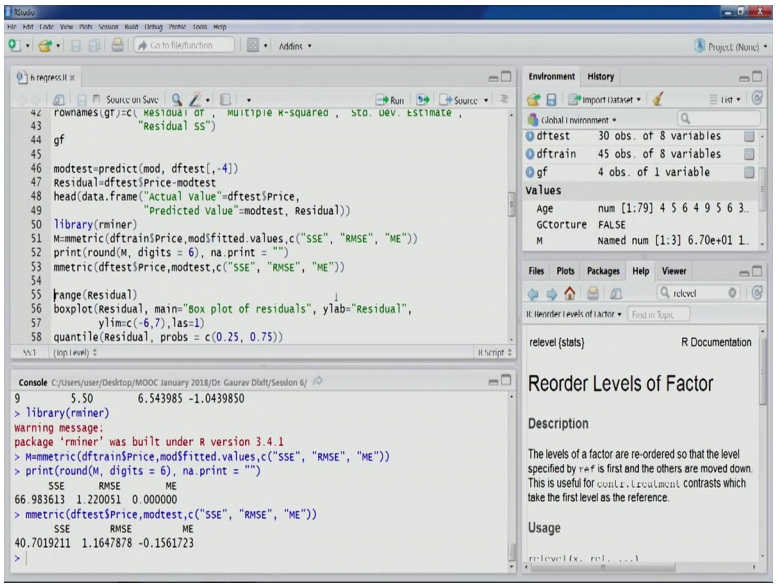
Business Analytics & Data Mining Modeling Using R
Dr. Gaurav Dixit
Department of Management Studies
Indian Institute of Technology, Roorkee

Lecture - 25
Multiple Linear Regression-Parts IV

Welcome to the course business analytics and data mining modeling using R. So, in the previous few lectures we have been discussing multiple linear regression and in the last lecture, we were discussing the regression results that we had produced. So, let us start with the same exercise. So, these are the results that we are discussing in the previous lecture.

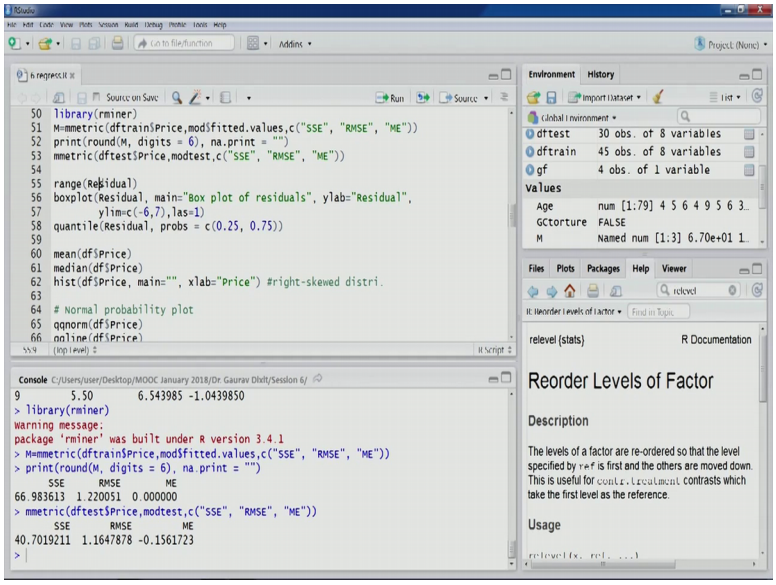
(Refer Slide Time: 00:41)



So, we stopped at this point when we you are comparing the performance of the model that we had built, on training partition and testing partition. So, we were discussing that the importance of the samples size, how where the point to be understood is that, the model that we have just build is performing quite well over the test partition as well; however, this has to be confirmed with the a largest sample size.

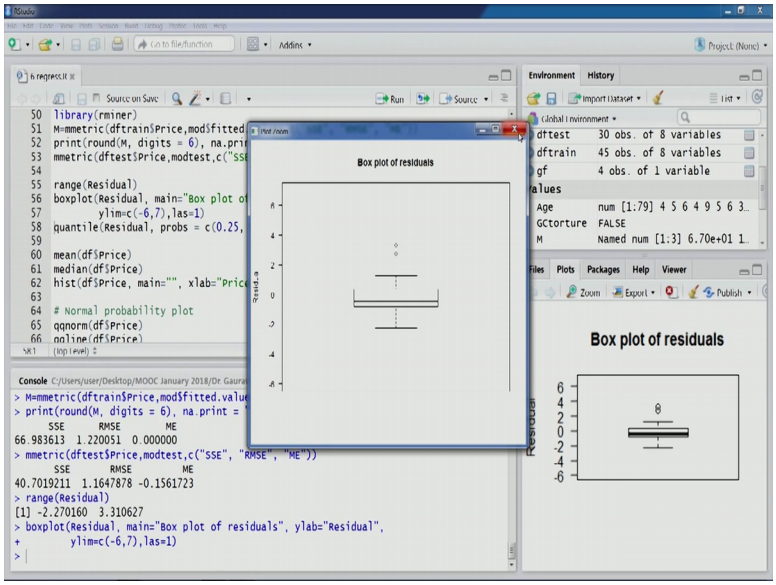
Now, a few more things that we can do about our model and to understand the results of our model for example, the box plot of residuals.

