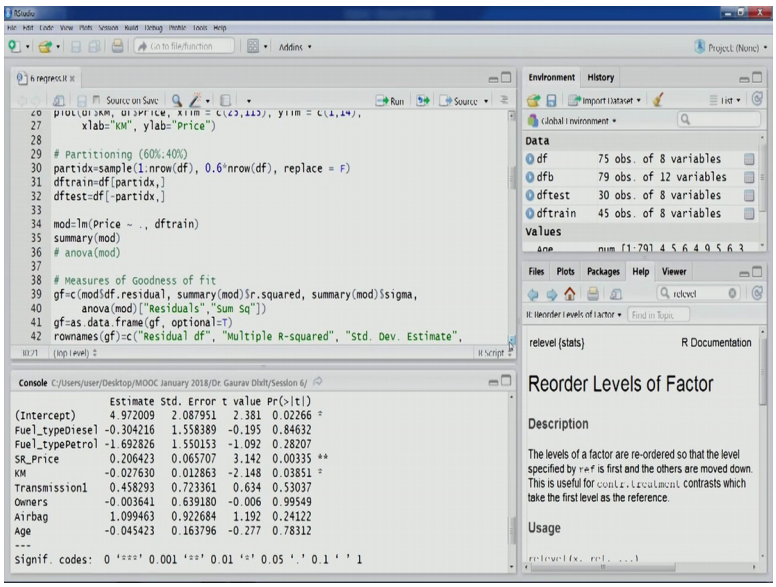
So, the significant relationship that we can see is SR price, that is show room price having 2 asterisk in this case. So, therefore, have being significant relationship at 99 percent confidence level. We look at the k m also. So, this is 1 astrick is there. So, therefore, this relationship is significant at the level of 95 percent confidence interval. Other relationship if we are ignore if we ignore the intercept term then the main predictors that are there others seem to be significant. Please also bear in mind that we are dealing with a data set.

(Refer Slide Time: 16:30)



That is just having 75 observations right this is a quite few number of observation in this case. So, therefore, the result or not that was not stable and therefore, the relationship so that we see the significance of relationship only 2 of them are being significant, that will also change depending on the partition we select. So, at this moment the partition that we have created 60 percent 40 percent, and the observation that had been randomly selected into the training partition; data in a way determining the relationship the significance of relationship and all that. So, therefore, it is generally advisable to have a much larger data set. So, that we can overcome the problems that could be there to do smaller sample size.

(Refer Slide Time: 17:23)

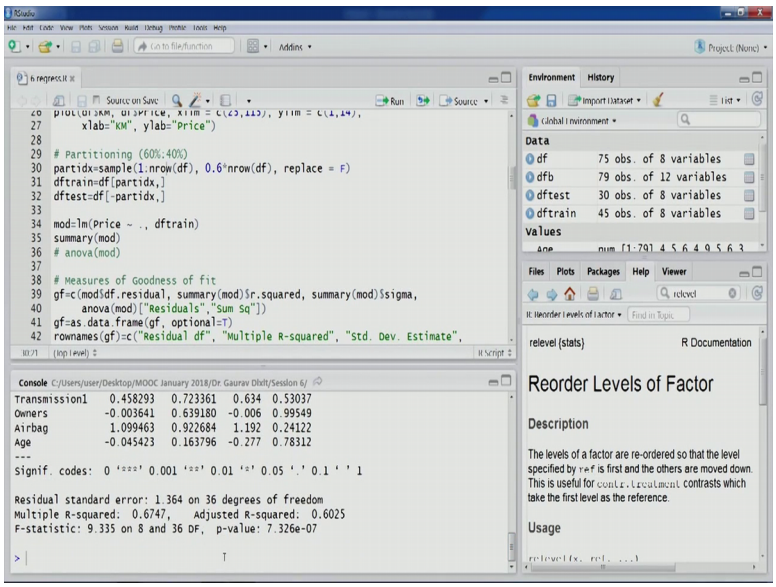


So, currently what we have is the smaller sample size and it might not be reflective of the true relations relationship between the outcome variable of interest, and the set of predictors that we have. Now other things that you can see in the regression results is, the estimate the this is regression coefficients for all these variables. So, we can we generally also look whether these estimates how you know if something is less than the value of residual standard error right, if something is less than that then probably even if those coefficients are having significant relationship right.

So, be even drop those variables. So, some of the insignificant variables that we see we can easily drop them to reach our final model, and even some of the predictors which are having this particular estimate the regression coefficient less than the residual standard error that is also a candidate for elimination. That all that predictors also could be candidate and those variables would reduce dropping of such variables would reduce insignificant dropping of insignificant variables and also variables having even though having significant relationship, but very small regression coefficient or small estimate as in this column that would improve the, that would actually reduce the variability in the prediction errors and which is the desirable thing for us.

