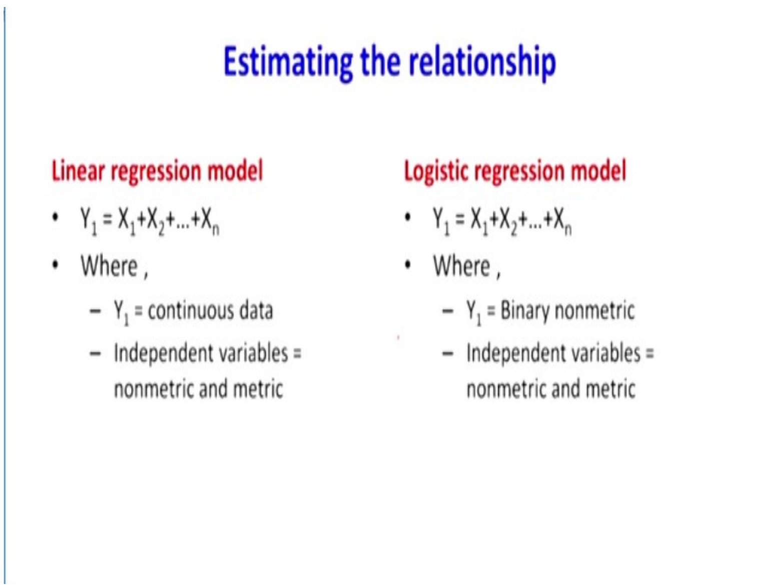
Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Management Studies
Indian Institute of Technology - Roorkee

Lecture – 40
Linear Regressions Model VS Logistic Regression Model

In this class, we are going to compare logistic regression versus linear regression because it is very important to understand how this linear regression and logistic regressions are connected. If you understand the relationship between this linear and logistic regression, it is easy to interpret the meaning of logistic regression. So, agenda of this class is comparison of linear regression model and logistic regression model.

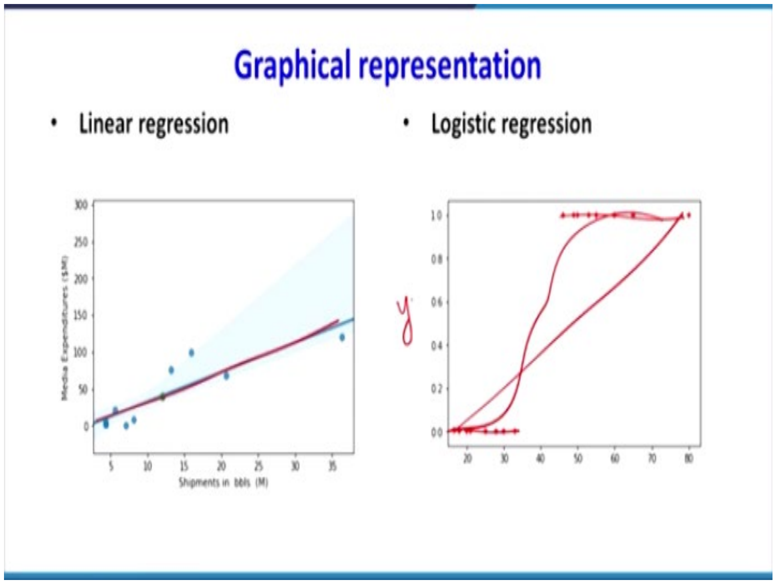
(Refer Slide Time: 00:56)



We will see the first relationship, first difference, estimating the relationship, when you look at the linear regression model, we used to write Y1 equal to X1 + X2 + and so on Xn, where Y1 is a continuous data that is dependent variable, X1, X2 are independent variable, this independent variable it can be continuous we can call it as metric, otherwise it may be a discrete, we can call it as non-metric.

