Business Analytics & Data Mining Modeling Using R

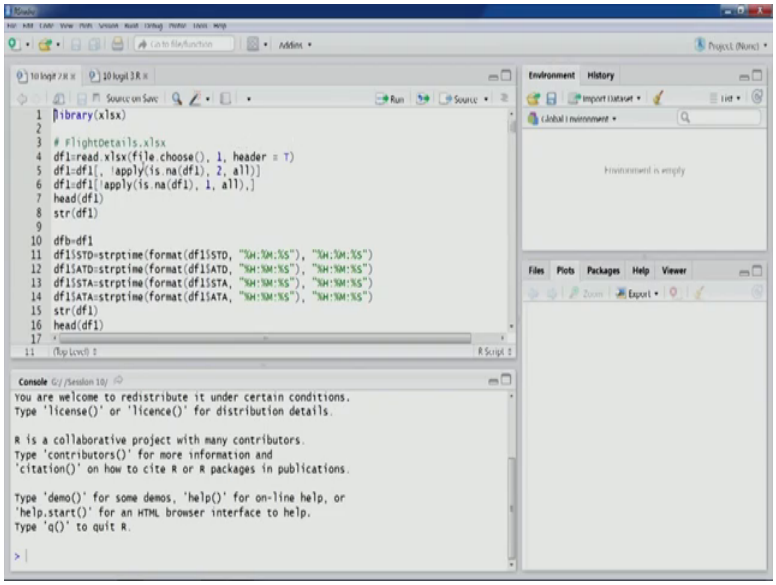
Dr. Gaurav Dixit
Department of Management Studies
Indian Institute of Technology, Roorkee

Lecture - 52
Logistic Regression-Part VII

Welcome to the course Business Analytics and Data Mining Modelling Using R. So, in previous few lectures we have been discussing different aspects of logistic regression so we will continue that discussion in this particular lecture as well.

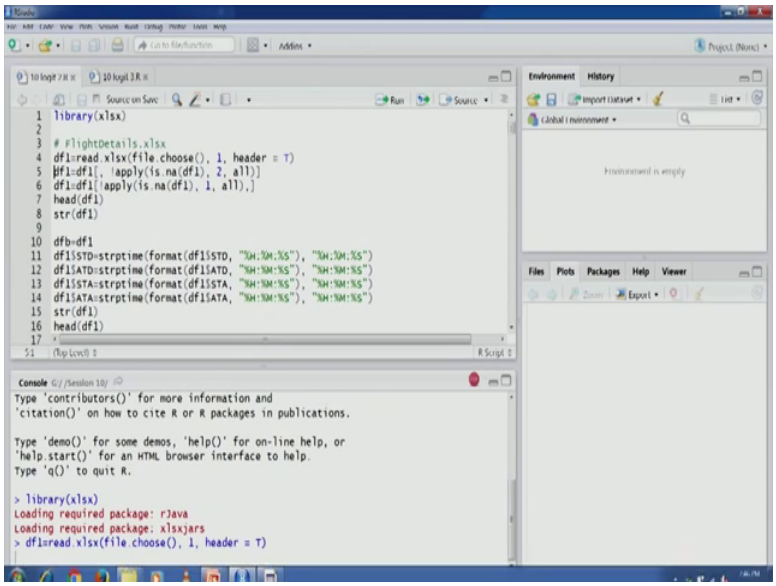
So, in previous lecture we will be used apply details data set and then promotional offers data set as well. Some of the details regarding modelling exercise we could not cover.

(Refer Slide Time: 00:44)



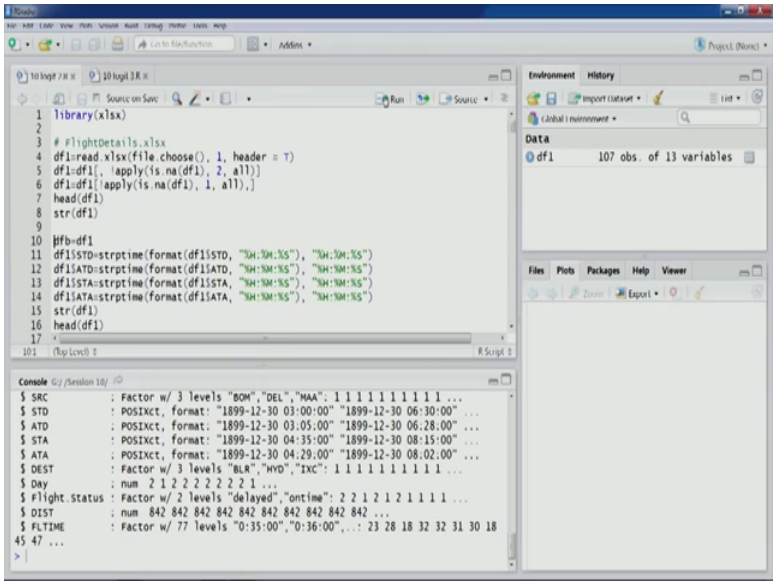
So, we will do that and discuss. So, let us move to let us import this data set. So, we will let us import the library xlsx. So, we are again going to use this particular data set flight details.

(Refer Slide Time: 01:00)



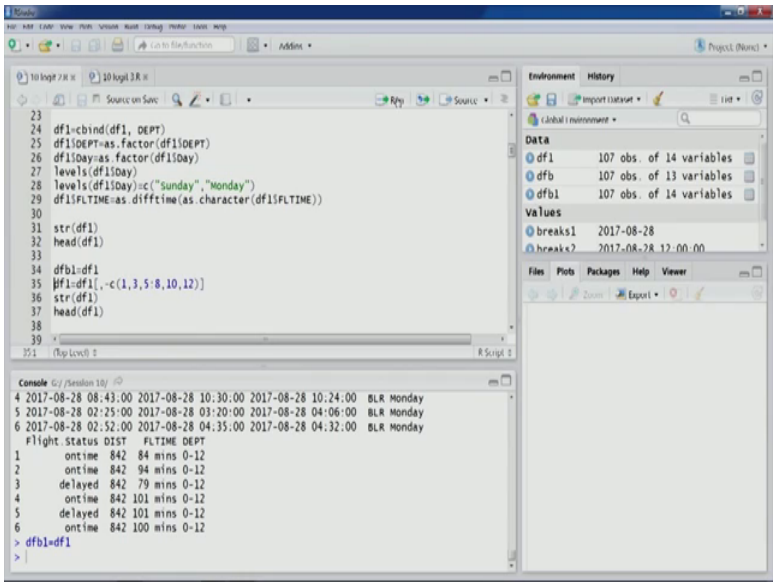
So, let us import this. Now, let us remove NA columns NA rows, let us look at these are the observations.

(Refer Slide Time: 01:18)



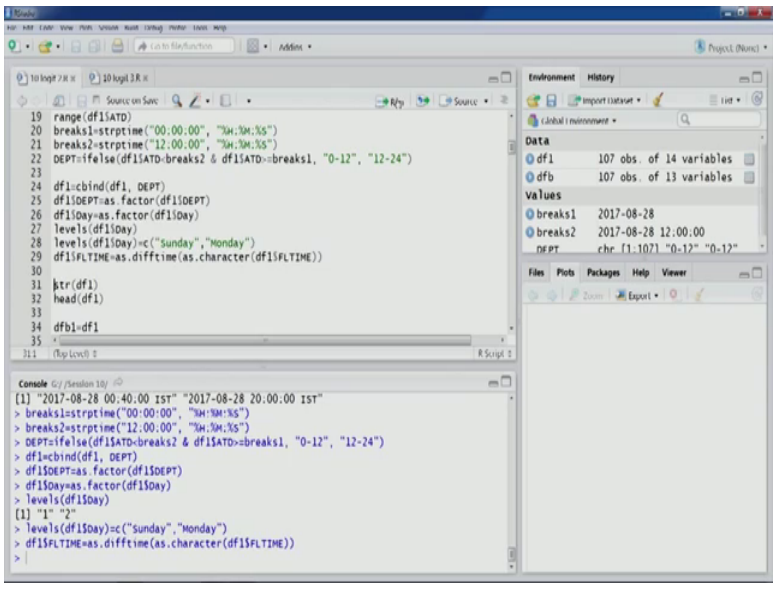
Let us the structure for the data frame. So, we will follow some of the steps that we have gone through in previous lectures as well, so we will just go through them.

(Refer Slide Time: 01:34)



Now, there is this particular exercise in the previous lecture we had used a separate grouping for departure time. So, again I have done certain changes into this grouping also, but this is not very important; however, let us run the model with a new grouping for departure time interval. So, this is the range we already familiar with.

(Refer Slide Time: 02:00)



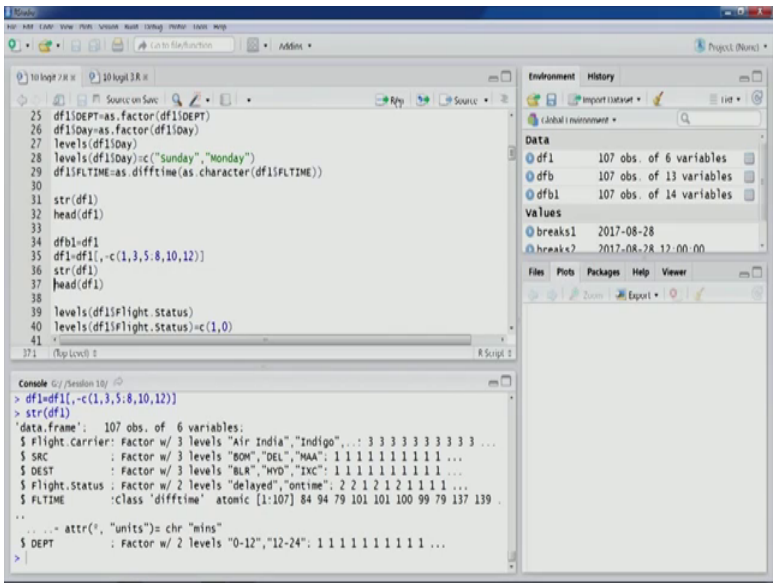
So, break is now 0 and 12, so these are the 2 breaks, 0 hours and 12 hours. Now this is how we are creating depart time variable. So, if less than a breaks 2, breaks 2, and breaks

1, then 0 to 12 so; that means, within if the timing is within these two you know hours 0 hours, and 12 hours.

Then first category that is 0 to 12 otherwise it is going to be the second category that is 12 to 14. So, let us create this variable let us append this to the data frame let us change it to a factor variable day variable as well let us cut the labels, change the labels flight time.

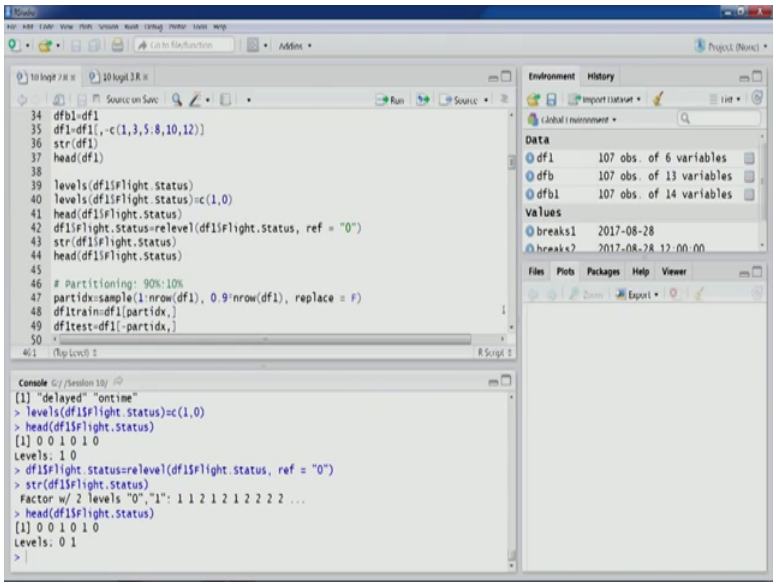
Let us also change it to appropriate format, now this is the structure that we have, now after taking backup we would not like to you know take forward some of the variables so let us get rid of them. Let us look at the structure again now these are the variables, now these are first few values for 6 values you can see everything is ok.

(Refer Slide Time: 03:04)



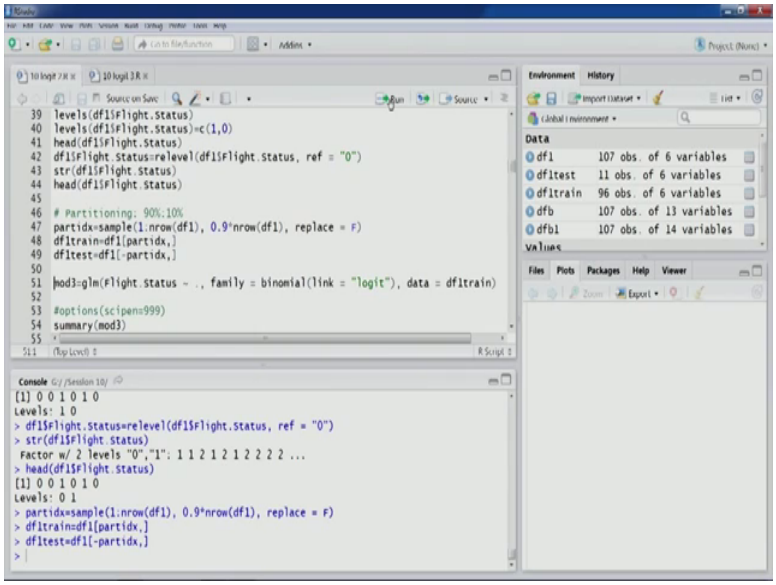
Now, let us work on the outcome variable like we have been doing in previous lectures. So, let us change it to numeric code, let us change the reference category, now this is what it becomes now this is ok.