So, this particular column gives us the regression coefficients for different variables. So, for example, in the SR price you can see 0.2 is the value, and for k m you can see minus 0.02 is the value. Then we have a standard error as well. So, for all these relationships standard erorr as also been given and the t value and p value is also given based on we can find out the significance of the relationship as discussed.

(Refer Slide Time: 19:33)



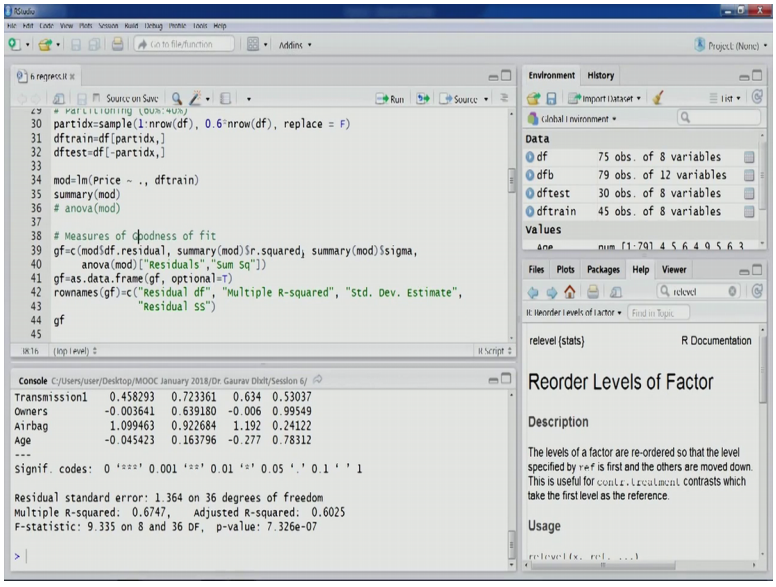
Other statistics that are available in these results are residual standard error and the degrees of freedom, we can also see the multiple R square which will discuss coming lecture.

So, multiple R square captures the proportion of explained variance. So, in this case we can see 67 percent of the variability in the outcome variable is being explained the value being 0.6747. So, 67 percent x the variance is being explained. We also have on the adjusted R square there is some difference between multiple R square and adjusted R square we will discuss this also in the this lecture are coming lecture, you can see adjusted R square value is slightly on the lower side that is 0.60 to 5 that is about 60 percent lower than multiple R square value.

So, we will discuss this why this value is lower and generally we prefer to look at the adjusted R square value this con considered to be a much better criteria to check the performance of the model. Especially in statistical sense we are generally looking for this R square value because there our idea is to find the model which is best fit to data. So, therefore, R square values whether its multiple R square and or adjusted R square indicating the proportion of variability that is being explained by the model, thereby indicating the fitness of model 2 data therefore, we always look for higher value, and then adjusted R square being a better indicator than multiple R square.

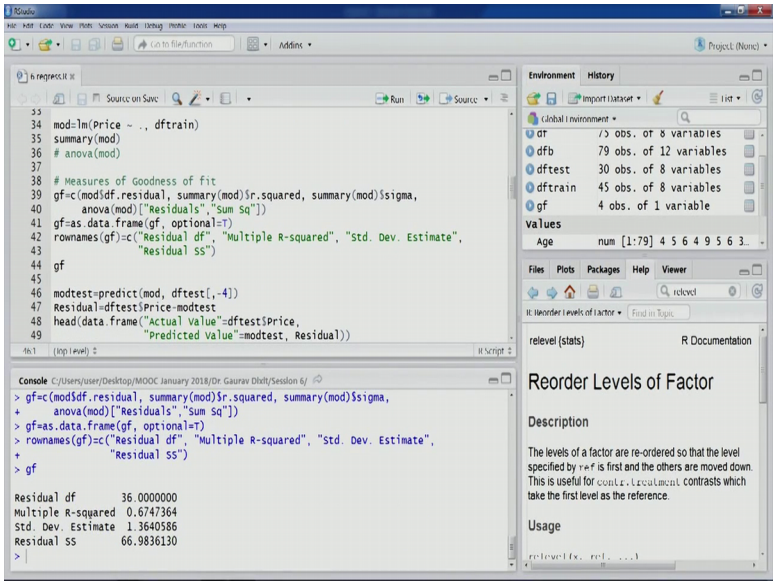
So, we always look for the adjusted R square value for different models right in different variation that we could have for this particular regression analysis that we are doing. So, we look for the model having higher adjusted R square value, then we have F statistics which be covered in our supplementary lecture. So, these values are also they were indicating the significance of the overall model that we have.