(Refer Slide Time: 01:22)



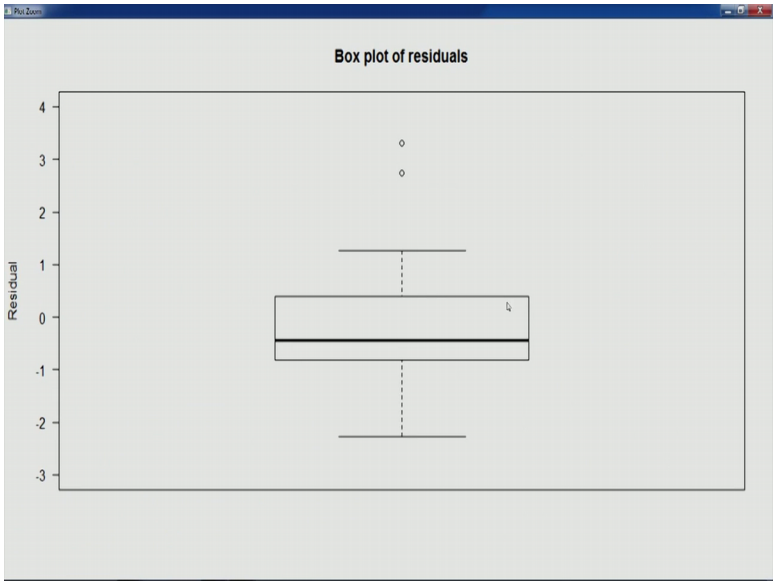
So, box plot of residuals also in a way gives of gives us some useful information about the model. So, let us look at this let us draw, let us create a box plot for residuals. So, range this minus 2.27 to 3.31 and you can see this box plot the first argument is on residuals, which we have already previously computed as we saw in the previous lecture. The titles and the labeling for y axis is also given appropriately, the limit that we already have is the range is going come fall within this is a particle range.

(Refer Slide Time: 02:09)



So, let us execute this line and get the box plot. So, this is the box plot that we have. So, this seems to be a bit more compacted. So, main reason being the range is slightly on the wider side and that is why in in every plot that we have been generating, we always focused on the limits. So, let us seen the limits appropriately let us make it 3, and this one as 4, and then we would have much regular box plot not the compressed one you can see the change in results, and the much better box plot in this case.

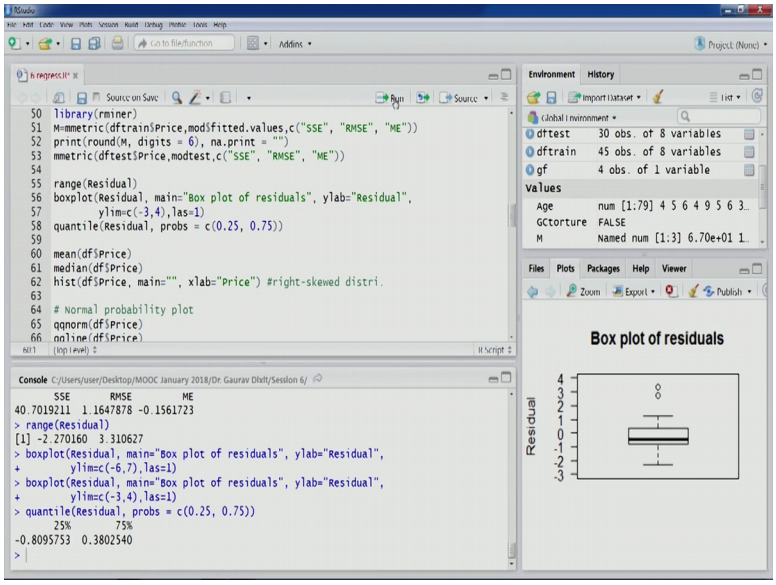
(Refer Slide Time: 02:41)



So, as we have discussed in the when we were discussing visualization techniques that the box plot contains is an some important formation for us, example in this case we can see that a different values the few outliers that are there right and the majority of values which are represented by the first quartile and third quartile, and the range where they are lying right. So, majority of value they seem to be this minus 1 to something more than 0 right. So, this range for box plot and from the look of it looks to be right skewed. So, we can easily find that out there is another important function that we are going to discuss the this is quantile function. So, if you want to compute the values for different quantiles. So, this is how you can do this.

