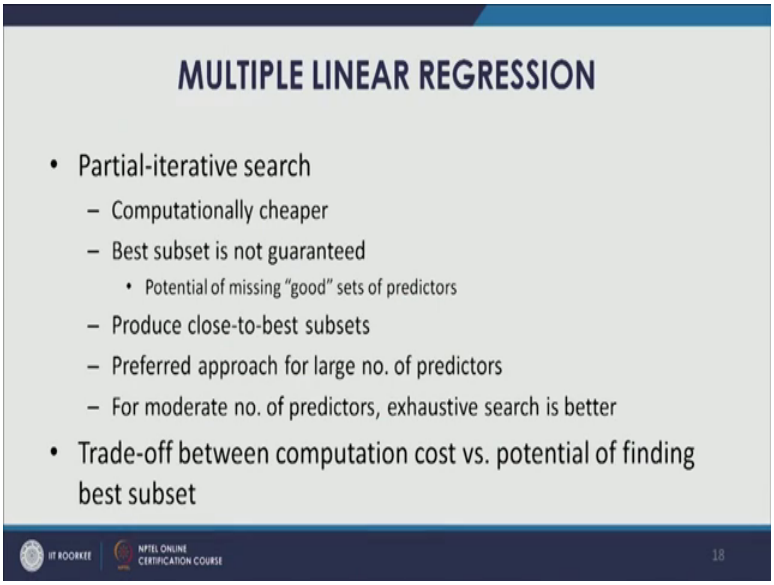
Business Analytics & Data Mining Modeling Using R
Dr. Gaurav Dixit
Department of Management Studies
Indian Institute of Technology, Roorkee

Lecture – 27
Multiple Linear Regression–Part VI Partial Iterative Search

Welcome to the course business analytics and data mining modeling using r. So, in the previous lecture, we were discussing a multiple linear regression. And we concluded our discussion on exhaustive search. So, the next approach that we use for variable selection and also for dimension reduction is the partial iterative search. So, there are a few algorithms under this particular approach that we are going to cover.

(Refer Slide Time: 00:43)



So, let us start our discussion. So, partial iterative search; this is a computationally cheaper in comparison to the exhaustive search that we do. So, in exhaustive search we try out all possible combinations of predictors. So, therefore, it is more like a brute force approach, and therefore, the partial iterative search which works on different algorithms. It is slightly computationally cheaper, but of course, there are some pitfalls like best subset is not guaranteed right.

So, there is always going to be this potential or missing some good sets or predictors. So, what we actually get is we produce close to best subset. So, that that is something that we

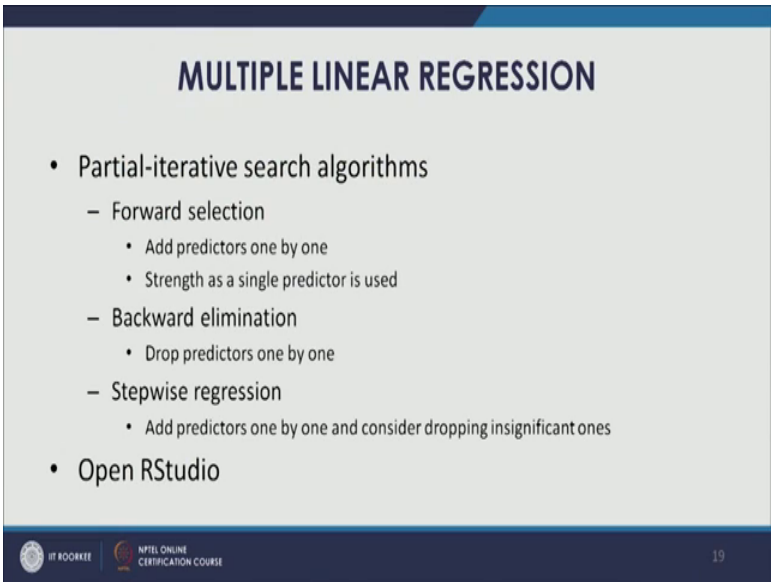
can say that close to a best subsets are definitely we are going to get out of applying partial iterative search.

So, this particular approach is preferred, when we are dealing with large number of predictors right because their the computational time that might be required an exhaustive search might be slightly on the higher side. So, therefore, we would prefer to apply partial iterative search in those situations. So, otherwise if we are just dealing with the moderate number of predictors or going low number of predictors, their exhaustive search is better, because we get a us you know some sort of guarantee of producing the best subset model.

So, therefore, with this between these between these 2 approaches, we can say that there is going to be a trade-off, between computation cost versus potential of finding best subset. So, if we want to you know minimize the computation time, and then probably a partial iterative search is the way to go if we want to you know, if we do not want to compro compromise with the potential of finding best subset, then probably we have we should apply we should employ the exhaustive search.

Now, under this partial iterative search approach we have 3 algorithms that we are going to discuss. So, first one being forward selection so, in the forward selection algorithm.

(Refer Slide Time: 03:00)



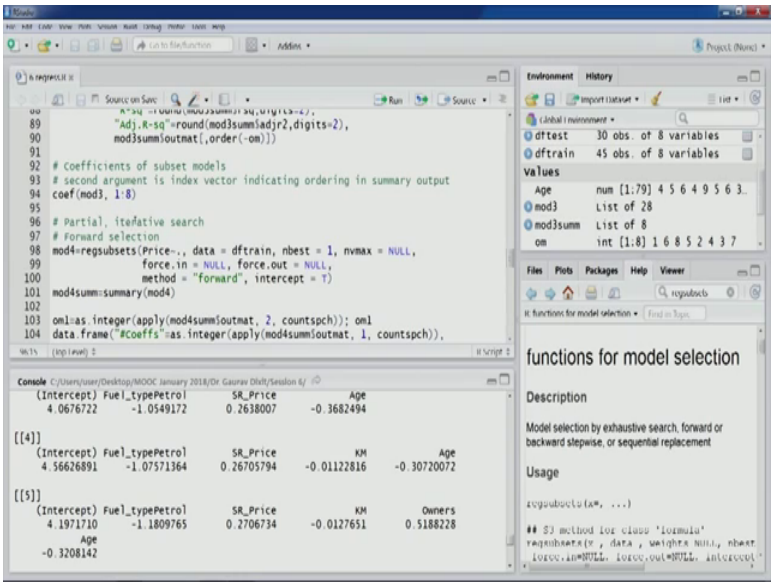
We start with 0 predictors so, we start with no predictor, and we add predictors one by one; so and then the as we go along, the main idea that what happens in forward selection is that strength as a single predictor is actually considered in this approach. So, if a predictor because we are adding one by one, and if it is significant then it would be it would remain in the model, if it is not done it would be excluded from the model. So, therefore, for a particular variable to appear in the in the model to be in the model in a forward selection algorithm, it is strength as a single productor should be on the higher side.