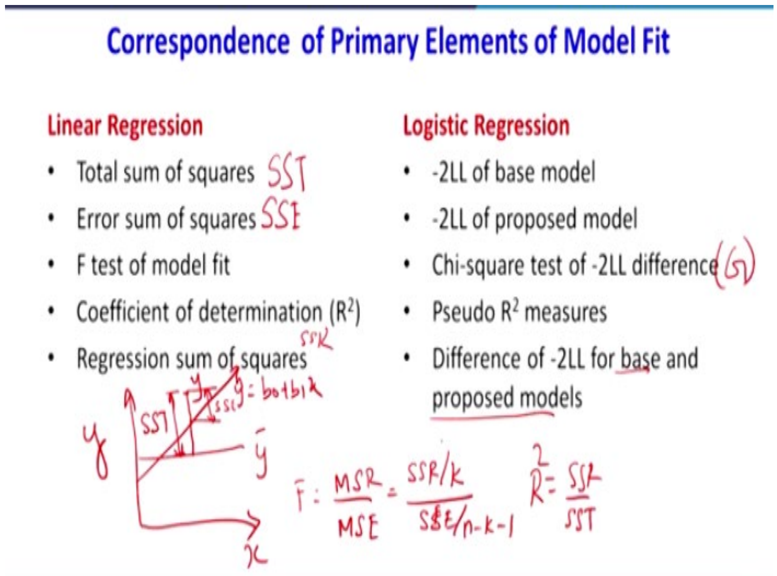
The linear regression model, if the independent variable is discrete variable, we can use the concept of dummy variable regression, whereas in logistic regression model, the general model is Y1 equal to X1 + X2 to Xn, where the Y1 is a binary variable, it is a nonmetric binary variable. The independent variable can be continuous or discrete, we can call it other way; it may be a metric or nonmetric that it is a basic difference between linear regression model and logistic regression model.

(Refer Slide Time: 02:02)



The another difference is if you plot a simple independent variable and the dependent variable for a linear regression, you may able to connect all the points this way but in logistic regression, value of Y can be only 2 possibilities may be 0 or 1, you may get this kind of relationship. What is a meaning here is you cannot form a linear relationship this way, you have to form a; a kind of an S shaped curve that is another difference between linear and logistic regression. What is that in the y axis, see 0 to 300 for linear for example, here you see that it is a possibility, not only that the y value is nothing but the probability but here their y value is the actual values.

(Refer Slide Time: 02:53)



Correspondence of primary elements of model fit between linear regressions, logistic regression, we used to write SST; total sum of square that is for linear regression, the equivalent term for logistic regression is – 2LL that is the log likelihood of base model. Here, we have to write SSE; error sum of square, the equivalent term in the logistic regression is – 2LL of proposed model. I have explained, what is the meaning of log likelihood in my previous lectures.

In this lecture also, I will show you the software output where we can get it this log likelihood value. We know that in a simple linear regression, the meaning of SST is like this, say this is y bar, this is y, suppose a line goes like this, this is our predicted value b0 + b1x, this is our x value, this is our y value. So, this distance was our SST, the equivalent value in the logistic regression is – 2LL.

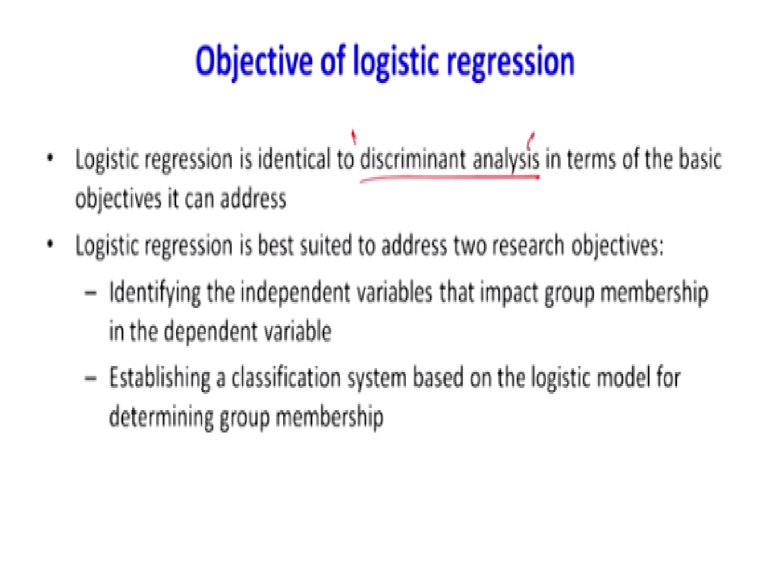
Similarly, we have seen SSE, this unexplained this length, this length in the regression equation is SSE; error sum of square are unexplained variance portions, in the linear regression model to test the overall fit; model fit, we have used F test. There what was the F test in the linear regression model, F is MSR divided by MSE, what is MSR; MSR is SSR divided by k; number of independent variable, divided by SSE n – k – 1.