So, for example, we are interested in first quantile and third quartile reason being that the box that we generally generated using box plot, that indicates that the majority of values are lying between first quantile and third quantile. Majority means 50 percent about 50 percent or even more our values are lying in between these 2 plots.

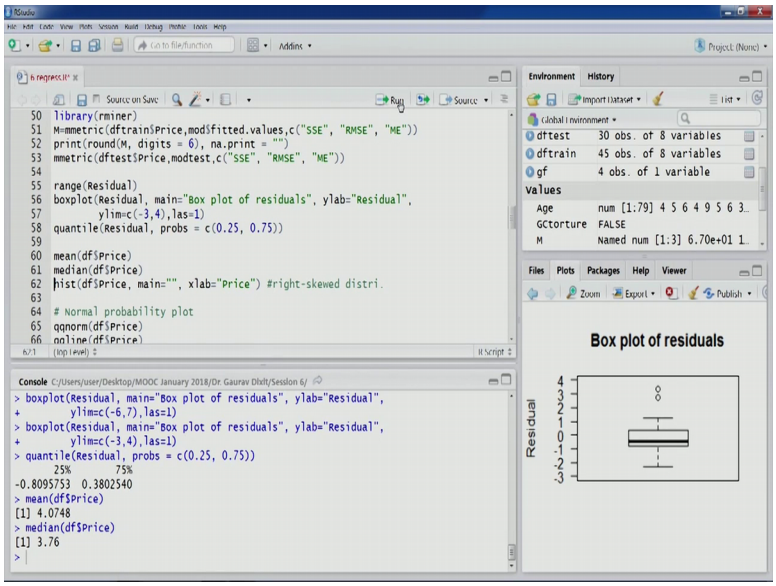
(Refer Slide Time: 04:00)



So, let us compute these quantiles you can see the number. So, if we have to compute the range for these. So, 50 percent values as indicated by these quantiles are lying between minus point approximately between minus 0.8 to 0.4. So, that is the range where more than 50 percent or approximately 50 percent of the values are lying

Now I with for talking about the skewness. So, we can do the same using by computing mean and median value.

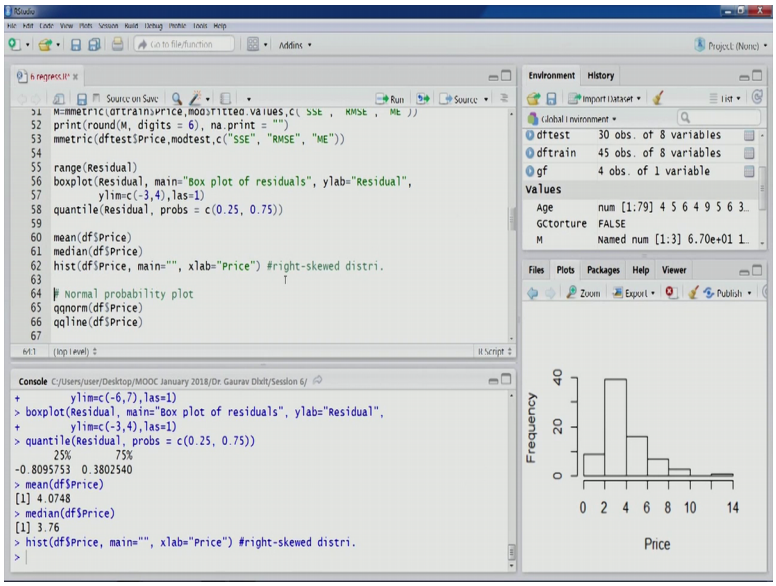
(Refer Slide Time: 04:33)



So, in this case we can see the mean value is slightly on higher side then mid median value and we can also see the some outliers there therefore, this particular you know residuals they seem to be following normal distribution which is right skewed. So, we can check this again by plotting histogram. So, now what we are going to do is we are going to plot histogram on the outcome variable.

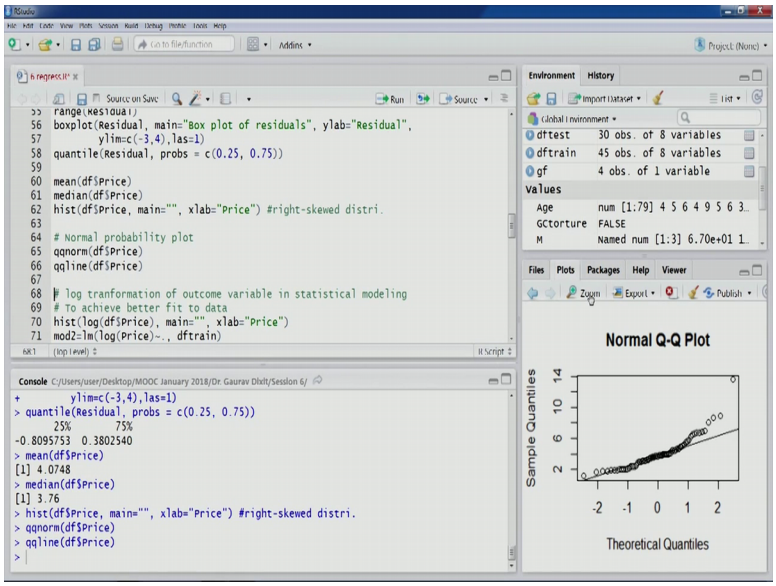
So, this is the price outcome variable is a price. So, let us plot this and you can see this skewness that this seems to be clearly right skewed distribution right normal distribution.