So, that also being the limitation of this forward selection approach. So, we start with no predictors and then we start adding one by one. So, the predictor is significant that it will remain there. Then the second approach is backward elimination. So, in this particular approach we start with all the variables, and then we start dropping them one by one. So, the insignificant mostly insignificant variables are dropped first, and in that order we keep on building the models.

And we keep on dropping till we reach the saturation where all the predictors that are present they are significant. So, this particular approach backward elimination, there is no you know obvious limitation of this approach, except that the computational time that would be requiring this approach would also be slightly on the high higher side in comparison to other partial iterative approaches.

And the third one is a stepwise regression so, in this also we start with just like forward selection approach. So, we start with no predictor and then we add predictors one by one; however, we can consider dropping in significant ones in this particular approach. So, as we move along. So, we keep on adding predictors one by one, and if we if there are some insignificant ones we can consider at any step whether we would like to drop them.

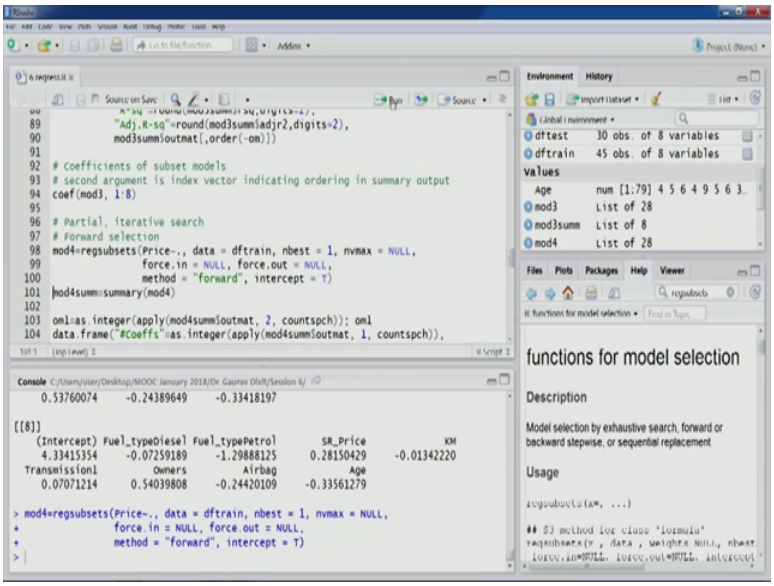
(Refer Slide Time: 05:23)



So, these are 3 main approaches. So, let us understand them a bit more using an exercise. So, a partial iterative search first we are going to start with the forward selection approach. So, in this as you can see the data set that we are going to use is the same; that is, a used car data set so, this is pre-loaded partitioning is already done, and we are dealing with the what number observation that we have is 75 up to excluding all the outlets that are there, and the variables often trust only 8 variables are there, ba one being the price that is outcome variable of interest. So, again you can see that we are using this dead subsets function, and the price and then this formula is specified as price tilde dot that includes all of the predictors in the data set. Now you can see the next important argument data is mentioned as d f train and then method is forward.

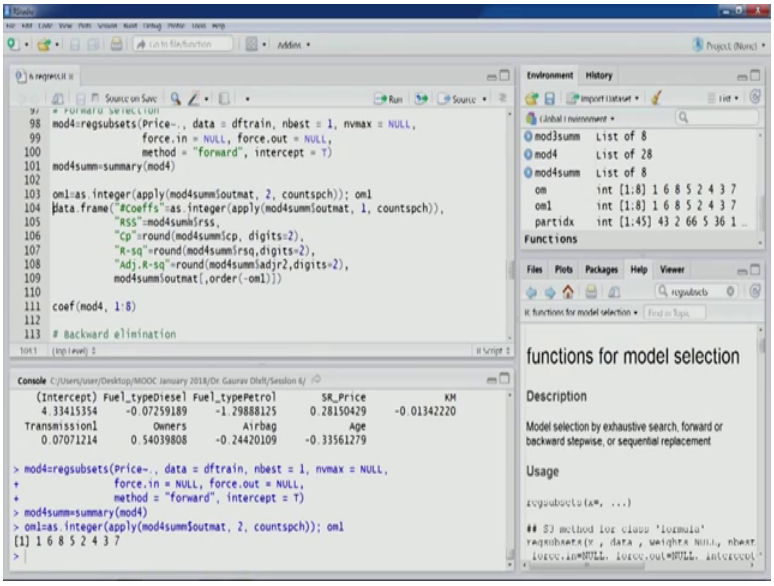
So, that tells the function that we would like to apply forward selection algorithm. So, let us execute this code.

(Refer Slide Time: 06:27)



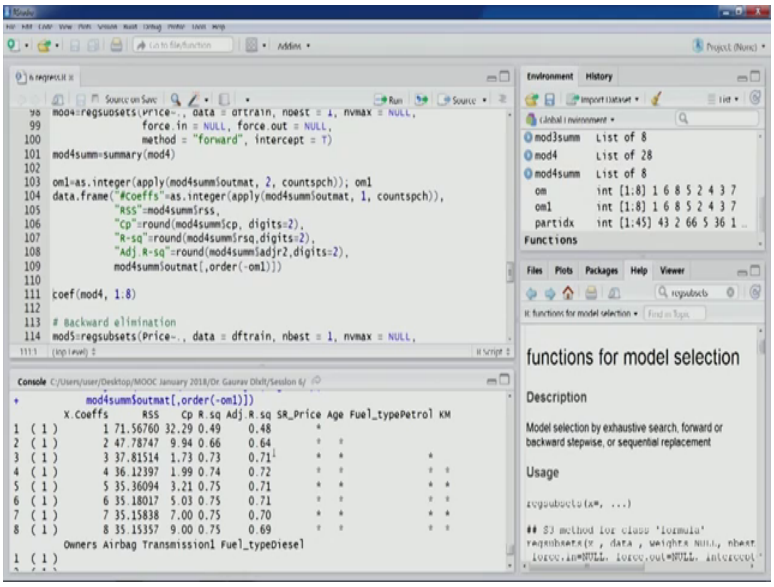
Now, let us generate the summary of this result this particular model. Now as we did in the previous lecture that we would like to count the special character that is asterisk in this case. So, that the particular function count is special character is already created here you can see here for the special character. So, we can use this particular function, and count the number of asterisk, number of asterisk that are there in this particular matrix for different columns. So, this particular result we are going to as we discussed in the previous lecture, this particular result is going to be used as an index vector for us later on.