(Refer Slide Time: 03:30)



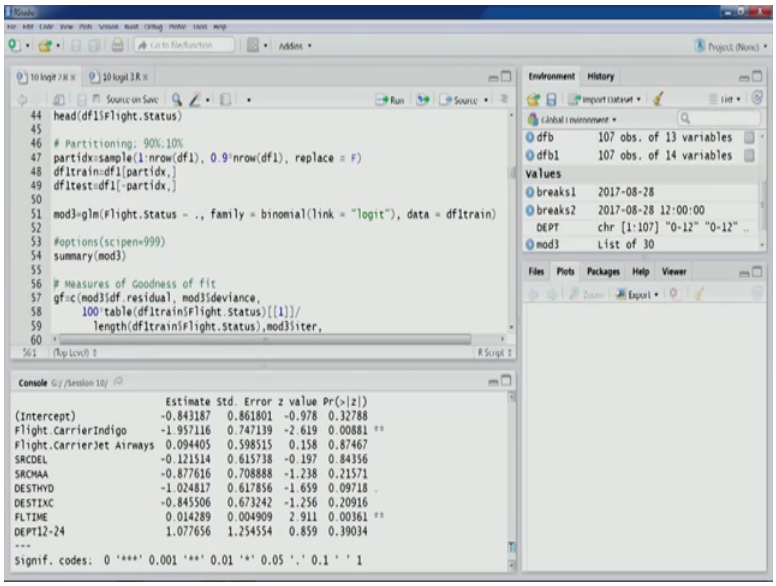
Now, we can move ahead let us do our partitioning 2 partitions, training partition, testing partition and 90 percent for training partition and 10 of observation for test partition.

(Refer Slide Time: 03:40)



Now, the same function glm that can be used to again model this. So, these are the results.

(Refer Slide Time: 03:53)

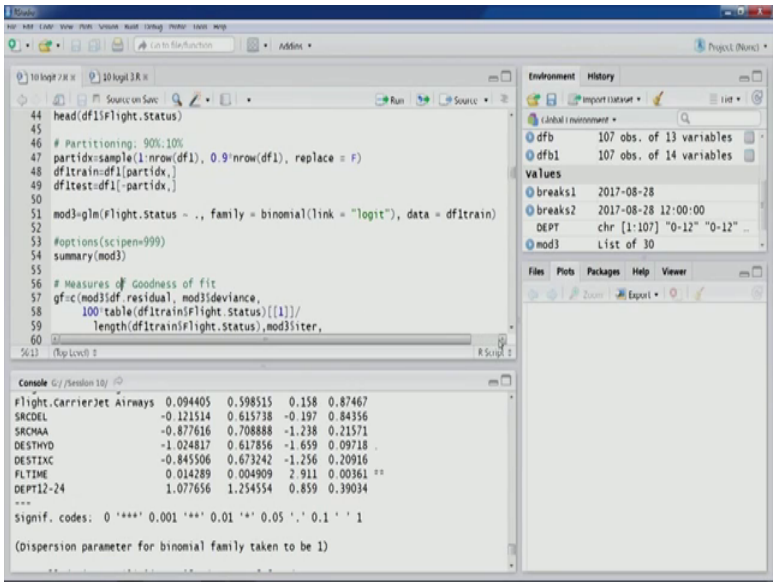


Let us look at once again now you would see every time like we have been running this particular model on the same data set and every time. The significance levels have been changing that as I have been explaining smaller data set and therefore, subject to change in terms of as we as the observation that are part of training data set training partition this results will also slightly change and mainly with respect to significance level.

Now, you can see again flight carrier indigo has become significant to star level right and we can also see that destination has also become significant right. And then we can also see flight time has also you know modes you know higher level of significance.

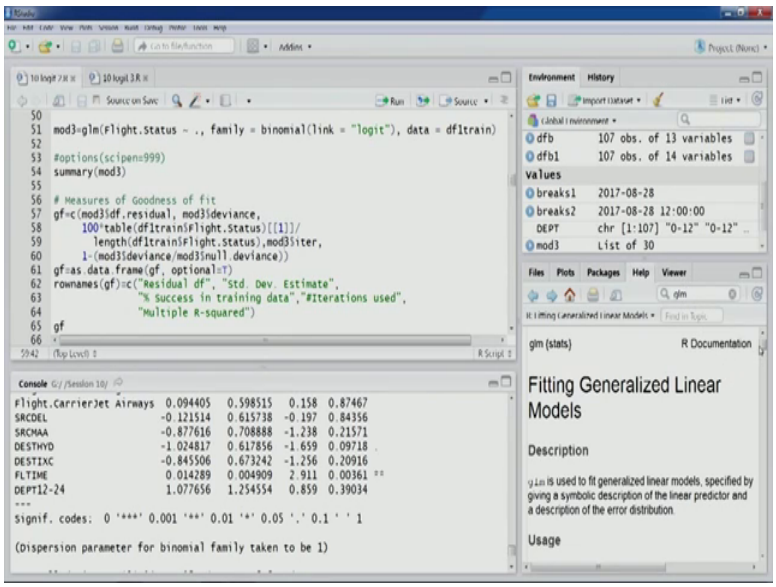
However, more important thing is look at the p varies we can see that this one source for madras as well this is also smaller p values this is anyway significant flight carrier and destination anyway this is significant at ninety percent this is also significant. However, the new grouping that we have created out of department departure time intervals that also not comes out to be significant; however, p value is now smaller.

(Refer Slide Time: 05:14)



So, with this we will discuss the important aspect of logistic regression so that is the majors of goodness of fit.

(Refer Slide Time: 05:19)



So, just like the linear regression multiple linear regression logistic regression also is a statistical technique primarily. And therefore, in a statistical modelling the main objective is to fit to data as we have talked about this aspect many times before. So, therefore, in multiple linear regression we have a particular metric called a multiple R square and adjusted R square which are used to assess how good the model is fitting to the data.

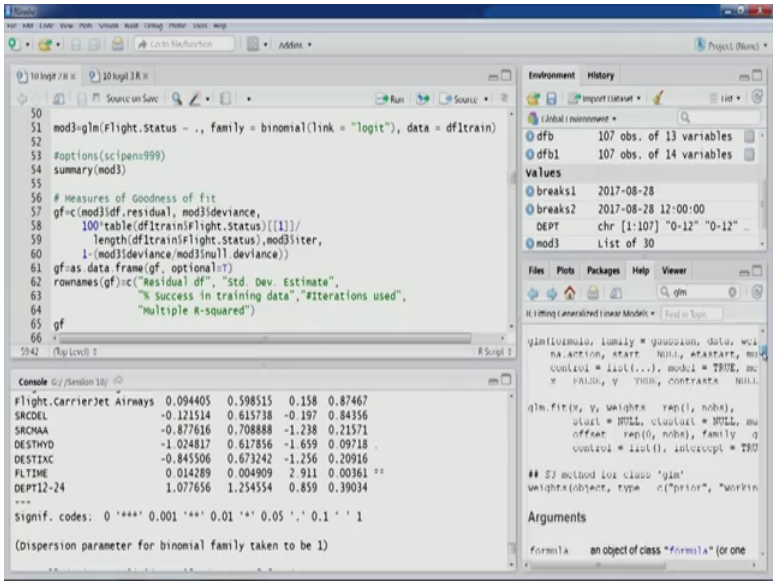
So, similarly since logistic regression is also a statistical technique so there also we are required to have matrix for good to measure the goodness of it how well model is fitting to the data. So, what are those matrix so because there is certain key differences in logistic regression and linear regression. So, we will talk about some of those matrix now.

So, you can see in the code that I have created a vector here gf first one is mod 3 the model that we have just computed and you can see degree of residual degree of freedom. So, this df dot residual is one of the value that is returned by and the glm function and gives us the residual degree of freedom then we have deviance.

So, this is again the returns the deviance value returned by the glm function right then few other things which are mainly for the descriptive purposes. For example, this table result which is for the outcome variable here and then divided by the full number observation that will give us percentage success in training data and then we have iterations.

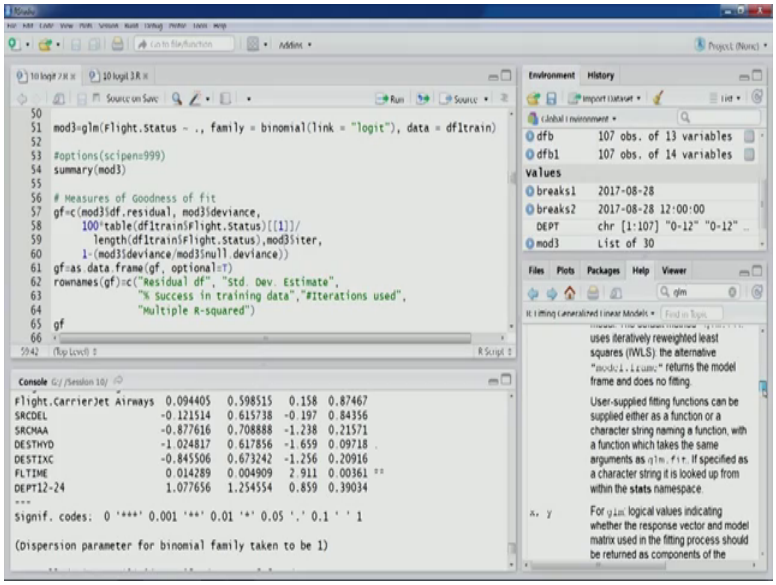
So, as we talked about the particular estimation technique that is used in that is used in logistic regression is different from multiple linear regression we can look at for more details we can look at here. So, we talked about that Emily is used for typically for used for logistic regression; however, you can see that for example, we have been using glm function, and within that if we go we look up for the some of the arguments.

(Refer Slide Time: 07:48)



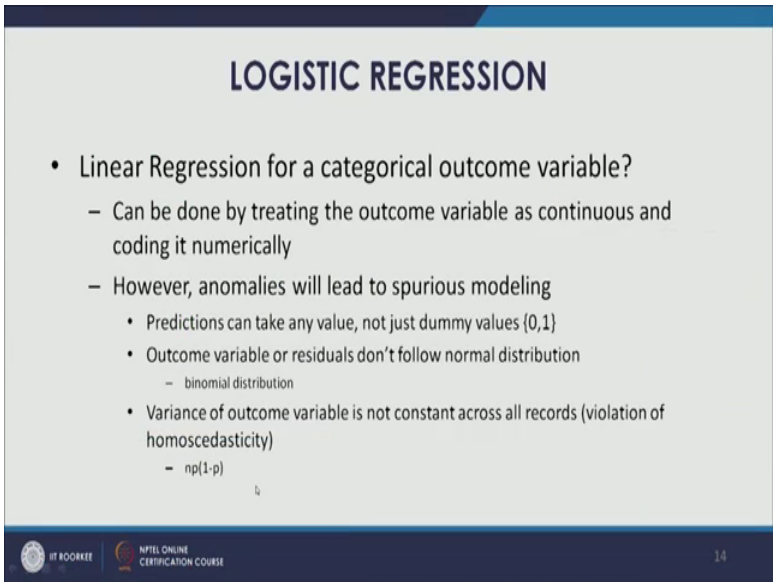
Specifically for this purpose the estimation technique purpose will get more detail. So, we will see that glm for dot fit.

(Refer Slide Time: 08:00)



We will see that glm dot fit method. So, this we can see iteratively re weighted least square is used the particular function that we are using glm iteratively re weighted least square function is used which is quite similar in approach with respect to Emily estimation technique that we talked about.

(Refer Slide Time: 08:41)



So, MLM techniques, sorry MLM maximum likelihood method that we talked about in our discussion as we can look in these slides as well maximum likelihood method MLM method that we talked about. So, this is quite similar to what this the discussion that we have.

So, in particular R's implementation glm function that we are using it is the iteratively re weighted least square that estimation technique that is used to estimate the coefficients that we have been doing quite similar to a MLM and as we talked about that number of iterations have to be performed to reach to estimate these parameters so that we can get the best model which is the model best model which is fitting the data.

So, number of iterations actually indicate that and then we have one matrix which is quite similar to what we have multiple R square and linear regression right. So, this is actually computed where using the deviance value. So, null deviance and the standard deviation estimate or deviance that we have can look at the returned values here.