(Refer Slide Time: 05:25)



So, we can again confirm the same thing using few different plots for example, one is one popular plot that is generally use is normal probability plot. So, we have Q Q norm function to generate the same. So, we are going to use this, this is the. So, we are applying Q-Q norm on outcome variable d f price.

(Refer Slide Time: 05:50)



And this is the plot that we get we want to have a reference line on which we can compare whether the plot is whether following normal distribution or not this is the line

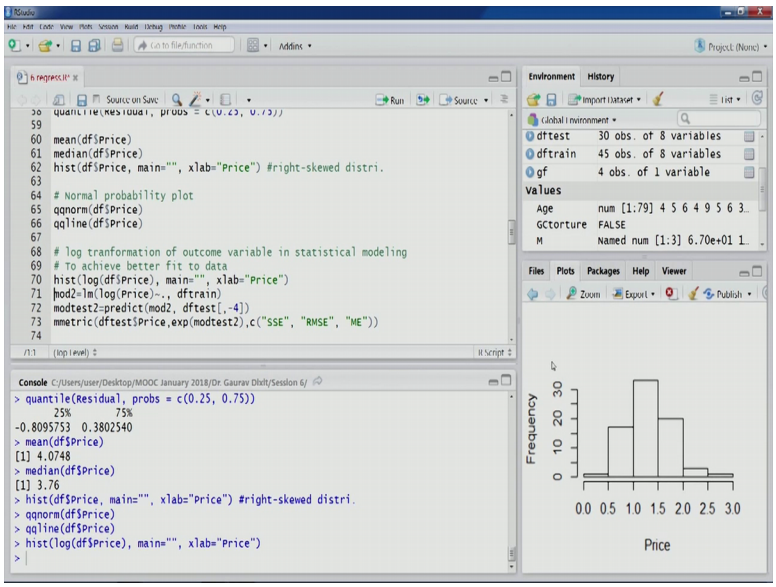
So, in the normal Q-Q plot we see that as we move from left to right on the right part, many observation they start deviating from this reference line that we have just drawn. So, that indicating the right skewness right for a normal distribution these points would be distributed along this reference line; if there is some deviation that tells that indicates the skewness in the plot in the normal distribution. So, as we would have as you see that we are mainly when we talk about the distribution specifically normal distribution, we are mainly interested in residuals and the output outcome variable in this case we have checked both.

Now, generally for our data mining modeling, the outcome variable or the residuals not following the normal distribution is not such a problem because generally the performance that we how we assess the performance is on using different partitions. So, therefore, we do not have to so, we get some relaxation from that assumption that we talked about, but if you are dealing in a few are doing a statistical modeling, in that case the people generally prefer to you know do some transformation. So, that there outcome variable of interest it follows normal distribution.

So, we are going to do the same. So, from the looks of it the we saw that outcome variable it seems to be rightly skewed. So, log transformation would be appropriate for it to make it more normal distribution. You would see that in a log form log curve log function, majority of the value you know they lie in this smaller range and then the there is a long range that is taken on the other values, which are slightly on the higher side. So, therefore, log transformation seems to be most suitable to make it look like more normal distribution.

So, if you plot a histogram we generate histogram on log of d f of dollar price, right then we will get whether will see that whether we getting a normal.

(Refer Slide Time: 08:29)



Now, you see this once we are generated; now they seem to be more of a normal distribution. If we compare this particular plot to previous histogram that we had plotted this seems to be following normal distribution. In the other plot there was a clear skewness right skewness in the plot. Now from the plot what we understand is now we can use this long log transformed value as outcome variable of interest in our regression model.