(Refer Slide Time: 07:13)



Now, as discussed before, first we will have the number of coefficient and different subset models that we produced that we generate here. then the R s s that is residual sum of a square then c p mallow c p then R square followed by adjusted R square, and then we will have the output matrix covering all the variables that are there and whether they are present in that particular model or not. So, let us execute this so, we get the output here.

(Refer Slide Time: 07:49)



As you can see, when in this in this case forward selection case, that are this adjusted R square will it starts from 0.48 and it keeps on increasing till a 4-variable model, right, this particular row 4th row it keeps on grazing. So, in terms of adjusted R square the result of forward selection and exhaustive search seems to be same.

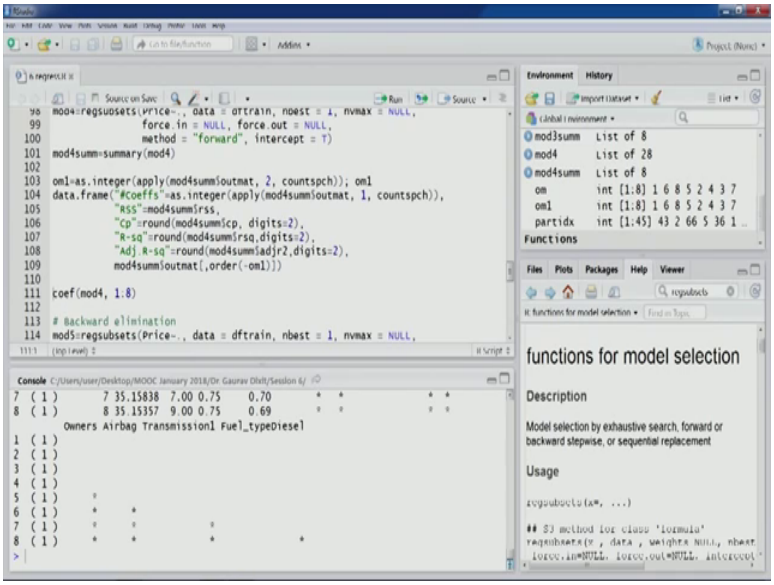
Then after that after reaching 4 variable model and this highest value adjusted R square value of 0.72 this adjusted R square value starts decreasing right as we discussed that there is going to be a penalty for increase in the number of predictors, and with respect to the contribution of information the amount of information.

So, 4 variable model is the model that we have to select if we follow the criteria of adjusted R square. If we look at the R square value, this is also a very similar to or is exactly same as what we had an exhaustive search 0.49, 0.66, 0.73 and then finally, reaching 0.74 and then 0.75. So, same result and the numbers are also same for c p value though we had the, we have with us the results of exhaustive search. So, you can see

these are this is this particular table is the results of exhaustive search, and you can see the same numbers are there the output that we have got from forward selection are happens to be happens to be same for forward selection. So, this is just in this in this particular case the data set that we have and the partitioning that we perform. If we change the partitioning the results might also change and also what we are dealing with a small data set. So, therefore, we might not see much difference between these 2 approaches.

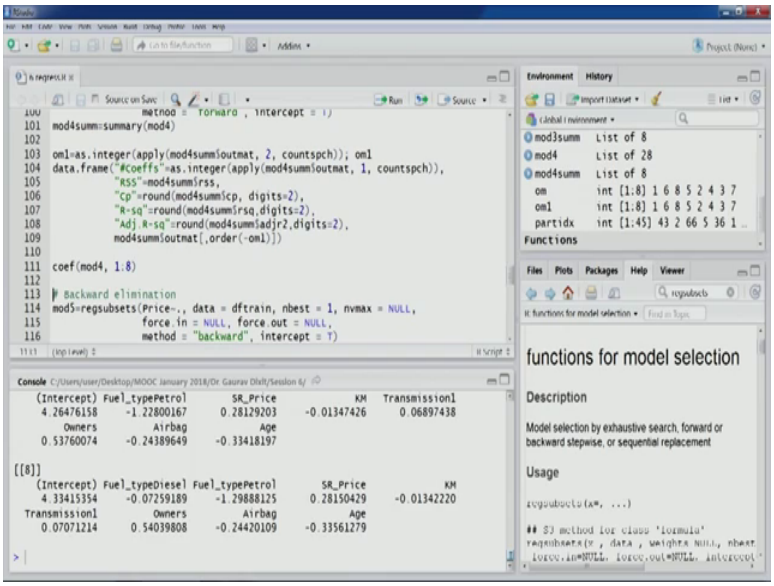
Now, if we look at the variables, you can see show room price a surface being present in the all 8 models, starting from one variable model to 8 variable models and all of them it is present then age is present in same in or such models starting from 2 variable models to 8 models. So, early fuel type petrol is there then k m is there. And then comes the owners and other variables that are there.

(Refer Slide Time: 10:05)



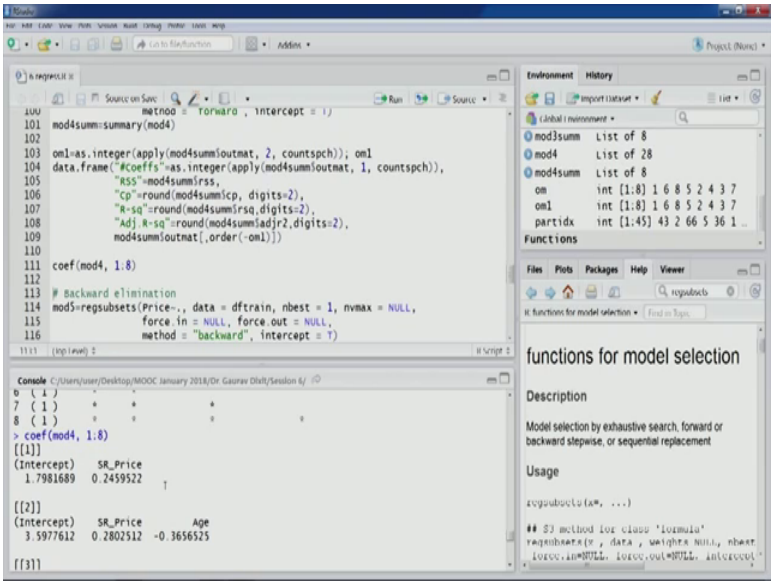
So, if we are interested in understanding the coefficient. So, let us run this particular function coefficient and pricing passing the model argument we will get this.