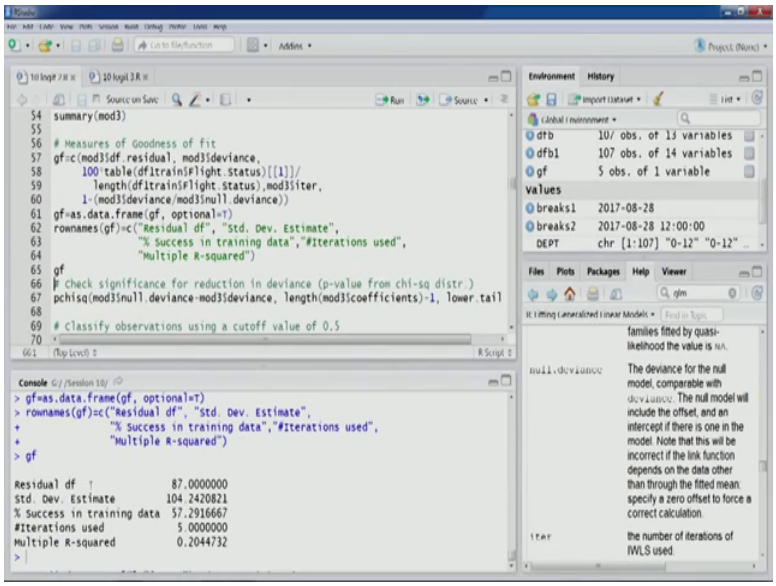
So, you can see deviance is the one of the return value so this is one and then we also null deviance is also a return. So, this is null deviance is come with respect to the naïve rule. So, 1 minus this deviance divided by null deviance gives us a value which is quite similar to what we have in multiple linear regression multiple R square.

So, this particular value will be will give us a metric which can be used to understand the goodness of fit follows logistic regression model and as on its own deviance also can be used it is quite similar to what we have there in as I see sum of squares error.

So, this is quite similar deviance is quite similar to that and then we can have one metric as I talked about similar to multiple R square. So, these metrics can be used to assess; the assess the fitness or model goodness of fitness goodness of fit of logistic regression model.

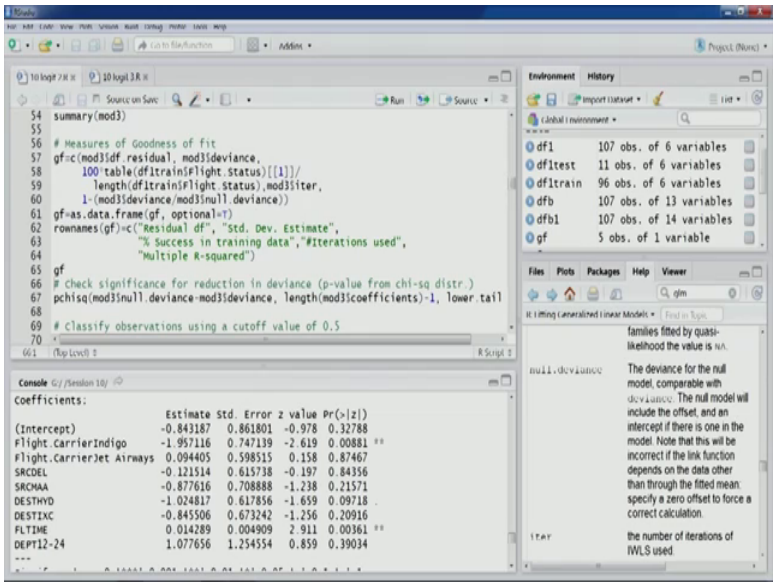
So, let us compute some of these things so, residual degree of freedom deviance which is similar to SSC in linear regression, then we have this proportions percentage in successive training data. Then we have number of iteration and then we have a multiple R square kind of metric let us compute this.

(Refer Slide Time: 11:34)



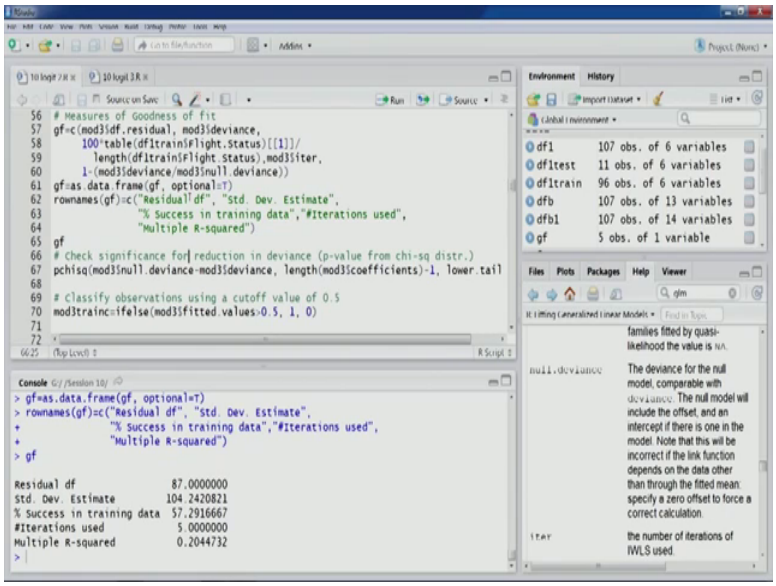
Let us create a data frame and row names we have given some and these are the values. So, you can see residual df df is 87 here, so as you can see let us again have a look df 1 training partition we have 96 observations.

(Refer Slide Time: 11:56)



And if we go back to our summary results right if we go back to our summary results we can see that how many variables we have here 1, 2, 3, 4, 5, 6, 7, 8; 8 8 variables. And we can see that 87 is the residual.

(Refer Slide Time: 12:06)



So, 87 plus 7 that makes it 94 that is n minus 1, so that is the computation that is how the degrees of freedom have been computed, so this is a correct value here. And then we have deviance value which is also called as standard deviation estimated by some software's some statistical commercial statistical software's.

And then so this is all similar to what we have SSE sum of squares error in multiple linear regression then we have number of iteration that have been used to arrive at the particular model and that we have that is not 3 in this case. Then we have a value similar to multiple are squares we can see 20 percent, of the variability in the outcome variable has been explained by this model. So, all we talk about that this is being computed by 1 minus mod 3 deviance, divided by null deviance.

Now, whether so in terms of on further in terms of deviance the null deviance represents the naive rule value. So, we have to see how much our model has been able to how much reduction in deviance has been done by our model and whether that that is significant or not so that can also be that can also be performed using this chi square test that we can do.

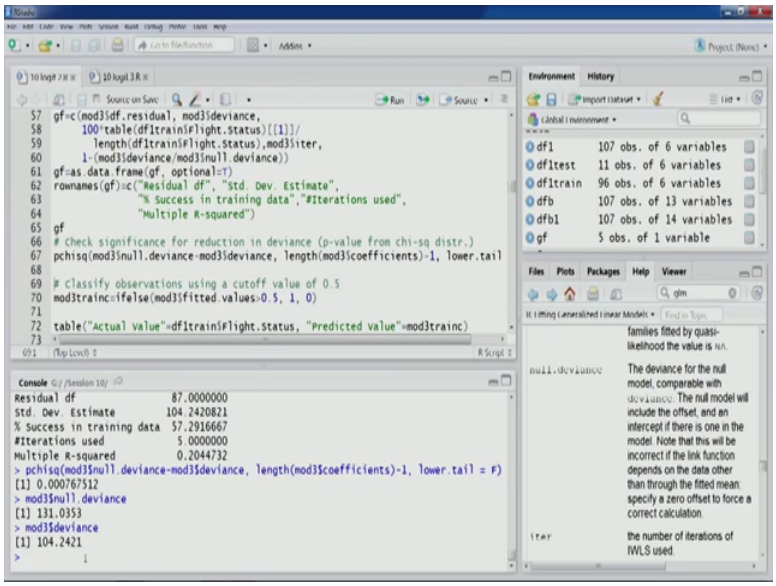
So, we have one function p chi square. So, there we can actually use these two values or we can take a difference of null deviance and deviance so that would be the reduction in deviance from a naive rule and you know how much deduction that our model has done.

And we can look at the number of predictors as degrees of freedom I could be used the number predictors that we use have used could be used degrees of freedom because these are freedom degree of freedom that had been used to reduce the deviance as we talked about from 95 minus 1 available degrees of freedom two we have reduced up to 87 that is residual degree of freedom so 7 predictors have been used.

So, that information can be use to perform chi square test and to find out the significance of whether the reduction has been significant or not. So, the third argument is lowered tail that is specified as false and we can compute this chi square value so we can see that this is a small value. So, therefore, it seems that this redis reduction is reduction deviance is significant which are also clear by the difference between the deviance and null deviance values.

So, we can also compute that we can see this is null deviance; this is null deviance. So, let us look at the value and we can have the deviance. Let us look at the value so you can see.

(Refer Slide Time: 15:17)



So, the difference is so there is good enough Reduction and deviance and that is why it also came as significant you know difference. So, these are some of the matrix that can actually be used to understand to a measure the goodness of fit to asses goodness of fit for a model as we talked about I talked about that in a statistical set setting. So, these are some of the matrix which would be more useful.

So, in a statistical modelling we stopped at when we build the model using the training partition. So, typically all the observations are used that are present and then we look asses the model with respect to some of these some of these matrix. Now, let us move forward to our next discussion point in logistic regression so that is let us move forward so that is this particular point whether linear regression can be used for a categorical outcome variable right.

So, there are there are some situations where linear regression can be used as a categorical outcome variable which we will discuss later; however, right now we are discussing some of the more important points with respect to overall general applicability of linear regression for a categorical outcome variable.

So, can be done technically it can be applied so we can treat the outcome variable as continuous. So, the categorical outcome variable can be treated as continuous variable. So, we can essentially do the numeric coding and keep it as a numeric variable and technically we can apply we will get the results; however, the results would be meaningful or not that we need to understand.

So, technically it can be applied we can read the categorical outcome variable as continuous variable we can code it numerically so that can be done. However, there are going to be anomalies that would lead to spurious modelling so what could be some of these things.

So, number one predict predictions can take any value not just any values so for example, that binary logistic regression model that we have been performing for on some of the data sets so they are the our outcome variable it is it typically has two classes class 1, and class 0.

And so, the values the remaining variable we will take is to 0 for class 1 and 0 and 1 for class 1; however, when we apply a linear regression model to a categorical outcome variable the prediction can take any values any real value can be taken and not just the dummy values 0 and 1 so that is one challenge.

How do we map some of these predators values which can which can be any real value to the actual values of the outcome variable 0 and 1. Now, outcome variable or residuals do not follow normal distribution. So, as we have discussed during a linear regression

that this is one of the important assumption that dependent variable that is outcome variable all residuals should follow normal distribution, but that is not the case as we can understand that categorical outcome variable will have just 2 values 0 and 1.

So, it is actually it actually follows a binomial distribution so this particular assumption would also be violated; however, we talked about that because we are in data mining modelling context. So, where as we talked about even if for you know prediction purposes even if this particular you know assumption is violated in terms of prediction it might not much matter much because generally check performance on validation partition and test partitions; however, this case is different.

The deviation from normal distribution is much higher it is actually different distribution binomial distribution so that is one problem. Now, the another assumptions that we talked about in multiple linear regression is homoscedasticity; however, if we apply linear regression to an a to a categorical outcome variable this particular assumption would also be violated. The variance of outcome variable that we that we expect to be constant across all the records that is the homoscedasticity property so we for to apply multiple interrogation we want our outcome variable to follow this to have this property.

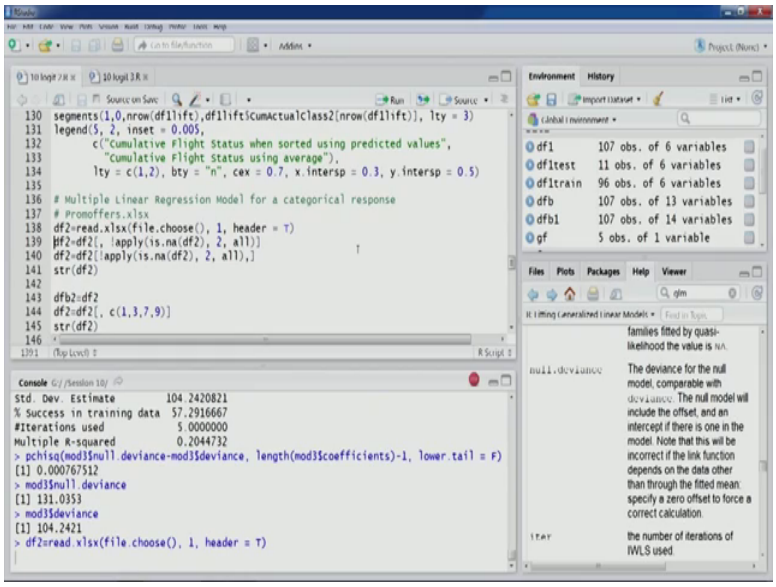
So, variance should be constant across all time; however, if we look at the variance for our categorical outcome variable it is going to be this particular value n times p into 1 minus p and as you can see because this is dependent on the value of p. So, therefore, the variance will change as the value of p changes.