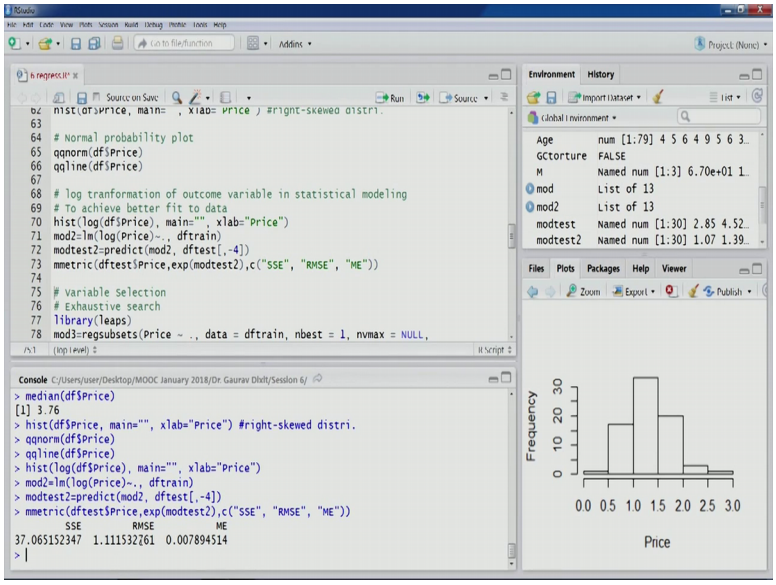
So, let us see this. So, we are again going to run one more models on this is on the transformed value log transform value. So, you can see in the l m function we have log and then we are taking log of price and this is the model is being run on log of price, and the set of predictors on the training partition. So, let us execute this. So, we get the model and now let us also score using this model let us also score the test partition. So, we can see that we are again using the predict function and we are passing on the new model, that is mod 2 in this for this time because you want to you score the test partition using this new model that we have just bulit. So, let us execute this code. So, this has been scored now we are again going to use this m matric function.

So, you can see the first argument being the out the this is for the test partitions first argument being the outcome variable for the test partition coming for the test partition and then you can see the values, that we have just computed mod test 2 we are taking an exponential of that value, because if you remember that we have done log transformation. So, therefore, this course that we would get on test partition, they would be in that scale itself.

So, therefore, we need to bring those value to the normal or regular scale. So, that we can done using applying the exponential function on this score values and therefore, the values would return to the normal scale, and then these values could be used to compute the matrix.

Because idea is to compare the performance of this transfer model with respect to the previous model that we have and that too on test partition. So, that is compute this these values and you can see the number. Now we look at the values you can see that the earlier the results we look at the RMSC value it is 1.16 right for the earlier model and now it is 1.11 right.

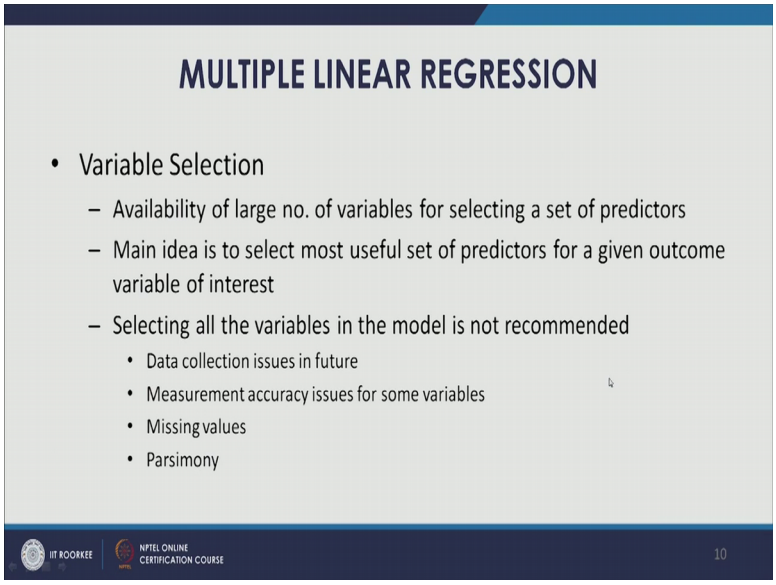
(Refer Slide Time: 11:24)



So, even in this case the error is further reducing. So, this might not always be the case was once you transform the typical scenario could be that the this RMSC value, the this error might slightly increase right, but in this particular instance the error is even when can we run a transform model when we transform about outcome variable and then even in that case the model is doing even better you can they see that even the SSE value is lower than the previous value that we had in this case it is 37.

And the previous value was 40 here. So, in this case the samples is a partition that we have, for after doing log transformation we are getting even better results; however, the word of caution would be that as have said in the previous lecture as well that we are doing this exercise on a small sample size having just 75 observations. So, therefore, the results are not that much reliable. So, therefore, we have to use a larger sample size. So, that the results are comparable in a much broader sense now. So, this was about the regular regression analysis that we have just done.

(Refer Slide Time: 12:56)



Now, let us move to our next discussion that is on variable selection. So, when we were discussing dimension reduction techniques in our previous modules and lectures. So, we talked about the principle component analysis, and how it could be used to reduce the dimensions. We also discuss that some of the data mining techniques right they could also be used to perform the dimension reduction. So, now, we also talked about at that time that the regression analysis could also be used to perform this we also talked about that cart could also used. In this right now we are going to discuss how the regression modeling can it be used to perform variable selection, that in a way also is for dimension reduction.