(Refer Slide Time: 10:19)



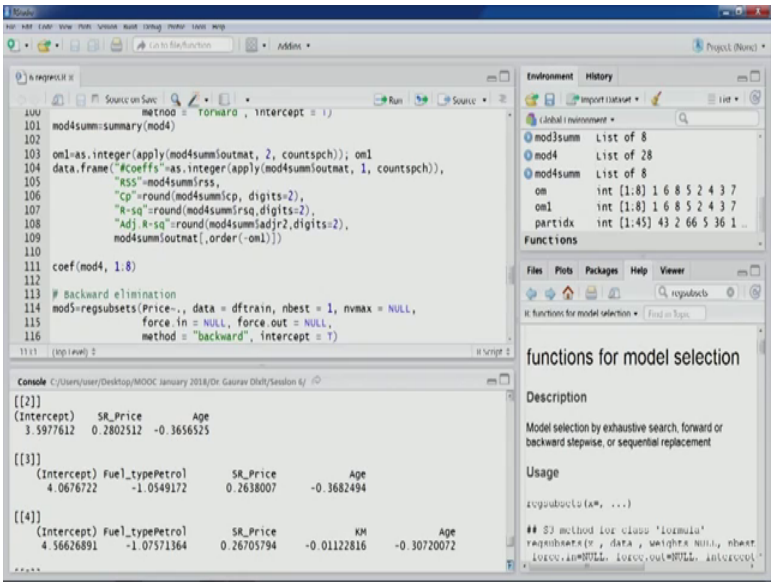
So, you can see the one variable model the coefficient comes out to be 0.24 5 9 5 this is the value.

(Refer Slide Time: 10:25)



So, s R price being the only variable, then we look at the 2-variable model, we have a s R price and age then you can see age is negatively related here.

(Refer Slide Time: 10:39)

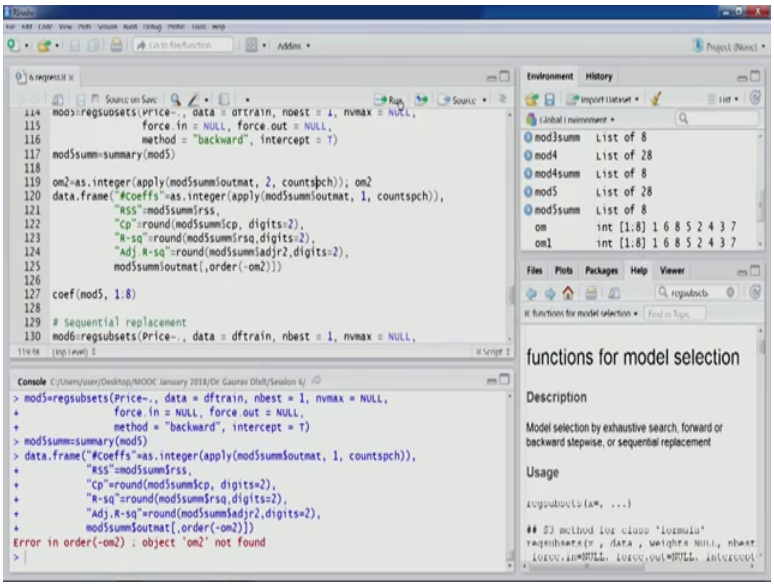


And then fuel type petrol and s R price age. So, just like in the exhaustive search. So, we are getting very same results the same models also. Then after fuel type petrol s R price we see the kilometers appearing in their. So, it is happening exactly like what happened in the exhaustive search right.

However, as I said if we change the sample of or if we change the partitioning, we increase the sample size already in change the partitioning I read on the redo the partitioning then probably the results will change. Now let us come to our next algorithm that is backward elimination. So, again for backward elimination as well we can use the same function less subsets.

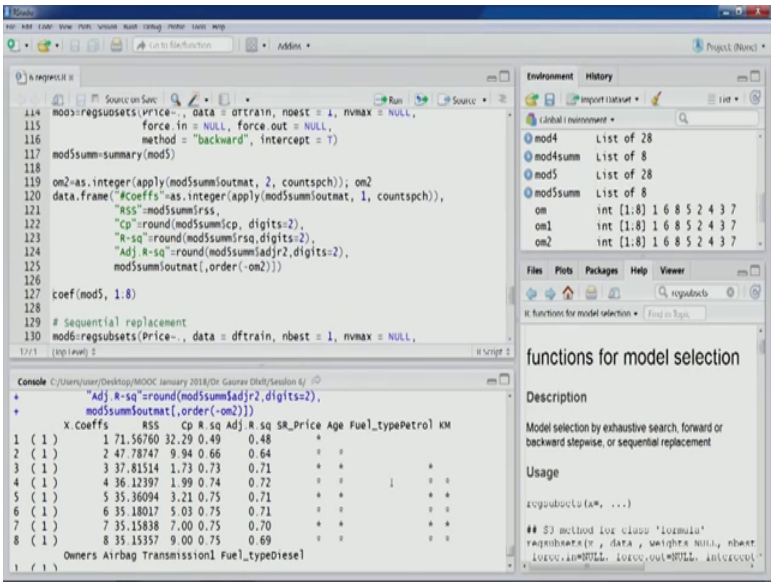
And you can see the one change that we have done here is the in the method argument where we have a specified backward as the algorithm in this case. So, let us execute this code, and let us also generate the summary of this output. Now here again we are going to do the same thing that counting of a special character and the other things remain same. So, let us execute this code ok.

(Refer Slide Time: 12:00)



Oh, we did not create O m 2. So, let us compute this, and then produce this data frame.

(Refer Slide Time: 12:10)

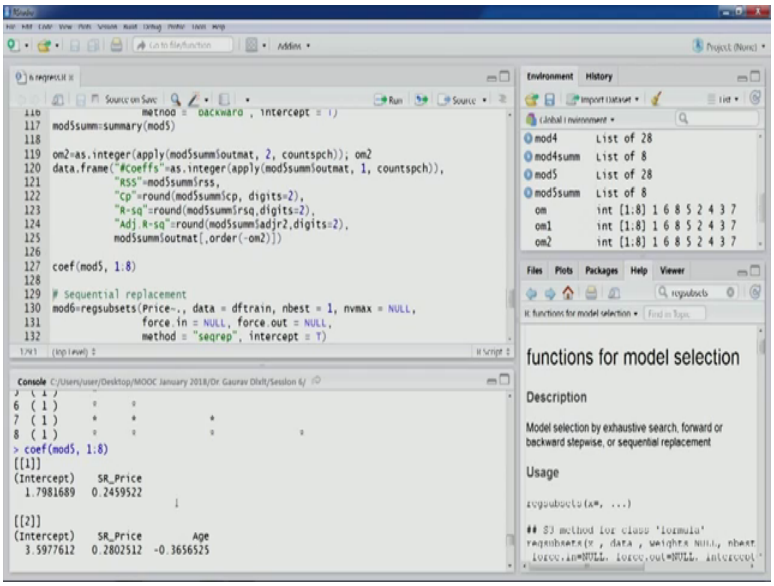


Now, again if we look at the results of backward elimination, then also we would see that same numbers are there again. So, adjusted R square value again starting from the same 0.48 0.64. So, the values these values for all 3 algorithm whether exhaustive or forward selection our backward simulation what we are saying is up to 2 decimal points. So, there might be some difference may be there may be some difference if we look at maybe up to 6 decimal points. So, the adjusted R square column you can see that speaking at 4

variable model 0.72 is the value. similarly, for R square also it is peaking at the you know 5 variable model 0.75 is the value. And the mellow c p is the numbers are same. So, we will have to select again 3 variable model, but the value of c p is one 0.73.