(Refer Slide Time: 21:53)



Now, let us look at few more measures. So, these measures that we are going to compute now they are mainly for goodness of fit. So, you can see that we are trying to find out the this residual value right for this model and d f degree of freedom for the degree of freedom for the residual, and then we have r squared value and then the sigma value right and then the residuals sum of square. So, all these well we are going to compute. So, let us execute these codes.

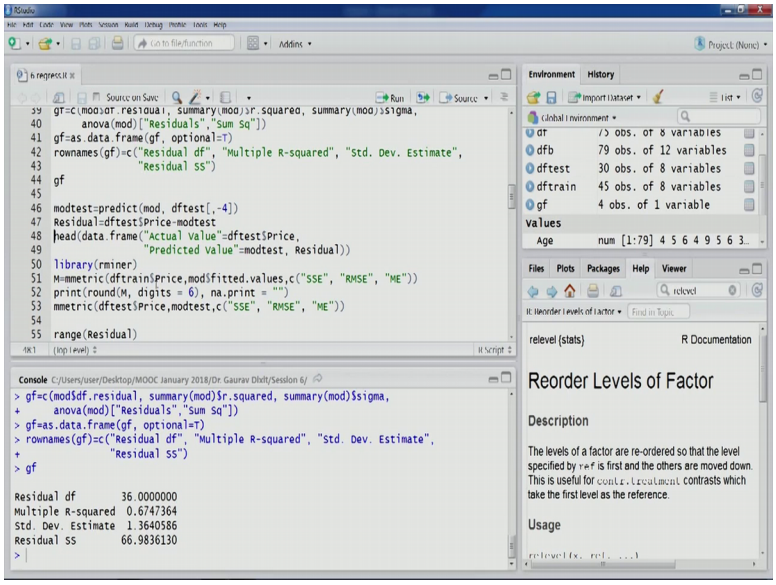
(Refer Slide Time: 22:29)



So, these are a few more statistics that we can discuss residual d f this was all 36 is there multiple R square value, and that is the same that we discussed before 67 standard divia deviation estimate is also there, residual as s s is also there residual sum of square is also there. So, sometimes we also like to a compare residual s s value, if there are more than 1 model if there are many candidate models.

So, some of these numbers we would like to compare and then we can decide the find the most useful model or best performing model.

(Refer Slide Time: 23:13)



Now, let us test the performance of this particular model on test partition. So, we are going to score the test partition first. So, you can see thus the function that we are going to use is predict. So, predict function its quite a generic function. So, for many different techniques for many different methods, this particular function just like summary function this is available

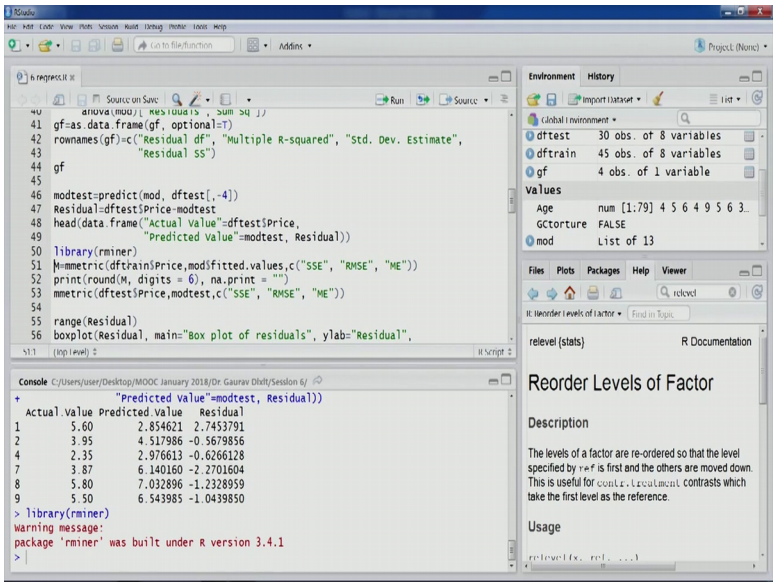
So, that on any new data, that we are able to score and in new data using the model that we have just built for example, mod is capturing mod is representing the model that we have just built using training partition, now we are going to score this data that is test partition d f test you can see that we are eliminating the fourth column which is actually the outcome variable.

So, we do not need that right. So, it is the model which is going to be scoring this, but data set having just the prediction information. So, just to clear that this that we do not need the outcome variable we are trying to predict the outcome variable and therefore, this particular column has been eliminated. So, let us score this test data set. Now we can also compute the residuals for test data set the error values for test data set.