So, generally when we are dealing with large data sets we also have large number of variables, you know and from those a large number of variables we will have to select the useful predictors for our prediction or classification task. So, how do we do that? Because the main idea for our modeling exercises, to select the most useful set of predictor for a given output outcome variable of interest. So, therefore, from those large number of variables the 20, 50, 100 or even 200 variables that we might have in our data set, how do we identify the most useful set of predictors of 5 6 7 8 to 10 variables for our modeling right.

So, variable selection that using regression models that could be the one alternative solution to perform this. Now at this point I would like to mention that is as explain in the slide also that selecting all the variables and the model is not recommended. So, now, those of the computing softwares that we have that immediately you might be tempted to have all the models in your all the variables in your model.

And then later on selecting and useful a variable from that from those results to that is not recommended for a number of reasons for example, data collection issues in future if you are having many variables, in the model and later on if you are required to rerun the model and because you would like to compare the performance of the model to the previous model as well, but in future you might not be able to collect the data on all the variables.

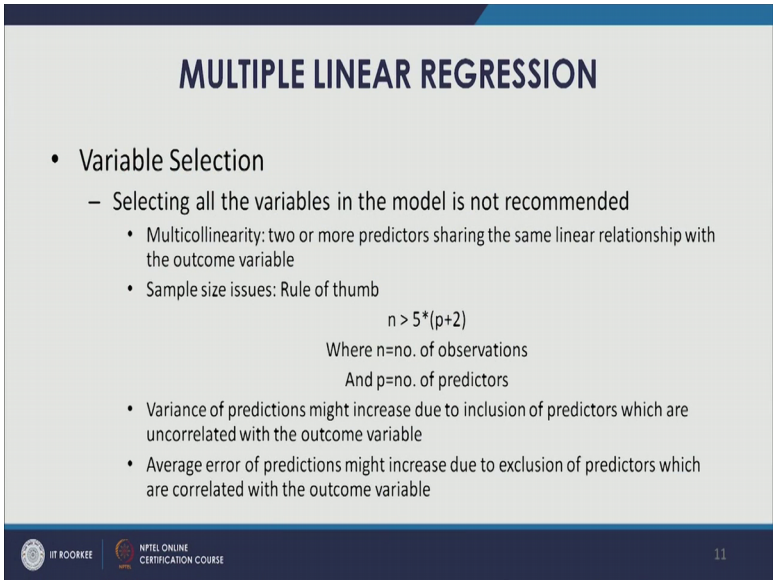
So, as you increase the number of available in your model, the collection data collection issues in future might hamper you know comparison the analysis. So, therefore, you have to beware in this thing in mind that there could be data collection issues, if you build your model using all the variables that are available. Now measurement accuracy is use for some variable. Now some of the variables because generally when we talked about the statistical modeling, generally we are dealing with primary data as we discussed in the previous lectures, some of this data some of these variables are measured use using the survey instrument.

Now, there are generally many of these variables of perceptual variables also there are generally the measurement issues measurement and the exercises about the data that we collect. So, because of that it is not recommended that whatever data that you have it is not recommended that you include all the variable all the variables in the model, you might choose to eliminate some of the variables which might be having some accuracy related problems measurement accuracy related problems. Now missing values if you are having more variables in your model, missing values could complicate the problem much more. If there are more variables in the model having missing value in even just 1 cell might lead to eliminating or removing more number of records right. So, therefore, if you have more variables in the model there are more chances of having missing values in the data set and then even more chances to remove you know more number of records, more number of observations or rows from your data set because of the missing values or depending on the impute imputation that you have.

So, usefulness of your model can also be under question. Now the other reason is parsimony. So, as we have discussed before the would like to follow the principle of parsimony. So, therefore, it is always recommended that we try to build a model with as few as predictors and is still to be able to explain the most of the vary variability in the outcome variable interest that is the ideal scenario that we want.

So, with respect to this principle of parsimony also we should be building our model on pure variables, and model building on all the variables is not recommended. Now few more reasons then other one being multicollinearity.

(Refer Slide Time: 18:26)



So, beta is the form multicollinearity in previous lectures as well, and I will do the same again in coming lectures as well. So, what is a multicollinearity. So, this is briefly this is about 2 or more predictors sharing the same linear relationship with the outcome variable of interest right. So, as we have talked about that one of the assumption in regression analysis is, that cases should be independent right.

Otherwise that is going to do you know produce might produce the or might affect the estimates of reliability of the regression coefficient those estimates right.

So, if the cases would be independent the same applies on the variables also the column side. So, on the row side also the cases would be independent, on the column side that is the variables that we are talking about we say that multicollinearity should not be there this is even more applicable to regression analysis and other statistical techniques. So, many predictors having the same relationship with outcome variables in a way is very similar to having you know 2 dependent rules. So, we do not want that in our data. So, we would like to eliminate the multicollinearity issues.