So, in this so, the R s s residual sum of squares number, there also seems to be same. Let us look at the variables you can see again a s R price present in all 8 models, and age is present and the 7 models followed by fuel type petrol present in 6 models, and followed by kilometer k m which is present in 5 models. So, if we have to we want to look at the coefficients for different variable models, we can you do the same using coefficient function.

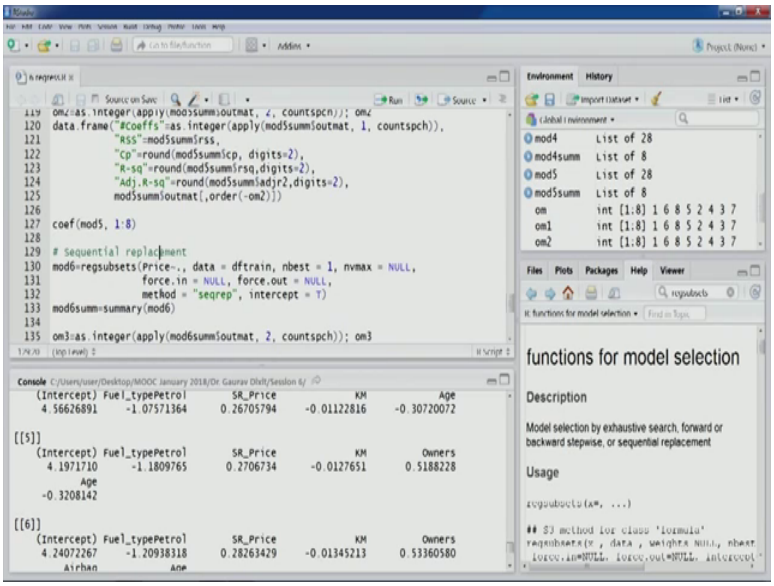
(Refer Slide Time: 13:42)



So, let us look at this again you would see that s R price is present in the first model and then 6 model you can look at the coefficient also age is again negatively correlated as expected in the 2-variable model, then in 3 variable model again we have the fuel entry of fuel type petrol with the negative coefficient negative re regression coefficient. And then you would see in the 4-variable model, we see the k m as well with the negative regression coefficient here.

In the same fashion the results again seemed to be very similar. So, let us move to our next algorithm, that is sequential replacement. So, for sequential replacement there is something that we have not discussed.

(Refer Slide Time: 14:29)

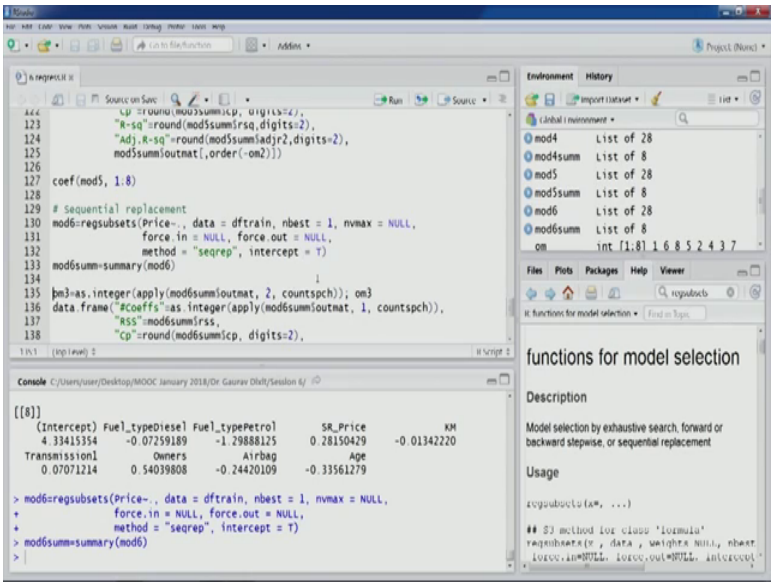


So, what happens sequential replacement is or we start with all the variables, we start with all the variables, and from there then we select you know, and then we identify the best subsets from their best you know subset model that would be there. Then for each of for each of the variable in the in the in that subset for each variable we try a different whether that can be replaced by some other variable right.

So, in that sense we start out, and then these so, in the subset model that is selected for each variable will try out for it is replacement. And then again, we will have for if there are 4 variables in the subset model, and for each variable we will try out different other variables as a as a replacement. And therefore, we will get 4 more models, and out of those 4 models, will again check the, which one is performing better if it happens. Then if then we select the best one, and then we proceed further the same thing is applied again and again till we are not able to find the best model.

So, this is what we call sequential replacement algorithm. And the stepwise regression approach would also be very similar to what is we specified as you can see the replacement.

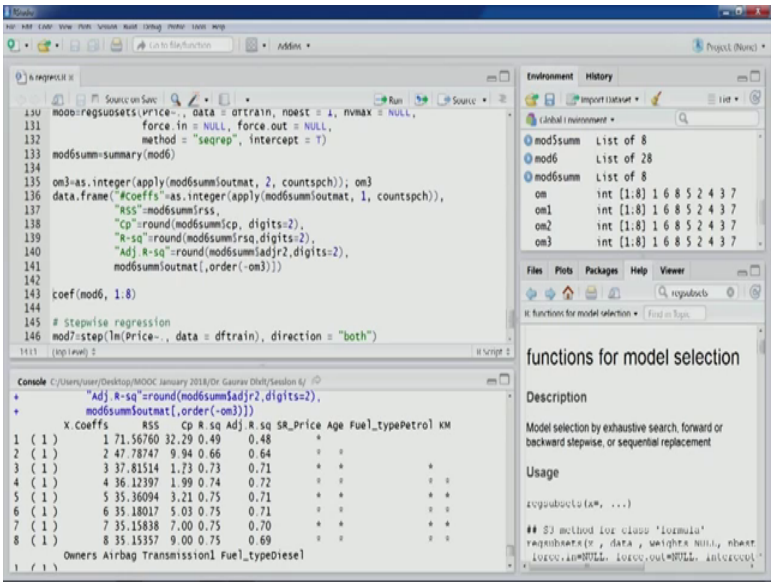
(Refer Slide Time: 15:56)



Some variations might differ, then with the name also, there are going to be very variation depending on the implementation that we follow. So, in this case as you can see we are again using the reg subset function. And the one difference that we can see is the method sequence replacement s e q R e p has been specified. So, let us execute this , it is run this model.