So, when the value of a probability value is actually close to 0, then the variance would be on the as lower side and when the value of p is approaches 1 then the variance would be on the higher side. So, therefore, the variance will be will not be constant and it will be it will increase as the probability value you know in case is from 0 to 1.

So, so some of these are some of the problems that we can see in that we can directly understand and why a linear regression and cannot be applied to category outcome variable in a general sense and the problems that we can see here. So, what we will do we will do an exercise in R to understand the same thing to understand this particular aspect.

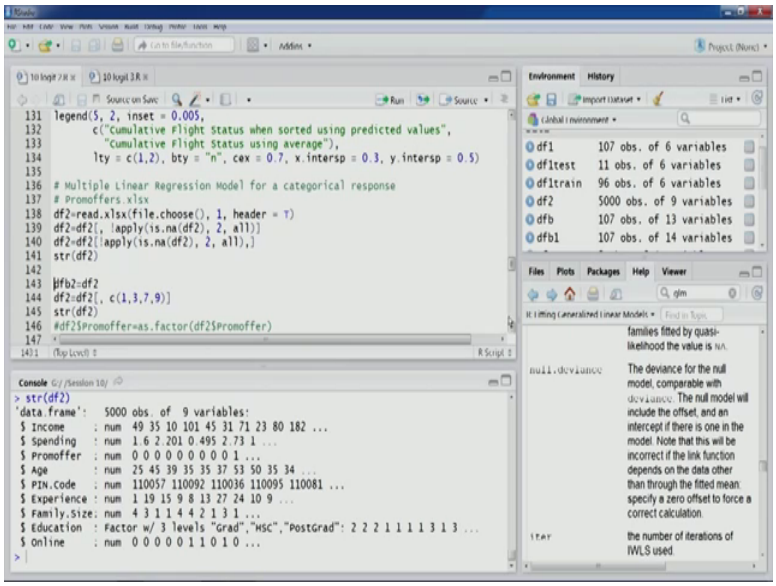
So, what we will do we will apply a linear regression model on a logistic partition and see the see its applicability and see some of the anomalies or violations that are that could be there. So, for this purpose we are going to use as you can see multiple here the comment is multiple linear regression model for a categorical response. So, promotion offers is the data set that we are going to use for this particular exercise. So, let us import the data set let us edit load.

(Refer Slide Time: 22:16)



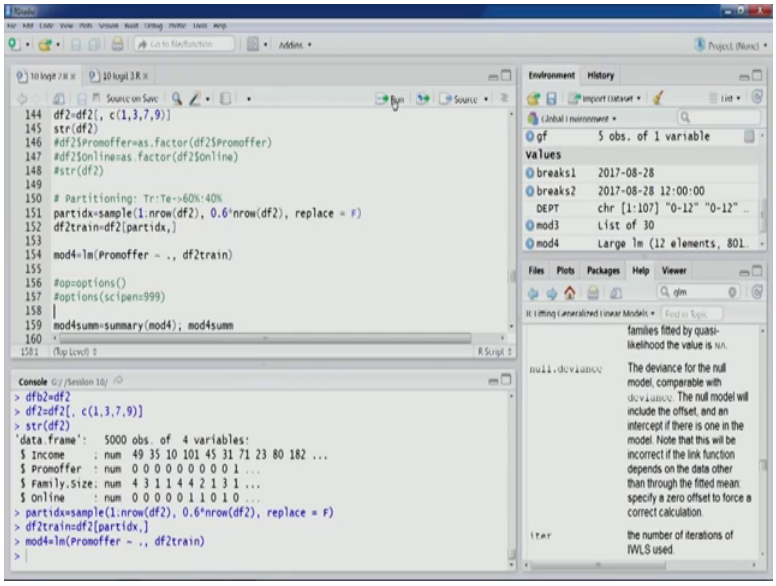
So, once the observations have been loaded into environment we as we will see in the environment section we will do some of these steps I think it is has been loaded yes. So, df 2 we can see 5000 observation, 9 variables so let us remove NA columns, or NA rows if there are any let us look at the structure so this is the data set. So, we are already familiar with this.

(Refer Slide Time: 22:41)



So, let us go through some of the some of the comparisons some of the transformations that we have been doing you know in previous lectures as well take a backup let us get rid of let us select just these 4 variables that is income promotional offer of our outcome variable then we have I think family size and online activity. So, let us select them. So, these are the variable selected for this particular exercise income promotion offer and family size and online.

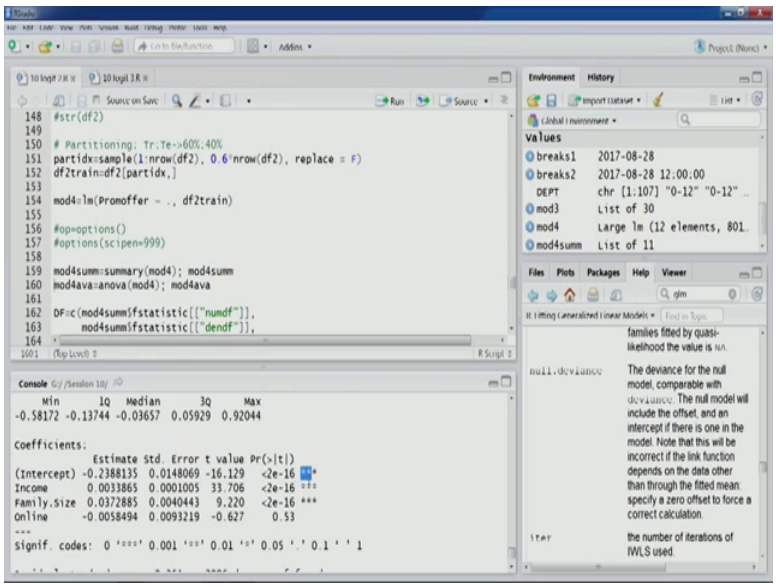
(Refer Slide Time: 23:12)



So, the variable that we are going to use for our outcome variable is going to be the promotional offer as you can see that we have commented out the lines of code, which we used earlier to convert these numeric variables going to convert numeric variables into the categorical variable. So, promotional offer an online so they are actually categorical variable factor variable, but we are not converting them into factor variable because we are going to apply a linear regression modelling. So, we will give them as numeric and we will apply.

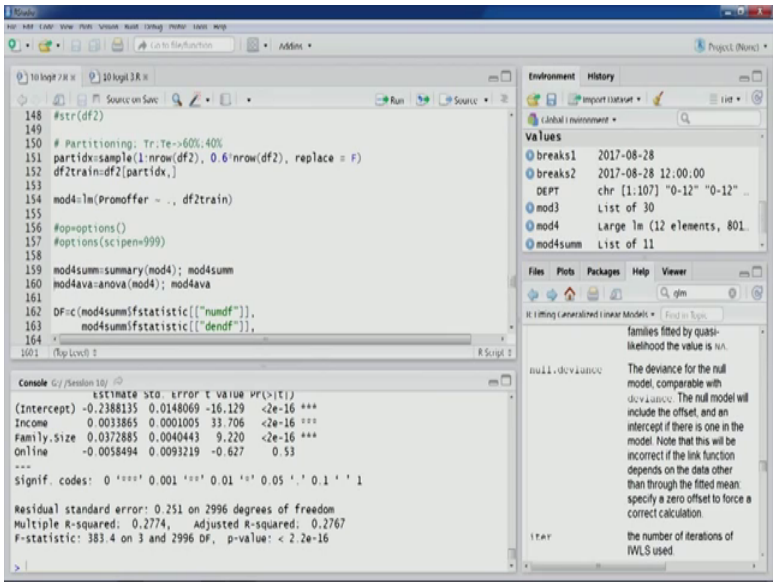
So, a partitioning is the same 60 percent, 40 percent in this case so we can see that let us do partitioning. So, df 2 train you can see 3000 observations, 4 variables. Now the same align function is going to be used now the promotional offer is going to be request against all the predictors that are present in this particular dataset df 2 train. So, let us run this.

(Refer Slide Time: 24:30)



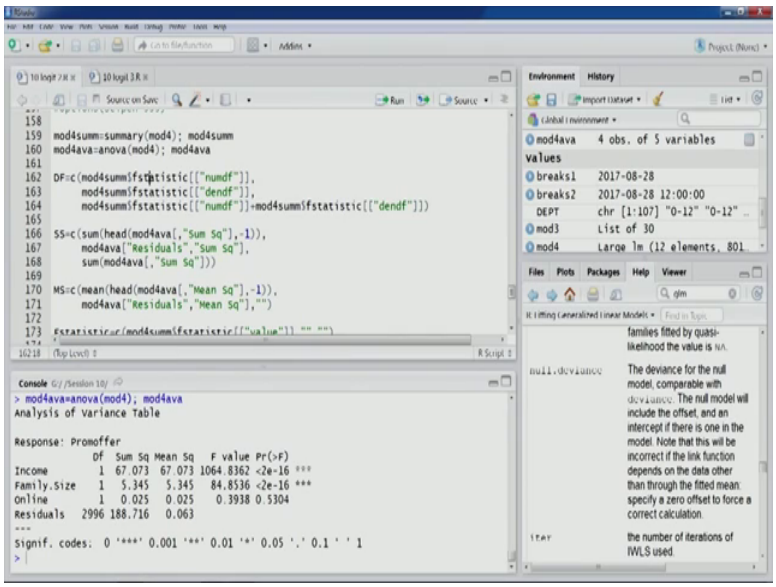
Now, what we will do we will look at the summary table. Let us look at the results. So, as we can see that income is intercepted significant income is significant and family size is significant we can look at the different estimates for example, income quite a small value from this family size is also 0.03, online is not significant it has not only found to significant as we can see.

(Refer Slide Time: 24:55)



And we can have a look at the other values we can see adjusted R square R square and multiple R square. So, we can see if we you know 27 percent multiple R square value we look at the p well it is quite as small. So, the model is significant ah; however, as we talked about certain problems as we have discussed could be there certain anomalies could be there.

(Refer Slide Time: 25:21)

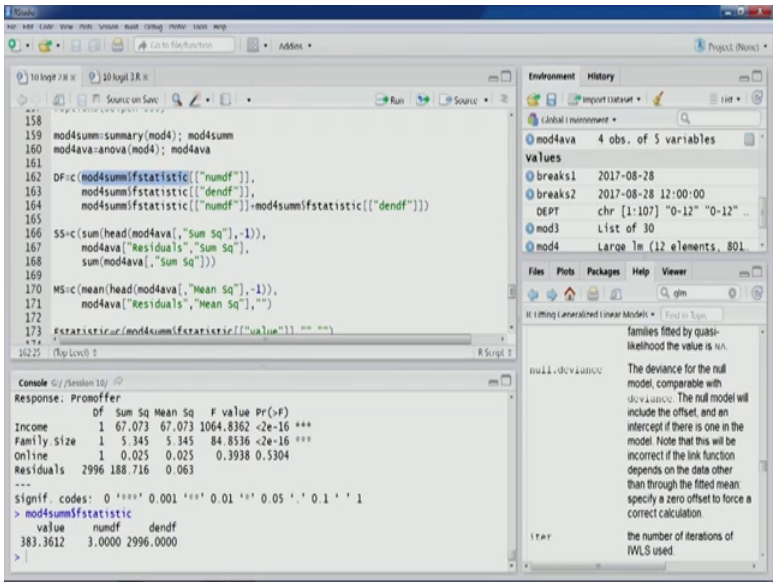


So, what we will do? We will look at look at computing some of those things to check whether those anomalies are there in this particular case. So, what we will do we are

running and over to extract some of the parameters here. So, we can see so sum of square errors mean of square errors and F value of a statistic and the probability values for these predictors income family size is given there in the ANOVA table.

So, what we are going to do is we are going to compute these particular values in a in a format that can be used for tabular presentation later on. So, mod 4 so on; we have a F statistics value. So, that is written as part of the summary function of models.

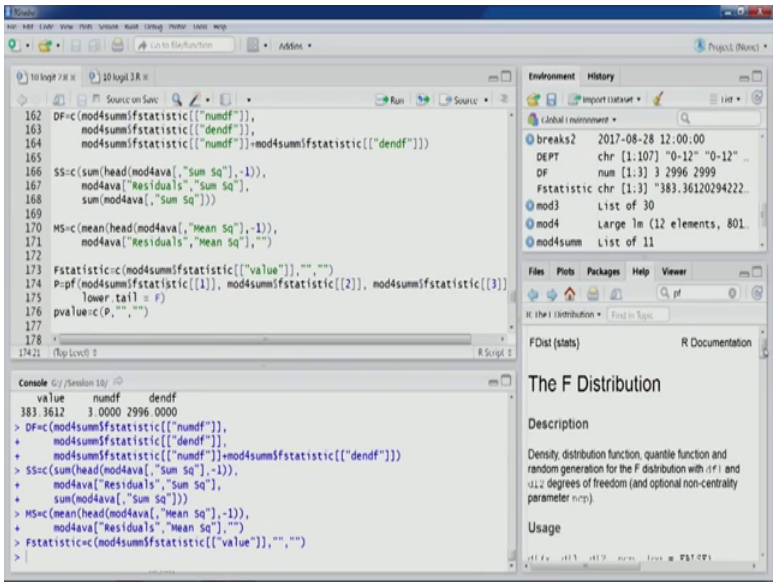
(Refer Slide Time: 26:19)



So, when we apply summery on the model object so you can see this is nothing, but F statistic value for the model and then we have degree of freedom and here. So, we can see the degree of freedom residuals degree of freedom and the residuals degree of freedom and the regression degree of freedom here so that is going to be stored in this so in this data frame.