Sometimes, some books they use k, some books they use p to explain the number of independent variables that is a F value. The equivalent test in logistic regression is chi square test of – 2LL difference that value is nothing but your G, in the linear regression to explain the model fit, the goodness of the model the term used is coefficient of determination, R square. So, what is R square?

R square is SSR divided by SST, regression sum of square divided by total sum of square, otherwise explained variance divided by total variance. The equivalent term in logistic regression is pseudo R square; I will explain what is the formula for finding pseudo R square in coming slides. Here, we use SSR; SSR is regression sum of square, the equivalent term for logistic regression is difference of – 2LL for base and proposed model.

I will explain the meaning of base and proposed model, base means when there is no independent variable corresponding log likelihood values called base value, when you introduce any 1 independent variable, after introduction of independent variable, the corresponding log likelihood values called model; model log likelihood, I will explain this detail in coming slides.

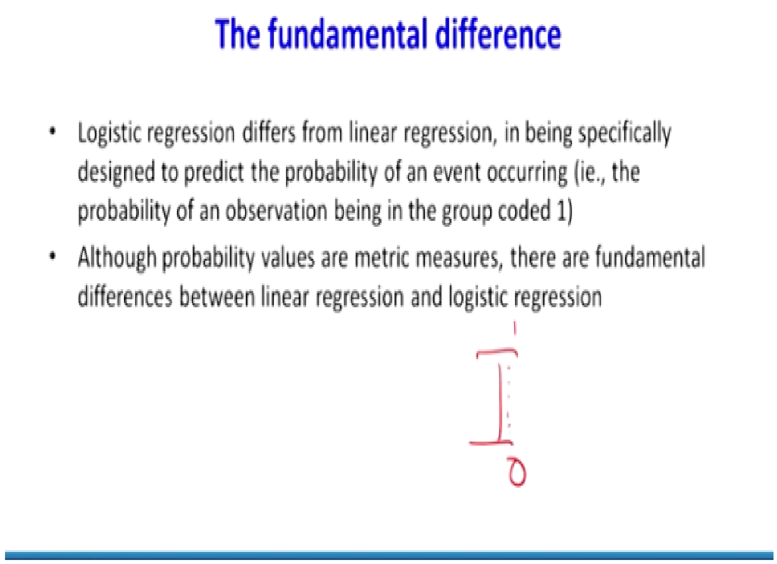
(Refer Slide Time: 06:54)



Then with respect to objective of logistic regression, how the linear and logistic regression differs. Logistic regression is identical to discriminant analysis in terms of basic objectives it can address. There is a one technique called discriminant analysis, the basic difference is in logistic regression, we had only 2 levels; 0 or 1 but in the discriminant analysis, we can have more than 1 level that case is called discriminant analysis.

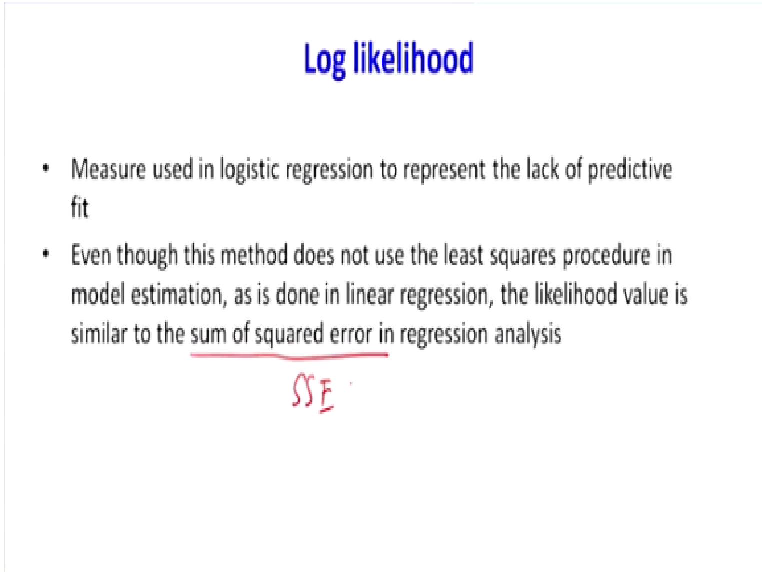
This I did not cover it but this is the concept behind of discriminant analysis, so logistic regression is identical to discriminant analysis in terms of basic objective it can address; still we go for logistic regression. If there are 2 category, we can go for discriminant analysis also but still we prefer logistic regression because it is best suited to address 2 research objectives, one is identifying the independent variables that impact group membership in the dependent variable. Another one is establishing a classification system based on the logistic model for determining group membership.