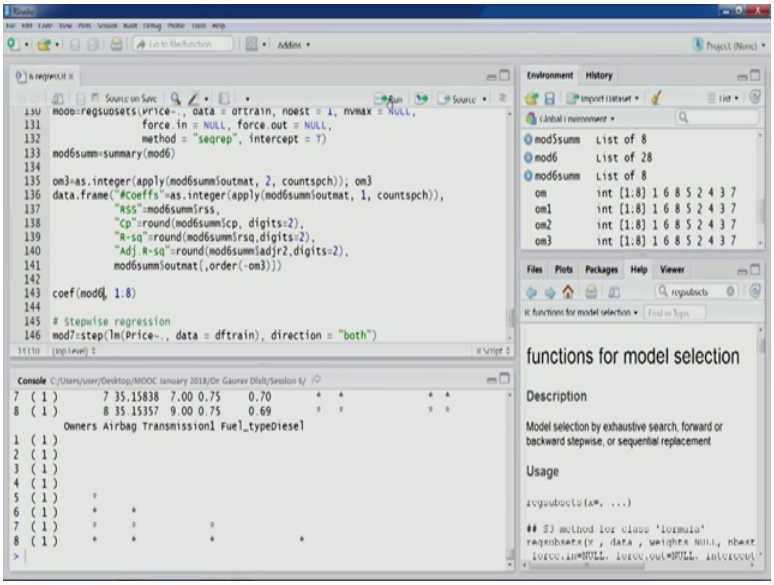
Let us also compute this number of this asterisk in each column of the output matrix. Let us compute this matrix. Now we can see again, in the adjusted R square value again it is speaking at the 4-variable model, and the same numbers are there for R square also speaking at the 5-variable model, and the same numbers are there again for cp also 3 variable is seems to be the more the appropriate one and the same numbers for R s s as well.

(Refer Slide Time: 16:56)



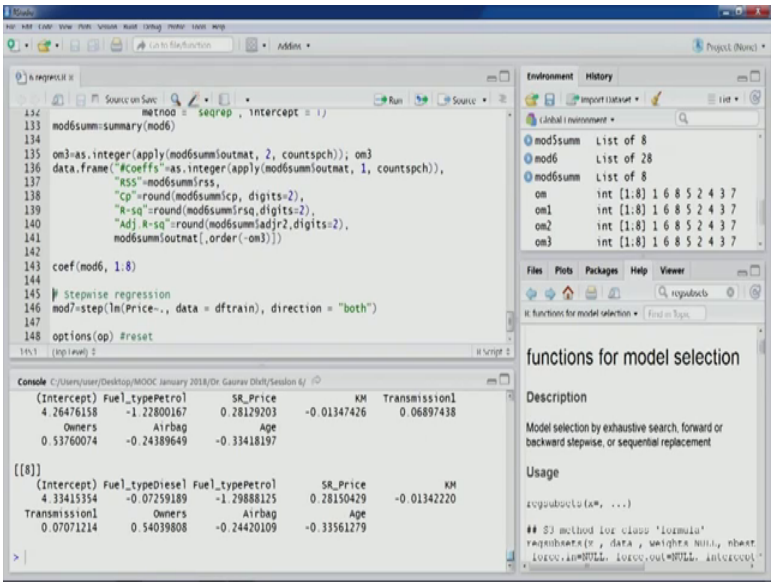
For the variables also, a s R price being presented in all the models then followed by age and then fuel type petrol and came the result seem to be exactly same.

(Refer Slide Time: 17:09)



Now, we are interested in looking at the coefficient value, then we can do.

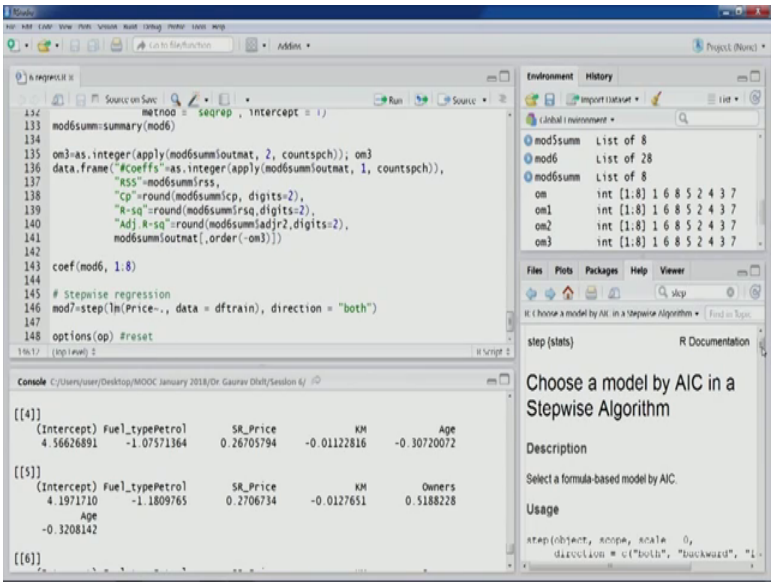
(Refer Slide Time: 17:19)



So, using coefficient function and again as we can see this, again the same results are there is no change s R price then s R price and age, then fuel type petrol s R price and age, and then fuel type petrol s R price and then entry of kilometer k m and the same fashion, the same result we get. Stepwise regression there is another function that is available to us for stepwise func regression that is called step.

So, in this case and this is a step function we have to pass on the again we have to pass on the l m function as the first argument. So, in the l m function we would be specifying the formula as usual in the first argument and then the data. The direction is specified as both so, we can add and remove. So, both the kind of operation can be performed. If you are understand in finding more about this particular function, you can go into the help section type step and choose them.