First we have the regression degree of freedom, then we have the residual degree of freedom and then we have the total so this data frame is about degree of freedom as is sustained by its named DF. Then we will compute some other square error so first we can see that in the ANOVA table, from the ANOVA table we are trying to extract out this these values.

(Refer Slide Time: 27:21)



So, first some other square for the you know regressors regression and then for the residuals, and then total. So, this would be recorded in this particular variable SS. And then we look at the mean or mean of square errors, so that is also being extracted from the ANOVA table results, you can see this particular column mean square and these values are being instructed.

So, first 3 values so this head function as you can see now this role of this head function is quite different and both these computations we have used head function.

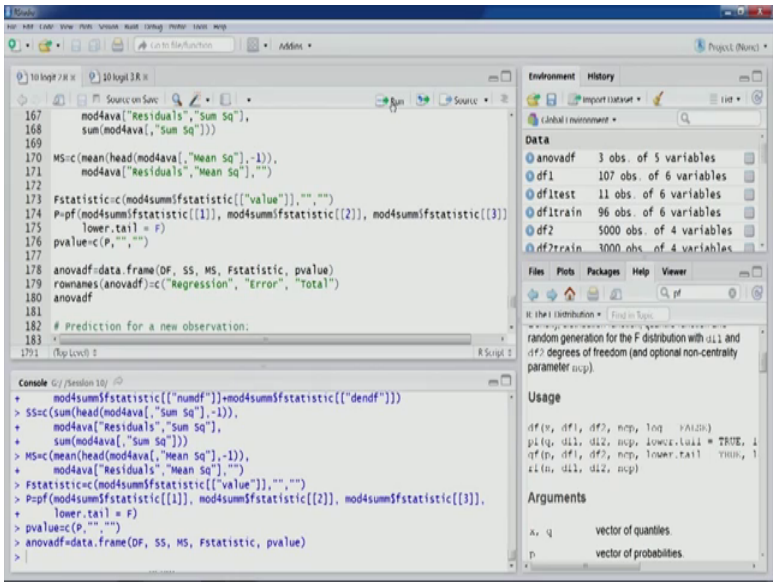
So, you can see the second argument is minus 1 so what it actually does is it gives us all the values except the last value in the vector. So, for example, mean square or some other square. So, there are 4 values and these 2 columns, and these 2 vectors.

So, except the last values that is corresponding to residuals the first 3 values are going to be written; that means, last value will be left out and the remaining and remaining n minus 1 values are going to be written. So, that is what we want. So, that will give us the corresponding sum of squares or mean sum of squares are for the deviation.

So, first that then residuals so let us compute this. Then let us also extract the F statistics, from the this particular vector that we have already seen. So, let us do that then what we are trying to do we are trying to compute the corresponding probability value.

So, probability value corresponding to the F for test so that that is how it is being committed pf is the function that can be used for more detail you can go into the help section and find out more information about pf. So, you can see this is F distribution.

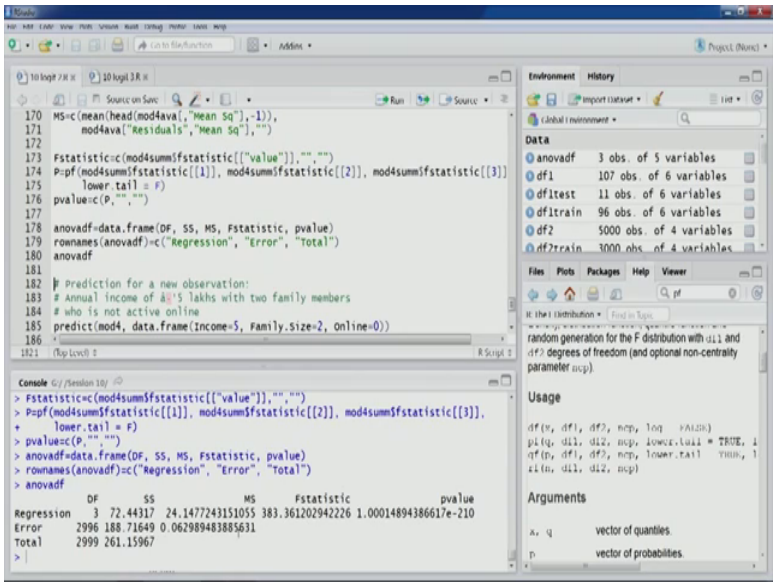
(Refer Slide Time: 29:09)



So, within F distribution we have F function which can be used to compute the corresponding p value. So, what we need is F statistics and the first argument then we need degree of freedom as second argument that is responding to a regression here and the second is then last one is residuals degree of freedom third argument and then we have a specified lower tail as false.

So, we will get the p value corresponding p value. So, let us record it in this format once this is done we can create this table data frame. And let us assign a names for this and let us have a look at this particular table. So, we can see that now once computed.

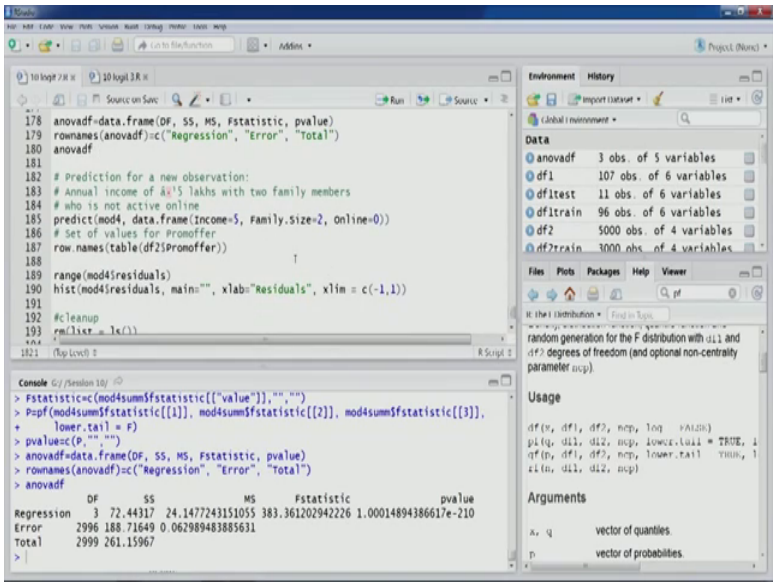
(Refer Slide Time: 29:50)



So, first column we have degrees of freedom. So, for regression there are 3 and then because as remember that we have used just 4 variable, so one being outcome variables, so 3 predictors. So, therefore, degree of freedom degree of freedom is here is 3 for regression then residuals it is the remaining that is n minus 1 is 2999, total number of observation and is 3000.

So, residual degree of freedom 2996 sum of a square for regression and for residual is also there. So, you can see that residual sum of square is much higher so that also is we can we saw that the variance was also on the lower side right earlier we saw that the variance that we had computed was on the lower side so that is also indicated here, mean square of errors is also there and then we have F statis F statistic and then we have p value is small value.

(Refer Slide Time: 31:00)

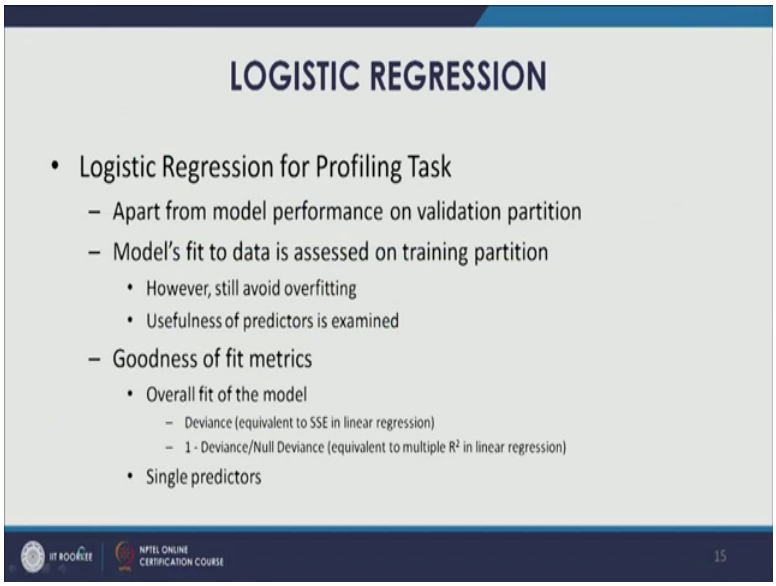


So, this gives us some information about the model and now what we will do is we will use this model to the score of a particular observation. So, let us do a prediction for a new observation let us say this is our new observation is annual income of rupees 5 lakhs with 2 family members who is not active online.

So, this is information about the particular customers customer whether is going to accept the promotional offer or not. So, annual income is 5 lakh family size is 2 and not acting online. So, you can use the predict function first argument is as usual the model object mod 4 then in a data frame we are trying to pass on the values operate predictors for example, income 5.

So, it should be in the same unit as was used for the modelling exercise in the training partition. So, you can see 5 and family size is 2 and then online is because this particular customer is not active so 0. So, once we do this we will be able to predict this particular observation. So, you can see this comes out to be a negative value. Now that was one of the first point that we discussed in the slide.

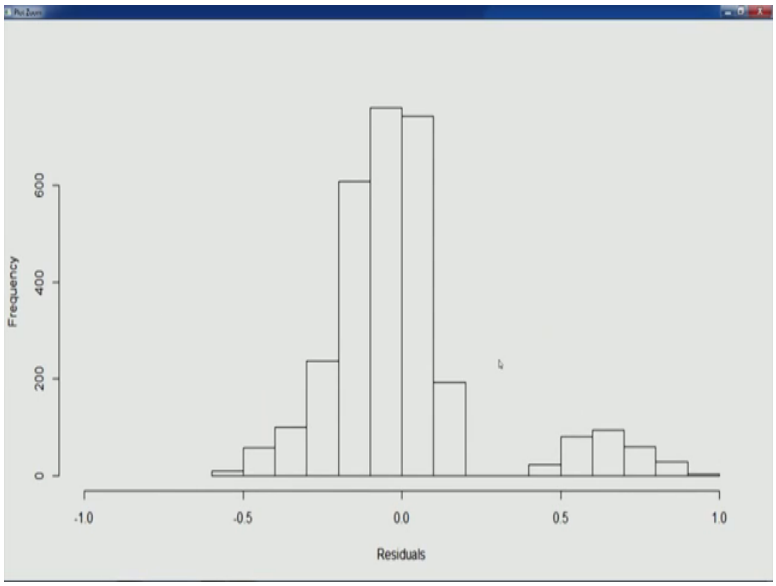
(Refer Slide Time: 32:11)



Let us go back you can see here that the predictions can take any value not just dummy values. So, we can see that a negative value has been taken here and so that is in one anomaly that we can clearly see. And set of values for them was no offer that we already know 0 and 1, and the value is that comes out to be predicted value comes out to be minus 0.14 now.

So, one difficulty is how do we do our classification in this case? Now, let us look at the residuals so second anomaly that I discussed that outcome variable of residuals do not follow normal distribution. So, let us look at that let us plot a histogram and find out whether this is being followed.

(Refer Slide Time: 33:07)



You can see residuals when we plot residuals when we get histogram you can see this is clearly not being followed normal distribution is not being followed one grouping here and other grouping where this is of this lower values, lower frequency and this is higher frequency. So, this is clearly looks binomial or different groups kind of thing. So, this definitely not following so normal distribution and distortions due to real binomial distribution can be seen here.

So, the typically what exercises and the discussion that we have been doing was pertaining to classification tasks and as we talked about that this is a statistical technique logistic regression being a statistical technique is also applied in so used in statistical modelling. So, the kind of task that we generally do in a statistical modelling are quite similar to what we can call profiling tasks.

So, as we talked about in the you know starting lecture of logistic regression and that it is about understanding similarities and differences between two groups. So, logistic regression can also be used in can also use to understand what are the variables which you know which can bring out some of the similarities or differences between group, so let us discuss that aspect as well.

So, in profiling tasks when we talk about profiling tasks the situation is slightly different. So, in the classification task we typically build our model and look at the performance of that classifier that particular model using the classification matrix using the overall

accuracy or overall error matrix and you know some deviations when we have a class of interest.