(Refer Slide Time: 08:18)



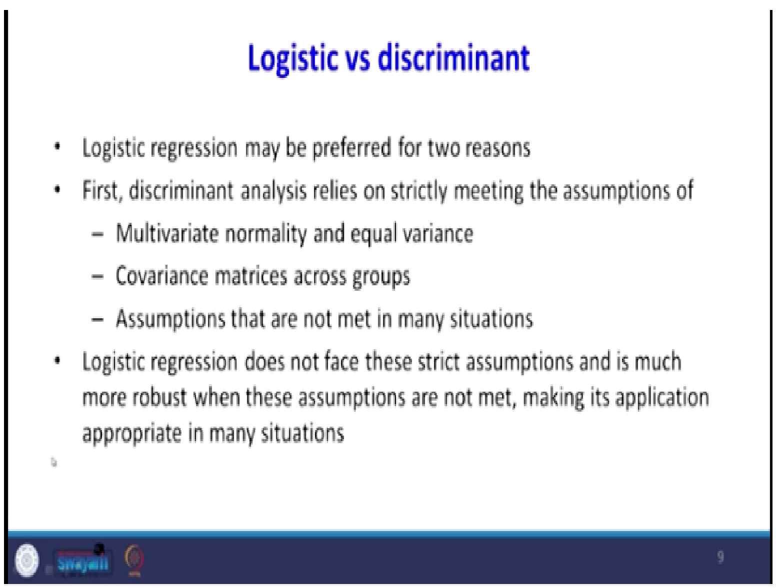
There are some more reason is there, the fundamental difference between logistic and linear regression is; logistic regression differs from linear regression in being specifically designed to predict the probability of an event occurring, so the y value is nothing but the probability that is the probability of observation being group coded 1 or not, although the probability values are metric measures, there are fundamental differences between linear and logistic regression. Even though, we can say the probability value is metric, so there is a 2 possibility, it may be 0 or 1, so we may get different probability, when we go for logistic regression.

(Refer Slide Time: 09:02)



Then, log likelihood; measures used to logistic regression to represent lack of predictive fit, so the log likelihood is used to measure how much lack of fit is there, even though this method does not use in least square procedure in model estimation as is done in linear regression, the likelihood value is similar to sum of squared error. If the log likelihood value is lesser, it is better because in the regression, we try to have sum of squared error SSE lesser it is better. Similar to that in logistic regression, if you are getting smaller value of log likelihood it is better that is a good model.

(Refer Slide Time: 09:52)



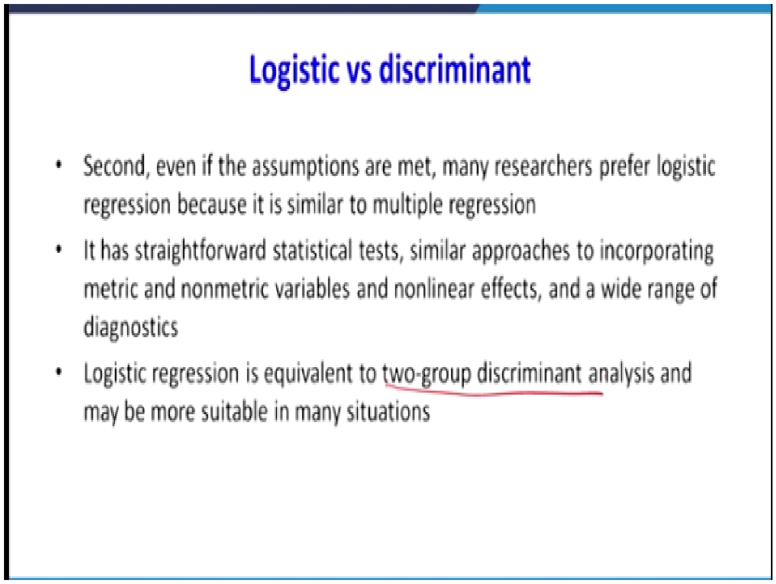
Now, we will compare when should we go for logistic regression, when should we go for discriminant analysis, even 2 slides before also, I have explained comparison between logistic and discriminant analysis. Discriminant analysis we can have more than 2 levels, 3 levels or 4 levels. The problem related to logistic regression can be solved with the help of discriminant analysis, where there are 2 levels.