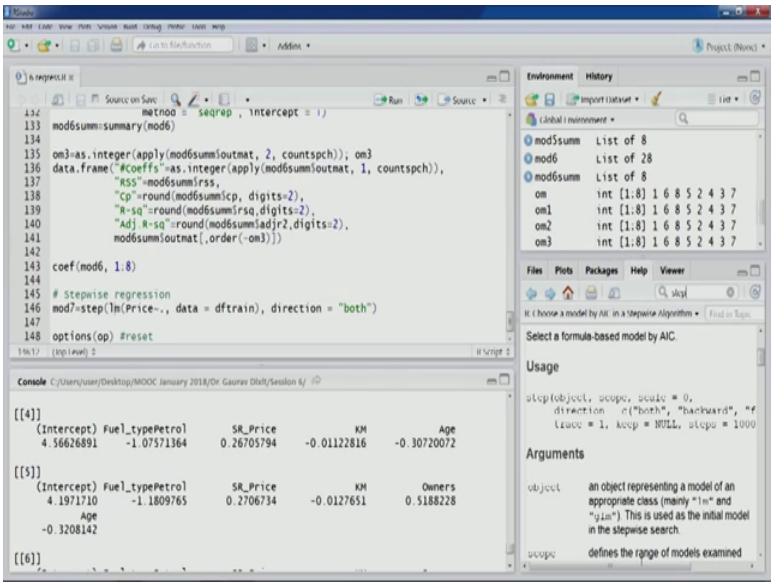
(Refer Slide Time: 18:16)



So, you can see the choose a model a i c in a stepwise algorithm.

So, you can see different options are there this particular method would also allow us to run backward forward and then both.

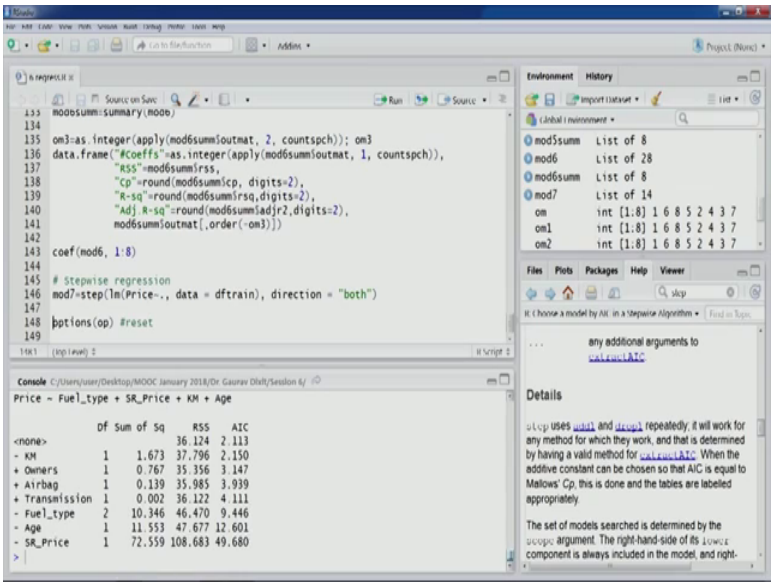
(Refer Slide Time: 18:33)



So, as we discussed that the stepwise regression is about a star just like forward regression, and at each step we can consider dropping off insignificant variables. So, that is why the both indicating meaning the same both indicating the step wise regression. And the other options using the same function we can also, build backward and forward

as well which we had just done using read subsets. So, in this particular function the a i c is used to find out the different subset models. For more information you can always look at the other arguments that are there. So, let us use this function so, let us compute this.

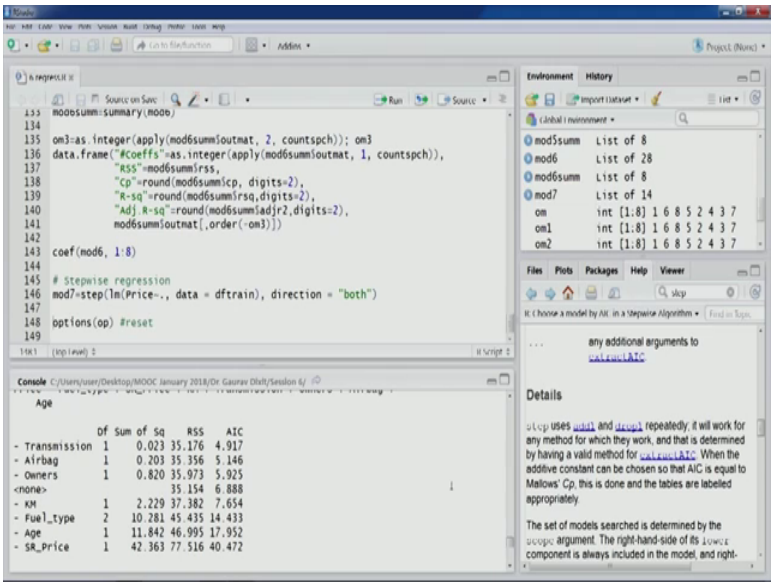
(Refer Slide Time: 19:30)



And this is the results that you get. So, you can see that we start with a i c value up 6.89, and the formula that we start with is price and then other variables, fuel type s R price k m transmission owners and airbag plus age. So, you can see all 7 variables are present, and we start with a i c value of 6.89.

Now if we look at the possible additions, and possible addition, or elimination that that could be there. You can see that if we eliminate transmission, we would gain will reach the a i c level of 4.917, and that would be and that would be a lower a i c value, and then followed by heavy remove airbag.

(Refer Slide Time: 20:23)



Then we will have this 5.146, and then 5.9 to 5, and then the the existing model that is represented by none. So, none means if we do not drop any variable, then this is what we will have, right, all the variables are there. And so, we have 3 candidate models here, at trans when we drop transmission when we drop airbag and when we drop. So, rather so, when we transmission followed by this dropping of airbag and owners, these are the more these are the options alternatives that we have.

So, in the next one if you see that we have fuel type s R price k m plus owners plus airbag and age, we have 3 variables one has been dropped. And you can see the one that has been dropped is the transmission. And that was the first one, right, because we could achieve a lower a i c value 4.917. So, that particular one has been selected and transmission has been dropped.

Now, again from this model also. If we further drop airbag we would achieve a much lower a i c value of 3.147, we drop owners then we reach 23.93 and 9 and then if we do not drop anything then this is the model that we have at present. So, 2 candidate models seem to be performing better than the present model, with respect to the a i c value a i c criteria. Now if we again drop airbag to get the first model. So, you would see , that we reach the a i c value of 3.15, and the variables that we have in the model are fuel type, then s R price then k m then owners and then age. So, if we look at the options that we have is the owners if we drop owners.

Then we will have the a i c value of 2.113 which is less than the value for the current model 3.15. Then the current model is as you can see the 3.147 that is 3.15. Then further we can drop kilometer or we can add airbag. So, this is step we have already followed we will reach to the previous models. So, probably we will select the first one and we will drop the owners.