If there are 2 predictors which are having the same relationship with the outcome variable, and we include both of them in the model, the result that we might get would be dominated by the information that is there in this 2 variables and therefore, our model would rendered useless. So, therefore, we would like to eliminate the multicollinearity issues and we are including more variables in our model, there are more chances for multicollinearity to appear in our model there are more available. So, of course, some of few variables might be highly correlated there are good chances of few variables, being highly correlated and therefore, multicollinearity could be there in the model.

So, therefore, for this reason also it is recommended that we should do our modeling with fewer variables. Now let us move to the next point that is sample size issues. So, few rule of thumb for sample size is we have been discussing before for example, this one particular is that number observation that we have should be greater than 5 times of you know number of predictors plus 2 value right. So, if there are you know if there are p predictors. So, we would like to have number of observation more than 5 into p plus 2 right. So, if we have a the main logic being if we have more number of predictor our requirement for having a number of observation will also go up.

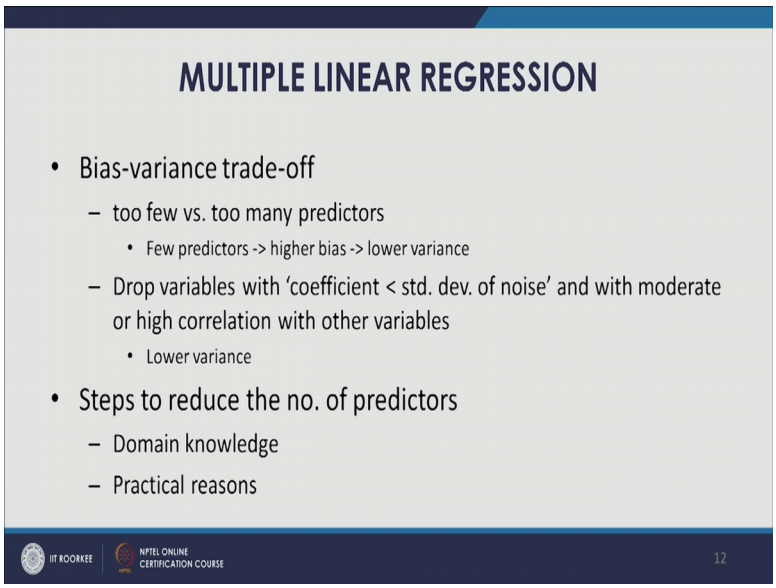
So, we have a 100 100 variables in our model, and you want to include all of them in our model. So, that would increase the number predictors that is p value. So, therefore, our requirement for the number of observations could also will would also go up. So, sample size issues could also be encountered, if we have more predictors in our model. Now few other things would also be there for example, variance of predict predictions that we do or after modeling might increase, due to inclusion of predictors which are uncorrelated with the outcome variable. So, we are having more number of variables in the model. So, there are chances there are more chances for including some uncorrelated variables.

Some predictor which are uncorrelated with the outcome variable and therefore, they will increase the variance of prediction. So, that is avoidable now another issues that me my face is the average error of prediction. So, average error of prediction might increase if we exclude some of the predictors which are correlated with the outcome variable. So, as I would said for us when we do any kind of analytics whether it is based on data mining on whether it is based on statistical techniques, we are always looking to build model we are always looking to predict values or classes in different task that we do. So, in all those situations if there are variables that we are analyzing, if they are correlated only then we would be able to do our job.

But as we discussed if the variables of highly correlated then of course, you would like to avoid that situation. Similarly if the variables are not correlated at all we would like to avoid that situation always that situation as well. So, therefore, we are interested in the mid-range, but the variables are slightly you know variables are slightly correlated less than moderately correlated and therefore, the our average error of prediction that is on that could be kept on check. So, if we in exclude predictors which are correlated with outcome variables, our average error of prediction will go up. So, therefore, we have to see on the we have to balance on both sides is would not like to have highly correlated variables because that would again those variables my dominate the results as we discussed.

Multicollinearity issues similarly if there are uncorrelated variables, they will increase the variance which is not desirable. So, we would not like to have those uncorrelated variables as well. So, we would like to have and you we also have to make sure that the variables which are correlated I know that are in the mid-range low or mid-range having low or mid-range correlation we would like to have those predicted in our model.

(Refer Slide Time: 24:19)



This brings us to another concept related to the same discussion that is bias variance trade off.

So, bias variance trade off when we is important especially when you try to include you know many variables, then what is going to happen is your variance will be a negatively impacted, when you have a low number of variables your bias that is the average error that is going to be negatively impacted. So, we have to balance between bias and variance. So, for example, so, this scenario sometimes is also referred as too few verses too many predictor.