So, those are the things that we typically do we also typically look at left chart especially when we have a class of interest to in left chart and Decile chart to see whether it is still despite you know higher error whether it still the model is useful in class of interest with respect to naive you know naive rule or a or a average case.

So, some of those things that we do in classification tasks; however, in profiling tasks what we do in classification we follow that. So, apart from model performance on validation partition we also asses models fit to data on training partition right because as I taught what in a statistical technique typically the whole sample is used; however, since we are using training partition for model building.

So, the models fit to data is assessed on training partition; however, model performance is assessed on validation partition for profiling tasks and models fit to data assessed on training partition. So, some of these things we talked about when we talk about goodness of fit measures we talked about the deviance, we talked about 1 minus deviance divided by null deviance that is equivalent of multiple R square some of those things you can see here as well.

So, models fit to data is assessed on training partition ah; however, we still focus on avoiding over fitting because as we talked about when we have matrix which look for and when we do modelling to achieve the you know goodness of fit then it can it can lead to over fitting. So, it still we would like to avoid over fitting and still be able to you know check the performance still be able to have good classification performance as well.

So, usefulness of usefulness of predictors is also examined in this particular case profiling so because it is about understanding similarities and differences between 2 groups. So, therefore, which predictor is more helpful in terms of bringing out those differences those similarities so when we build our model. We also look at the significance levels of some of these predictors we look at which predictors are significant which are not significant whether the not significant predictors can be dropped from the model.

And this has also should be looked at from the perspective of model performance because as we have taught about data mining model we would like to keep the insignificant variables also in the order if they provide some practical importance in terms of scoring new observation.

How and logistic statistical modelling we just drop the insignificant variables because we are just interested in understanding the phenomena, understanding the underlying relationship. So, profiling is quite similar to you know that that approach; however, because we are doing data mining modelling. So, we have to balance between these two, we have to balance between performance and also the main profiling tasks.

So, we have to really see whether the predictors can be dropped just like in statistical training or whether they have to be kept in the model because they also provide some practical significance for scoring new observation so that balance has to be achieved.

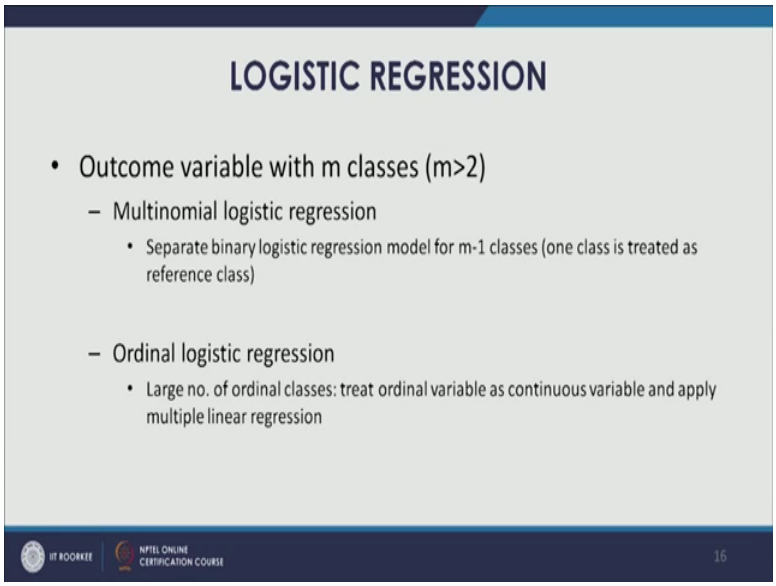
So, we have to avoid over fitting, we have to look for using usefulness or predictors in both the context data mining context prediction point of view and also statistical context in terms of understanding you know finding understanding variables which differentiate the groups, which bring out similarity or differences between groups.

So, this kind of exercise is done in profiling tasks and goodness of fit matrix that through an exercise in R that we have already understood overall fit. So, we look to understand first we look to understand the overall fit of the model, so if the overall fit of the model is good only then we go ahead and look at the individual variables.

So, first step typically is in profiling on a statistical modelling we look at the overall fit of the model. So, in this particular case logistic regression as we talked about the deviance is the metric we taught one previous lecture that could be used and we also said that this is equal to SSE in linear regression and 1 minus deviance is divided by null devian that is equivalent to multiple R square in linear regression.

So, these are two matrix that could be used to assess the overall fit of the model and then once this is done then we look at the single predictors. We look at whether they are significant or not as I talked about and whether we can strike a balance in terms of prediction performance, classification performance, versus the profiling that is and also statistical modelling context.

(Refer Slide Time: 40:04)



So, with this we move to our next discussion that is about so till now the exercises that we have been doing they were mainly focused on binary classification, and binary logistic regression model we just had 2 classes class 0, and class 1. Can logistic regression we extended to a scenario where we are dealing with more than 2 classes where we are dealing with you know m classes. So, yes it is possible so we will discuss some of those things.

So, first one is multinomial logistic regression so multinomial logistic regression. So, the categories the classes that we have they are you know so the categorical variable is nominal, so, in that case we can apply multinomial logistic regression. So, what happens in multinomial logistic regression first out of those m classes that we have we have to select one we have to pick one as the reference category, and for the remaining m minus 1 classes we create separate binary logistic regression model.

So, for each of the m minus 1 classes apart from the reference category class. So, we will have n minus 1 classes so for each of the n minus 1 classes will create separate binary logistic regression model; that means, for a class 1 we will have the scenario where the observation probability of belonging to class 1 and probability of long not belonging to class 1 so that kind of binary scenario we will have and that for each of the n minus 1 classes.

So, we will be dealing with m minus 1 binary logistic regression equations and so using that we can compute all those probabilities values with the help of the predictors and then the remaining reference class of that probability for that can always be computed by the m minus 1 probabilities values for the m minus 1 classes.

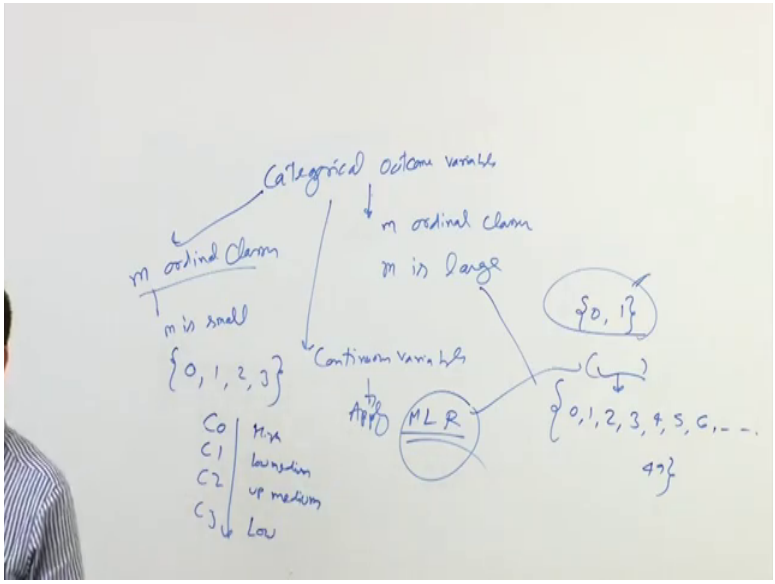
So, we can just subtract that from one for you know and then we can get for the probability value for the different class. And once we have the probabilities values for all m classes then we can apply our most probable class method routine where the class having highest probability value would be assigned to the new observation.

So, this is how we can go about applying logistic regression to an outcome variable with m classes. So, this is called this is a multinomial logistic regression and this is applicable mainly to nominal categorical nominal variable right, sos the categorical variable having nominal classes.

The second scenario second scenario is about when we have a categorical variable with ordinal classes we have an ordinal variable. So, in those cases we can apply ordinal logistic regression. So, so again within this ordinal logistic regression we can have 2 scenarios. So, as we have understood in some of the initial lectures and supplementary lectures that ordinal variables they have order among different labels is also important that is also meaningful.

So, less than or equal to or greater than or equal to operations they are also applicable in this case, so the first scenario is large number of ordinal classes. So, if our outcome variable which is categorical variable with ordinal classes if that variable have is having large number of ordinal classes, then one solution is treat that ordinal variable as continuous variable and apply multiple linear regression right.

(Refer Slide Time: 44:10)



So, when we have a categorical variable with ordinal classes so we have a categorical variable. Categorical outcome variable when m ordinal classes, and m is large then I as I talked about we can treat this particular variable as a continuous variable and apply multiple linear regression.

So, multiple linear regression can be applied and reason is one reason one justification for this is as we talked about earlier that in binary situation we have just two values for a categorical variable and the predicted values up using MLM can range anywhere any real value so, that was the one main problem here.

But when we have m is large; that means, set of values could be you know many more. So, it could be you know if there are 50 groups let us so in this fashion we can go on up to this. So, the number of values that can be taken by this particular this particular ordinal variable are many more.

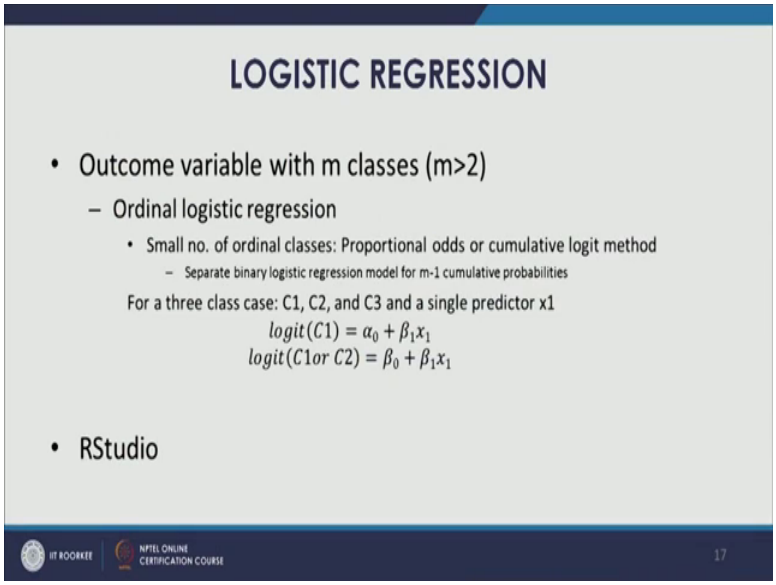
So, therefore, the predicted values this can be easily mapped to some of these values could be close to some of these values, and probably multiple linear regression can be still applied so this is one way. When we have a ordinal variable with many number of classes with large number of classes, when m is large then is still you know instead of logistic regression we can apply multiple linear regression so that is first scenario.

The second one is whatever we have a small number of ordinal classes. So, if m is in this case m ordinal classes that we have if this is a small m is small, then probably we will we will run into the same problem like for binary classification. So, similar problem would be there we will have only few values 0, 1, 2, 3 let us say these many.

So, again small number of classes small number of ordinal classes, so we will have I will run into same problem. So, so we what we do is we use a different version of logistic regression called proportional odds or cumulative logic method as indicated in the slide, so small number of ordinal classes so we would like to use proportional odds, or cumulative cumulative, or logit method.

So, what we do here in this particular method is we create separate binary logistic regression model for m minus 1 cumulative probabilities, so, we talked about that when we so, when we discussed multinomial logistic regression; So, for all m minus 1 classes will have separate binary logistic regression model.

(Refer Slide Time: 47:50)



However, if you see that here will have separate binary logistic regression model not for m minus 1 classes not for presence of that class or absence of that class, but m minus 1 cumulative probabilities.

So, let us understand what we mean by that so let us take an example for a 3 class case as written in this slide for a 3 class case C 1, C 2, and C 3. Let us say these are the 3 classes

ordinal classes C 1, C 2, C 3 and a single predictor x 1 that is being used. So, our logit equations could be something like this logit for C 1 it could be alpha 0 plus beta 1 x 1 and logit for C 1, or C 2 so; that means, from C 1 first logit equation is just for you know observation belonging to C 1. The second is observation belonging to C 1, or C 2 so that gives us the cumulative sense. So, as we talked about that ordinal the order is important so that that means, you know different classes can be compared.

So, therefore, C 1, or C 2 is a you know is a meaningful here in the sense that if we look at the rights part of the equation beta 0 plus beta 1, x 1 you can see that beta 1 is same and both these equation x 1 so you can see because the this is the comparison can be done.

So, the coefficient so the intercept R only difference because the comparison so one is when we talk about ordinal, so one particular class this ordering this ordering is you know this is meaningful. So, when the ordering of classes is meaningful; that means, one can be you know said you know less than one particular class the or higher than one particular class just like the categories that we might have high, you know low medium, upper medium, medium and then low.

So, this kind of this kind of classes we might have all we might have you know a strongly agree to strongly disagree. So, those kind of ordinal classes where the order is meaningful, where the order is meaningful then in those regression we can have something like this cumulative probabilities values, and beta coefficient is can be used the same beta coefficient can be used for both these logistic regression.