So, now you can see that outcome variable this is representing the actual values now we can subtract the predicted values and find these errors or residuals now let us create this data frame and let us have a look at the first six observations. So, you can see the actual value and the predicted value and the residual. So, for first 6 observations out of the 75 total, we can see these 3 value actual value predicted value and the residuals now if we want to as we discussed that we wanted to the main idea being why in the main idea behind scoring the test partition is we wanted to compare the performance of our model in training partition with test partition

So, for this we are going to require this particular library r miner. So, this particular library has functions, which can be used to compute various matrix that we talkeed about in our previous lectures on performance matrix. So, let us load this particular library.

(Refer Slide Time: 25:45)



Now, M metric is the function that is used. So, in this particular function the actual values are the first argument. So, you can see d f train price is the first argument.

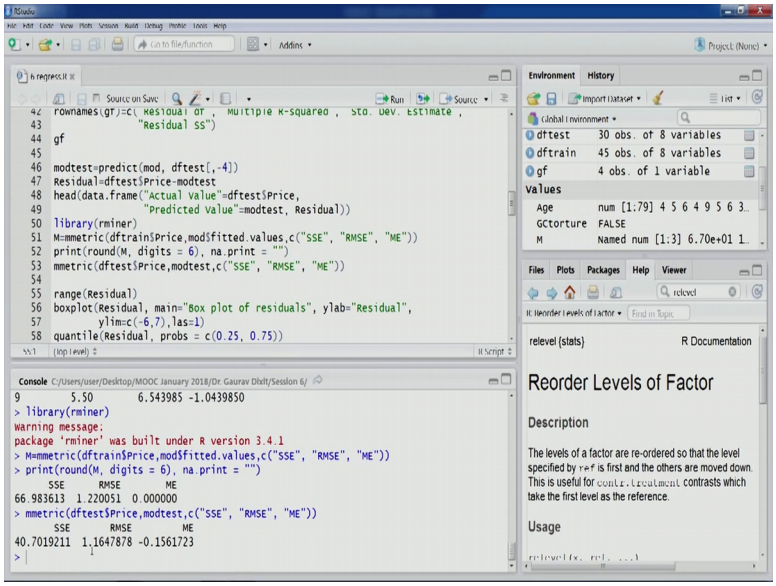
So, here you can see you are also computing the matrix containing partition as well, and then the fitted values that are predicted values right. So, that is the second argument and then the third argument is a character vector indicating the matrix that we want to compute. So, in this case we want to compute SSE that is sum of squares error then RMSE and ME mean error. So, all these matrix these 3 matrix is specifically we want to compute.

Now, once this has been done you can see we are using another functional print, that is just 2 because we are not interested in the actual value till the last decimal point, we would just be interested in few decimal points and because the main idea is to compare. So, print function I am using and rounding of the results that we have got gotten got in the previous line and we are restricting the numbers to 6 decimal points right. So, let us print these numbers.

So, these are the numbers and the value for SSE RMSE ME and if you look at the previous value the residual s s that 66.98 that we had this is nothing, but SSE again the same number is coming there the new matrix that we have computed RMSE and ME.

Now, let us look at the our values on test partition. So, again we are going to call this function and m metric and you can see first argument now for the test partition. So, the actual values and then the mod test this is these are the predicted values which we have just computed using the predict function right. So, this is scoring has already been done and the same matrix. So, we want to compute the data for the same matrix and you can see the numbers.

(Refer Slide Time: 27:55)



So, in this case you can see that the SSE value that we have is even lower in the case of test partition. So, the model seems to be performing better in the test partition in comparison to training partition. So, this is fault because we are having a smaller data set. So, this seems to be more because of the chance thing. So, it is always advisable to have a larger data set. So, that we are more sure about the results that we have.

So, in this particular instance the exercise that we have done, the our model seems to be performing better in new data in the test partition. If we look at the RMSE value that is also on the lower side. So, RMSE value is a much better indicator as we have discussed in the previous lecture, because it gives us the number in the same unit as the outcome variable. So, we can see that RMSE value for the test partition is also on the lower side in comparison to training partition if you look at the ME value.

So, we can see that in the training partition the ME value is 0. So, therefore, the ME on an average level model is neither over predicting nor under predicting for the training partition, but if we look at the ME value for test partition we can see that this is negative. So, on an average level model is under predicting the values. So, from this we can see the results that we have, we can say that the model is robust model is giving stable performance or even better performance.

But; however, the main caution being that this exercise that we have done is on a smaller data set and therefore, the results also depend on the partitions that we created and the observation that are selected and those partitions. So, therefore, it is always advisable to have a larger sample size and then trust your results. So, will stop here and will continue our discussion on multiple linear regression in the next lecture.

Thank you.