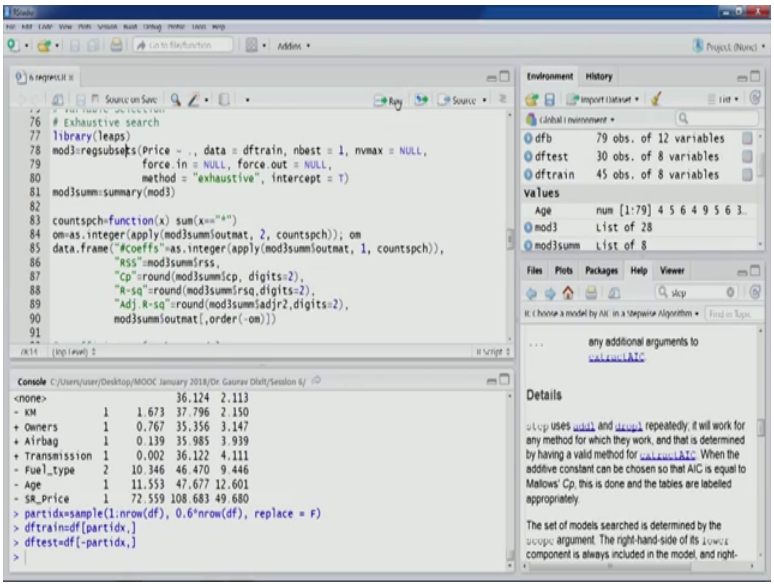
And you will see that reach this step a i c value of 2.11, and now the variables that we have is fuel type plus s R price plus k m plus age. So, these are the 4 variables is superior you remember. The model the final model that we got using reg subsets of function, that we had there also using adjusted R square criteria criterion the final model that we selected was of 4 variable model having the same variable, right, a s R price k m age and fuel type right.

So, in this case also as you can see the first the current model having 2.11, and there is no other model we can see the first row right among the options that we have the first row none that is the same model. So, no other model can improve this further. So, using a different criterion. So, we talked about the mallow c p adjusted R square and R square , using different criteria like a i c we also get the same results and in this case by running stepwise position.

Now as I said that in the in the results that we are getting here, they are they are with respect to the sample that we have very small sample we are dealing with very small samples and the small number of observation and also the partitioning. So, as I said if we change the partitioning, the number of the observation that are randomly selected in the training partition if they change the results might also change. If you want to see what will happen if we change the partitioning, we can repeat few of the models that we have just done. So, let us change the partitioning again, again if we do this partitioning again.

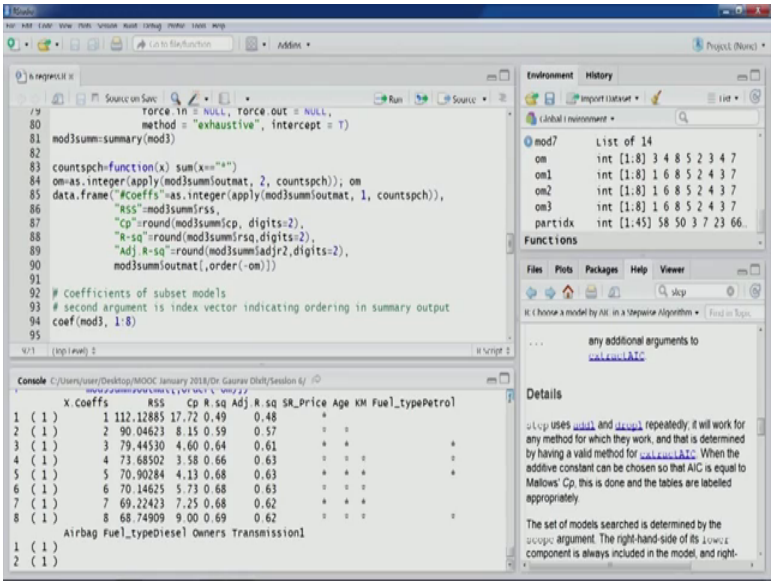
So, we have reek generated these partitions, also as you can see regeneration has happened. Now if we look at the model let us look at the let us go back to the same point variable selection if we look at the exhaustive search. So, once the partitioning has been done.

(Refer Slide Time: 25:22)



We can read on this and you would see that, the results that we might get or might be slightly different. So, you will have to look at this.

(Refer Slide Time: 25:35)



You can see how the numbers have completely changed. At least the numbers of these different criteria have changed you can see different partition you can see adjusted R square now I start with the 0.48, right. The earlier one was different right, we have the previous results as well and the output. So, if we go back you can find out t v s results,

yes, we started with 0.48, then 0.64 then 0.71, 0.64 than 0.71, and then 0.72 now there is just that we have now you can see that start with 0.48, then we these 2.57 than 0.61.

So, we do not reach to that 0.72 level, and you see that 4-variable model we reached the a peak value our gestures by 0.63. And then it remains 0.63 for 5 variable model and 6 variable model, especially if we look at only the 2 decimal point. And then you would see for 7 variable model and 8 variable model the value drops again.

So, we will have to select you know 4 variable model to 6 variable models we had 3 options, in this case you can see just by changing the partitioning, instead of you know using the adjusted R square as the criterion. Now we have to pick from these 3 model instead of just one in the previous scenario previous partitioning that, we did we look at the R square value that is also. There also the results in is remains to be the same that 5 variable model having 0.68 value is going to be selected if we look at the c p value.

So, those numbers have changed significantly. Now we look at c p value 3 variable model the value is 4 points I 6, and then the this value we need to compare is 4, and difference is point 6, right. And we look at the next value. So, it will be compared with 5 and the value is this one 3 point 5 8. So, therefore, difference of for more than 1.

So, probability variable model is again going to be selected here in this case as well, but if we look at the variables now, the column for sr price you would see it is still present. So, it still present and then age is also present and followed by km. Now if you see that km is present in 5 of the models and you see the 5 fuel type of petrol, let us present in the 4 and 4 in the models, if you go back to the previous results that we have had, it was the fuel type of petrol which represent in the 5 you know 6 of the models, and k m was in the 5 of the models right 5 of the 8 models. So, that has changed.

So, fuel the s R price and is they are still present in 8 models and say and models respectively. But k m and fuel type of petrol they have you know they have changed their places right, who k m coming into 5 models and a fuel type of petrol in 4 models. So, you can see that once we change the partitioning the results changed. And this is mainly due to the small sample size that we are dealing with if we had a much larger size probably it would not change the spite partitioning, because we have we would have more number of observations to learn from to build our model from. And therefore, the results are going to be more robust right.

So, with this the same thing you can apply on other algorithm that we have discussed, and that with the change in the partitioning the results would also change. So, with this we will like to stop here.

So, this also concludes our discussion on multiple linear regression all right. So, in the next session will start with k n n.

Thank you.