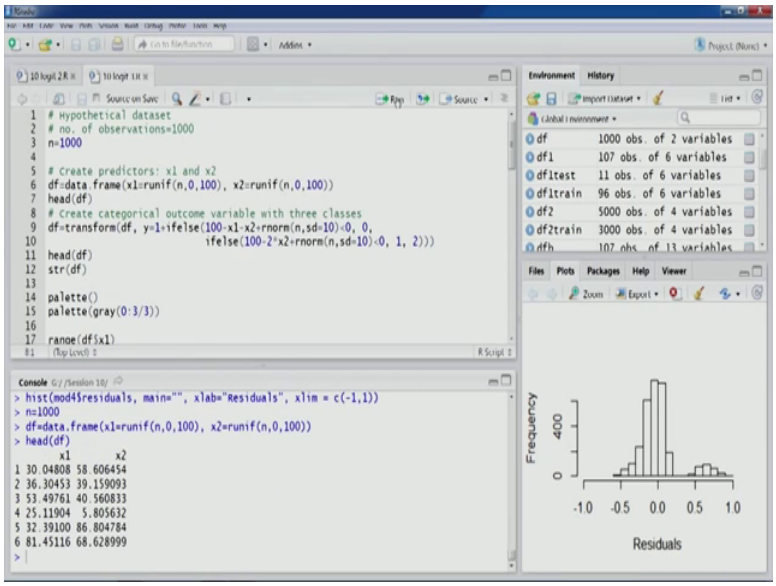
And from this we can compute the cumulative probabilities values and once these cumulative probabilities value values for two of the classes for these classes have been computed the actual probabilities value for C 1, C 2, C 3 can be derived from using the probability value that formulation that we have so, it is from logit, we can derive the probabilities values.

So, once these cumulative so from there once we have the probabilities values for all these classes then we can again apply the most probable class method and also assign the class based on the based on the probability value and again this so in this fashion in this fashion we can apply ordinary logistic regression to a scenario with fewer number of ordinal classes.

So, what we will do we will go through R Studio and do an exercise for this when we have classes more than m is greater than 2, so a logistic regression modelling for classes greater than 2. So, what we will do we will create a hypothetical data set here, so number of observations are 100 in this case, so as you can see and it is 1000.

So, let us create this now what we are going to do is we are going to create a data frame having that data frame where we have 2 variables x 1 and x 2, x 1 as you can see we are using run if for x 1 and x 2 both so the observations. So, the val values would lie between 0 and 100 and n number of values would be created that is 1000 values would be created lying between 0 and 100.

(Refer Slide Time: 52:23)



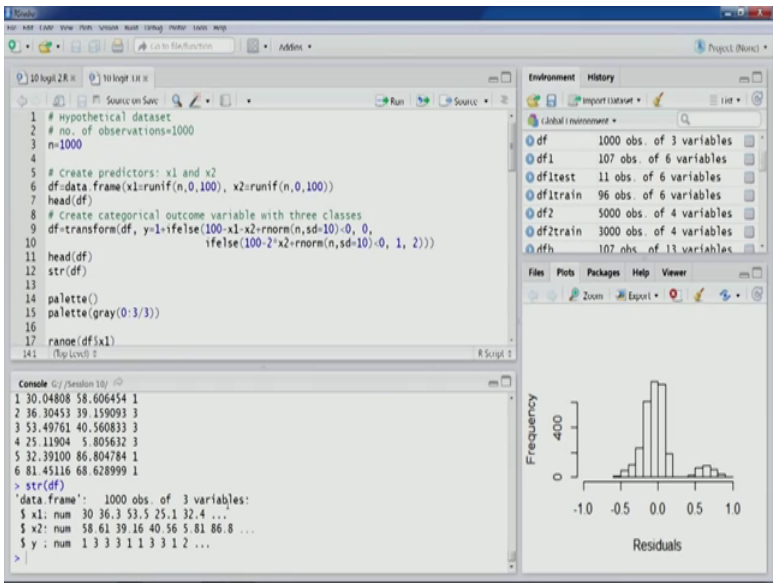
So, let us create this data frame, let us look at some of the observation. For 6 observations you can see all the values they are lying for both the variables x 1 and x 2; they are between 0 and 100.

Now, what we will do we will create a categorical outcome variable with 3 classes. So, this particular data set that we are trying to create we are going to use it for both the scenarios multinomial scenario, and ordinal scenario right, so this is just for illustration purpose so we are not specifying whether the variable is ordinal or not s, we are just going to use it for both the scenarios.

So, what we are going to do is we are using transform function to create our outcome variable, categorical outcome variables. So, you can see that we are trying to compute y as 1 plus if else and if the value is less than 0, this particular value 100 minus x 1 minus x 2 plus again a certain value is being taken from normal distribution no standard deviation 10000 values.

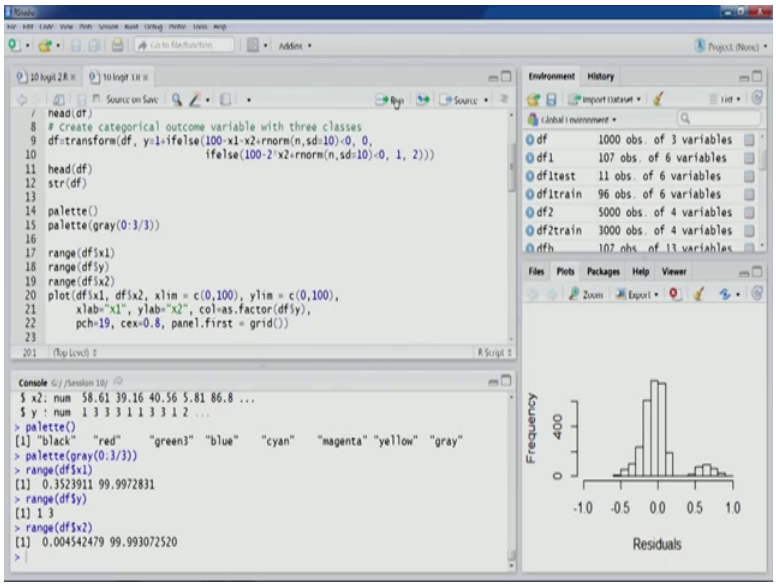
So, if it is less than 0, then you know we are using this information and that is information based on x 1, x 2, and certain additional computation we are trying to assign it a class 0 or then the second if that is not true then second computation so it will get a class 1 or class 2. So, let us compute this once this is done let us look at the observations you can see another variable y that is categorical variable has been created having you can see 1 and 3 2 values are being taken. Let us look at the structure of this particular data frame.

(Refer Slide Time: 54:00)



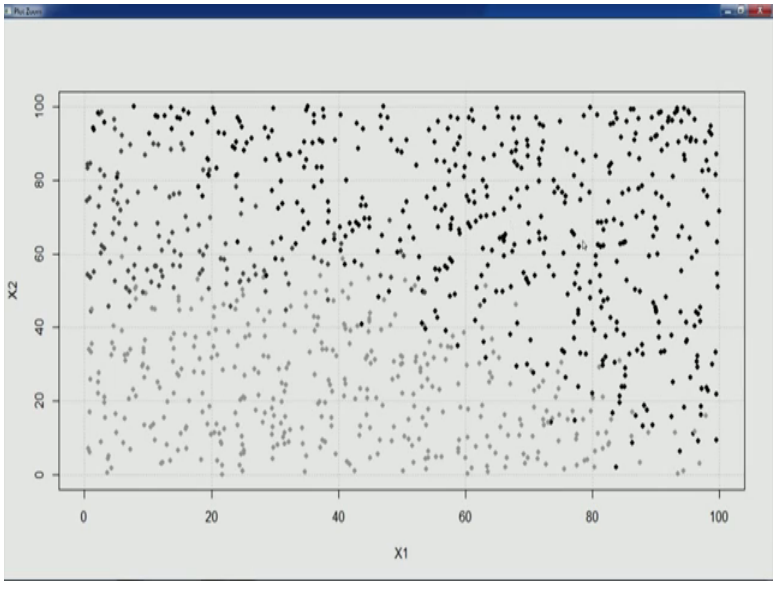
So, these are 3 variables x 1, x 2 values between 0 and 100, and y is taken value 3 values 1, 2, and 3. So, let us plot this particular data set. So, this is our default palette so; however, I would like to use this palette agree 3 sets. So, in this fashion we can have sets of no number of sets that we require. So, let us look at the range of x 1, y, and x 2 which is already.

(Refer Slide Time: 54:25)



Because we have just now created these variables studies actually clearly understood as well so limits 0 to 100, and 0 to 100 and then colouring is with using this particular factor y. So, you can see as dot factor we are using this particular variable and let us create this plot.

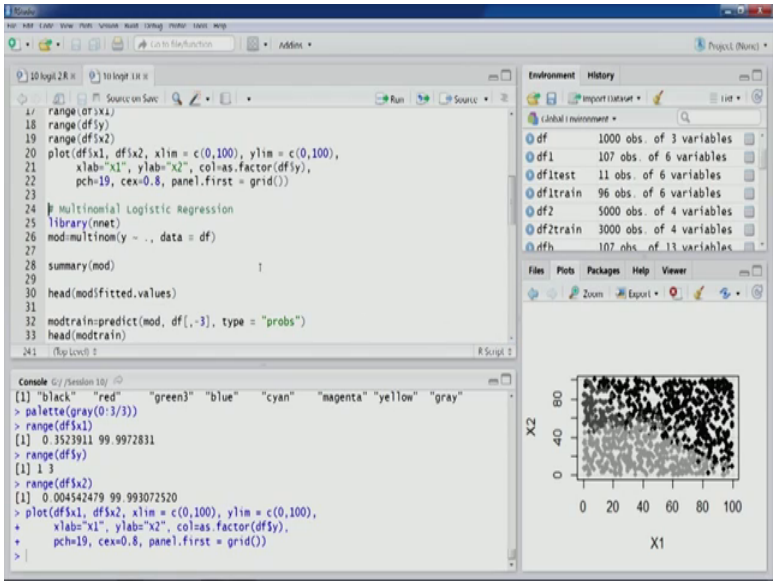
(Refer Slide Time: 54:58)



So, you can see so this is our plotting. So, a one group is here, the second group is having some medium level gray set here and the third group is here, this is lighter gray code

colour. So, these are the values that we are going to use x 1, x 2 this is plot between x 1, and x 2 and the outcome variable has been color coded, so, 3 categories.

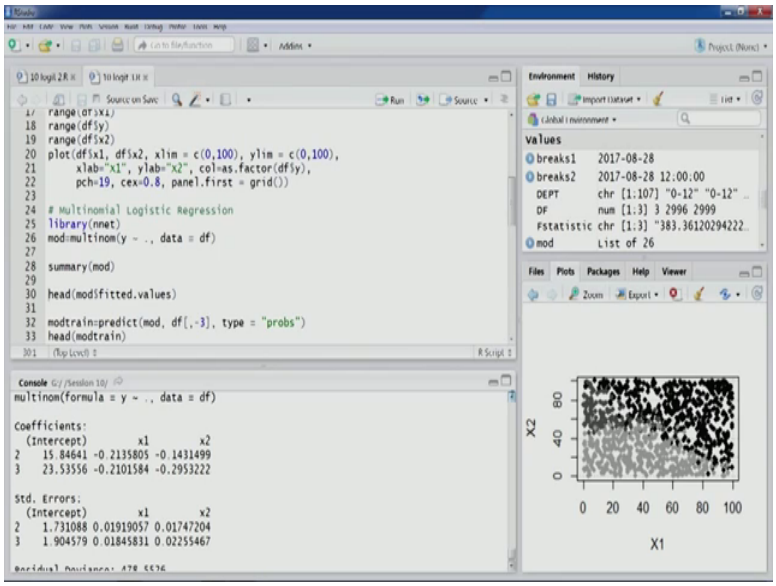
(Refer Slide Time: 55:25)



Now, what we will do is we will use the multinomial logistic regression. So, for this we need this package the library, and net package this package actually for a neural network, but it provides us offers us this function which can be used for multinomial logistic regression.

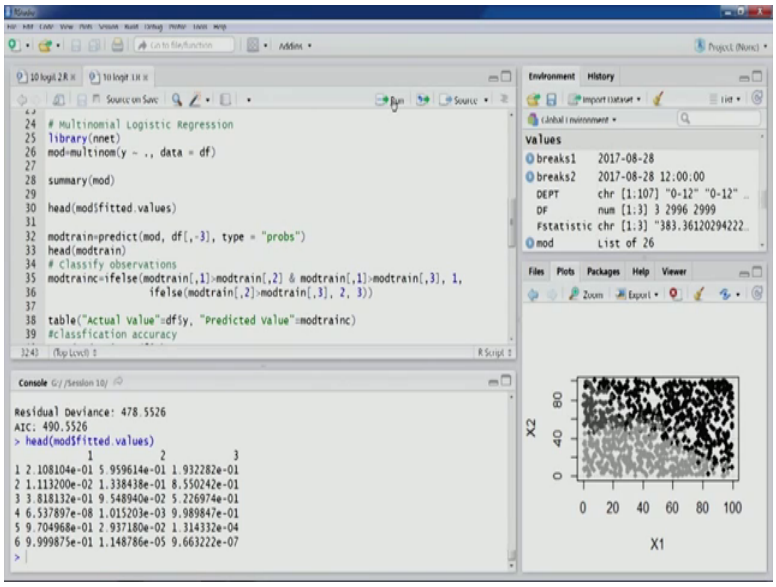
So, a multi norm is the function so what we are now going to do is the outcome variable y we are going to regress it with the remaining variables that is predictors in this particular data. So, all the observations we are going to use here, so let us run, this let us look at the summary.