Whether we should have too few all too many predicts both the suggestion are avoid avoided should avoided we should have a balance approach, balance bias variance trade off and balance has to be maintained. So, if you have few predictors that is going to be leading to higher bias, that is at we discuss higher average error and therefore, lower very variance. So, the lower variance is desirable, but then you would also do not know would not like to have the higher bias. So, therefore, that is the tradeoff that we need to do that we need to perform.

So, what could we done. So, we can drop variables with coefficient with are less than standard deviation of noise, and with moderate or high correlation with other variables. So, as we discussed that if the other variables that our predictors they are moderately are highly correlated with the other variables right, and they are also having coefficient which is less than standard deviation of noise, then probably they are good candidates for dropping. So, as we have discussed when we were discussing regression results. So, some of the variable which are having coefficient value less than standard deviation of noise, and also being correlated highly correlated or moderate highly correlated with the other variables. So, they could be dropped and if this is done, then will achieve lower variance which is desirable.

So, sometimes we accept even a bit more bias to achieve lower variance. Variance is more desirable for our modeling exercises. So, sometimes we accept even more bias to achieve lower variance therefore, most of the recommendations that you would see they are generally given to it is variance achieving lower variance. Now this brings to our next discussion that is steps to reduce the number of predictors that we have been discussing. So, what could be the steps that we can take? So, domain knowledge is one. So, using your domain knowledge having been having you know done some work on the same area, having some knowledge about the different relationship, different phenomenal constructs variables that will give you that will put give you some more expertise to find out the predictors, which are more which are more sensible with respect to the task that we had that is prediction of classification or even the statistical modeling.

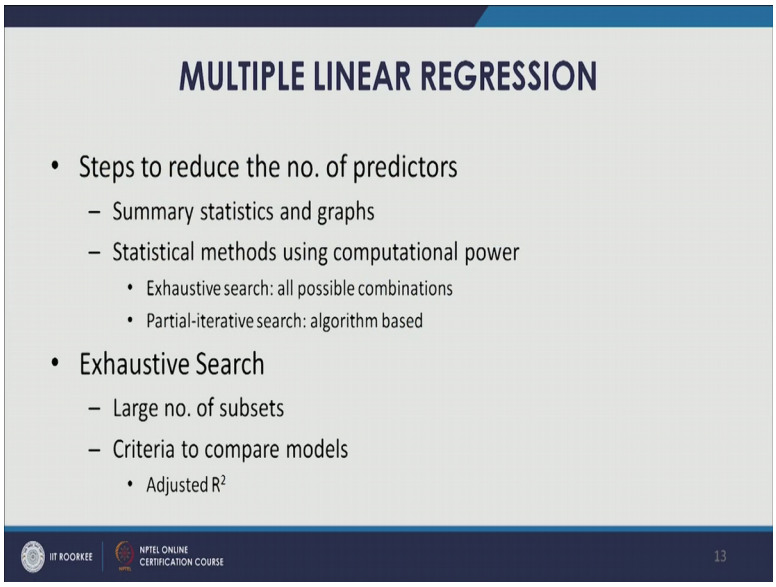
So, domain knowledge is always going to help you in terms of identifying which variables are going to be useful for the modeling exercise that you are performing. So, the first reduction you can do based on your domain knowledge. For practical reasons would be the another approach. So, most the discussion that we did just before all of them would be applicable on the practical reasons.

So, discussion in the previous slide that we had done for variable selection why we should have few variables, all these points they would also be they will also give us some direction from practical reasons that we talked about for example, the first one that we discussed that data collection issues and missing values right. So, all those all those point gives give us provide a some practical reasons to avoid to reduce the number of predictors.

Few other things that we have discussed in previous modules also summary statistics and graph that we have done in our visualization techniques lectures right, this could also be done. So, in this particular topic multiple linear regression our focus is on statistical methods, and to also explain the computational power that we have nowadays. So, that 2 approaches are common or popular exhaustive search that is to search all possible combination of predictors and find find out the best subset which is fitting the data. So, that is one approach this is more like brute force approach.

The first one is exhaustive search is like brute force approach, where we are checking all possible combination of predictors and you would like to identify the subset of a subset of those predictors, which is giving the best fit to the data. Now the second approaches partial iterative search.

(Refer Slide Time: 29:36)



This is more of algorithm driven, your optimized approaches right. So, we are going to discuss these 2 approaches in coming lectures. So, first let us start with the exhaustive search and few points and then will continue on the same in the next lecture say exhaustive search large number of subsets we are going to examine.

Because we are going to check all possible combinations criteria that we generally used to compare models, different subsets and comparison between those model subset models would be generally be based on adjusted r square; and r square and value CP that we are going to discuss in coming lectures. So, will stop at this point and our next discussion is going to be around the exhaustive search, and the criteria that that are used to perform that, and then will also do an exercise in R studio to understand how this is done. We will stop here.

Thank you.