(Refer Slide Time: 56:00)



So, we can see coefficients x 1 and x 2 right. We can also see the error values residual deviance and the AIC values, are also indicated here. Now, fitted values also we can look at first 6 observation here.

(Refer Slide Time: 56:19)



So, we will get these fitted values which are nothing, but estimated probabilities values for these 3 classes 1, 2, and 3 so what this is what we have. Now, if we had another partition validation or test partition you can use predict function to the score those

partitions and have probabilities values, estimated probability values, for those new observation.

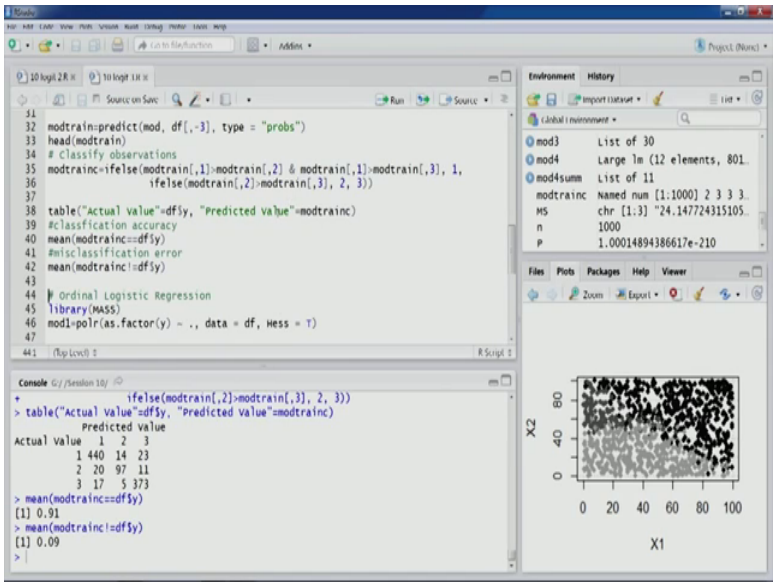
However for the demonstration purpose, we are applying predict function on the training partition that is the full data set itself. So, you can see another argument that we can see is type which is props which is proper probabilities values. So, let us run this and you would see we will have the same probabilities values which were fitted. So, we can see here that same values you can see first row it is same, same values. So, the fitted values and all we have got the same one using predict function.

Now, let us classify these observations. So, this is how we can classify if the probability value for 1 is greater than the probability value for class 2 and also greater than the probability value over class 3, then of course, this observation has to be classified to class 1.

Otherwise we will again compare the probability value for class 2, with probability value for class 3, if again it is greater and I know class 3 value divided then the class 2 is assigned, otherwise class 3.

So, in this fashion we can assign all the observation into appropriate classes so this is implementation of most probable class method. So, once this is done you can see mod train see for all 1000 variables have been created and observations have been assigned to go to appropriate classes.

(Refer Slide Time: 58:13)



Now, let us look at the classification matrix. So, we can see actual value, predicted values, now till now the classification matrices that we have been observing that we have been creating they had 2 values 0 and 1. And we had 2 by 2 classification matrix now this time we are seeing 3 by 3 classification matrix.

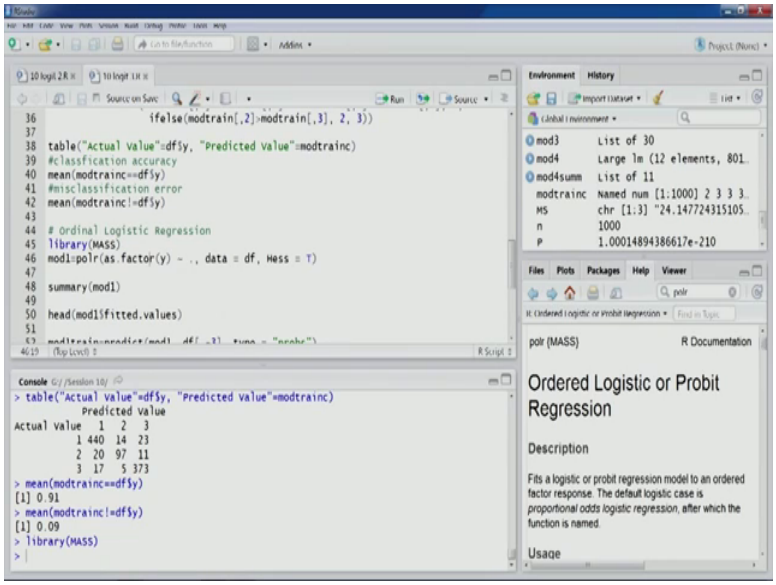
So, 3 possibilities for actual values 1, 2, and 3 and the predicted values and we have the corresponding numbers here. So, you can see that again the diagonal element; that means, these 3 elements they represent the values records which have been correctly classified, and the remaining observation off diagonal elements you know they have the records, they have the counts of record, which have been incorrectly classified right.

So, from this we can compute the classification accuracy that is 91 percent in this case, and error that is the remaining 9 percent. So, in this fashion we can apply multinomial logistic regression to a particular data set. Now, let us move to next part that is ordinal logistic regression.

So, in this product case ordinal logistic regression as I talked about 2 scenarios, so 1 where m is large; So, in the in that scenario we can apply multiple regression one exercise that we had done in previous lecture, we had applied multiple linear regression, but that was for the class which had just two you know for a variable category which just had 2 classes 0 and 1.

In the same fashion for the first scenario the multiple linear regression, can be applied. So, there is not much much different in terms of applicability. So, well what we will do we will do an exercise for this scenario where we have to apply this ordinary ordinal logistic regression the cumulative probabilities method, cumulative logit method so for this we need this package mass.

(Refer Slide Time: 60:33)



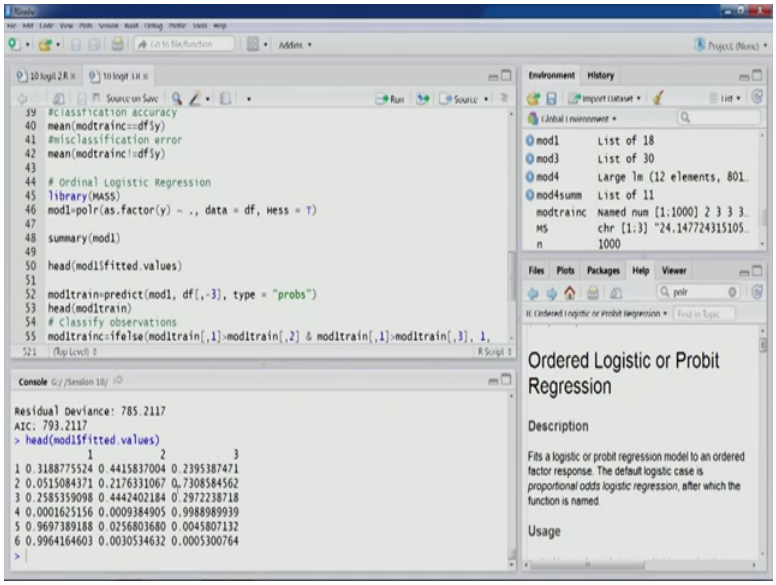
So, let us load this particular library and we have polr is the function for this. So, this is actually for probit however, it can also be used for it can also be used for the cumulative logit method. So, we will just see in the in the in the help section. So, you can see polr this is ordered logistic, or probit regression right.

So, what we are interested in ordered logistic regression, so so here ordered logistic regression. So, you can see in the method argument the first one is draw logistic this is actually ordered logistic method. This is also called proportional odds logistic regression which we have discussed right.

So, let us go and build this model so polr is the function. So, first as you can see y variable I have converted into a factor variable and then the request against all other variables that which are predictors data is full, then we need this argument as well as which is mainly if we want to apply for summary function later on the model object, so, this is also true.

So, now let us apply summary, so we will get the results we can you can see the coefficient values here x 1, and x 2, the value and the error and the T values are also there. And we have the residual deviance, and AIC value as well for this particular model and we are interested in finding the fitted values so that is also returned by the model. So, let us look at some of these values.

(Refer Slide Time: 62:00)



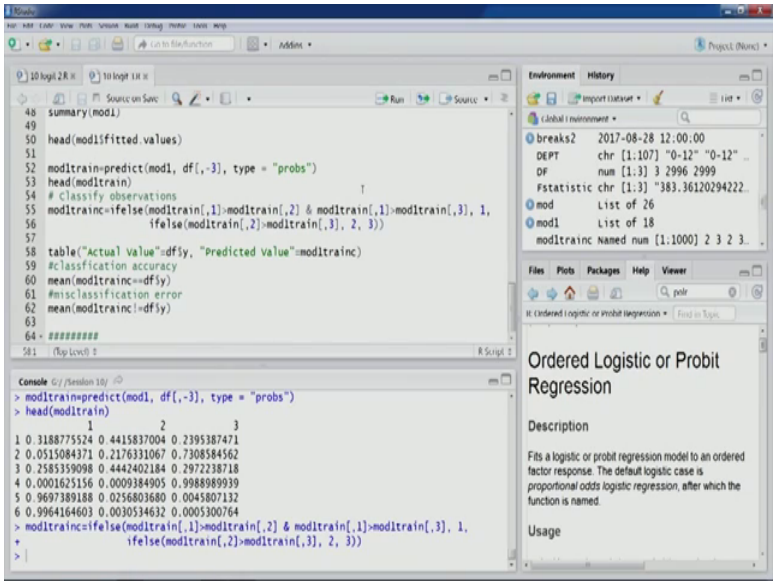
So, you can see for each of these classes class 1, 2, and 3, so these values are actually probability estimated probabilities values. Now using these values we can again apply the most probable class method so first we need to compute the first we need to compute and do the assignment as per the most probable class method like we did in previous exercise.

So, as I talked about predict function can again be used to score new data. So, in this case we are scoring the training partition itself again. So, we expect to get the same values as you can see last row, you can see here, here also the same values are there. So, it is scoring off for the training partition itself. So, we will get the same observation.

Now, what we will do classify observation. So, as we did in previous exercise mod 1 train for class 1 and probability value for class 1 greater than value class 2, and greater than class 3. Then assign it to class 1 otherwise again we look on more comparison the probability value for class 2 is greater than probability value class 3, then assign it to

class 2, otherwise class 3. So, in this way we will have the appropriate classification scores.

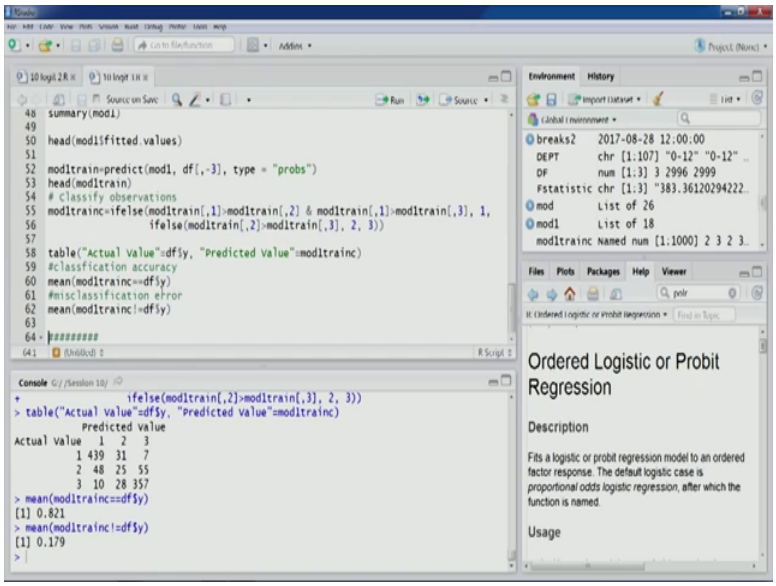
(Refer Slide Time: 63:17)



Now, let us generate the classification matrix here. so you can see 3 by 3 matrix we have 3 classes actual values 3 possibilities, predicted values, 3 predicted classes, so again diagonal elements they represent the correct classification values and off diagonal elements they represent the incorrect classifications.

So, what we can do is so let us compute a classification accuracy, and you can see eighty two percent and the remaining is error.

(Refer Slide Time: 63:49)



So, with this we have completed our discussion on logistic regression and so today we have been also able to cover the scenarios where more than 2 classes are present in our categorical variable. What happens when the classes are nominal and how we can apply logistic regression when classes are ordinal, so we have seen that we have also done an exercise in R.

So, we stop here and we will continue our discussions in next structure for a new technique.

Thank you.