We can say logistic regression is a special case of discriminant analysis but we would not go for discriminant analysis, we will go for logistic regression, there are some reason is there. See that the logistic regression may be preferred for 2 reasons. First; discriminant analysis relies on strictly meeting the assumption of multivariate normality and equal variance that is the first assumption for the discriminant analysis.

That means, the data has to follow normality and it has to have equal variance and the covariance matrices across groups is necessary, when we go for discriminant analysis. Assumptions that are not met in many situations, a real time problems we cannot have this assumptions, in that situation whenever there is a 2 level in the dependent variable instead of going for discriminant analysis, we can go for logistic regression.

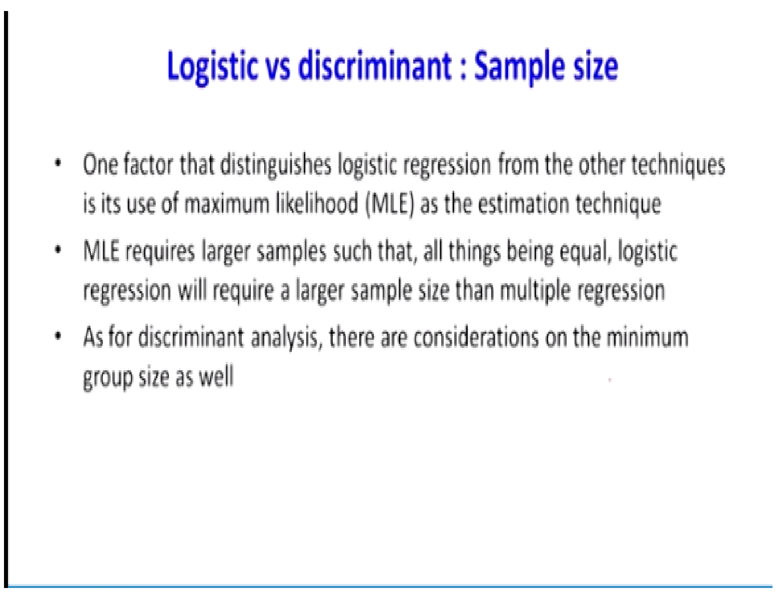
Because these assumptions are not required for logistic regression otherwise, we can say logistic regression is more robust than discriminant analysis when there is an only 2 category in the dependent variable. The next point; the logistic regression does not face these strict assumption and is much more robust when these assumptions are not met, making its application appropriate in many situations, that is why we are going for logistic regression over discriminant analysis.

(Refer Slide Time: 11:56)



Another point is even though, the assumptions are met, many researchers prefer logistic regression because it is similar to multiple regression, many possibility to interpret the result, it has straightforward statistical test, similar approaches to incorporating metric and nonmetric variables and nonlinear effects and a wide range of diagnostics. Logistic regression is equivalent to 2 groups; this point which I am trying to say, 2 group discriminant analysis may be more suitable in many situations.

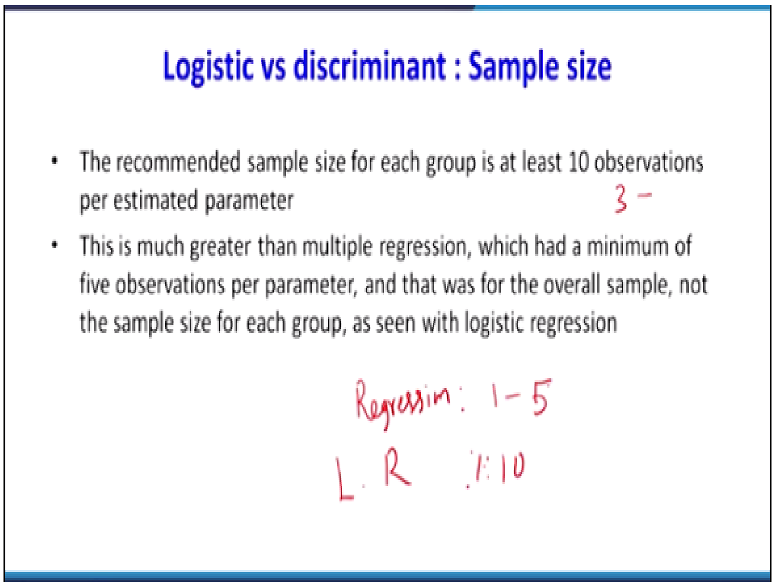
(Refer Slide Time: 12:38)



With respect to sample size which is better; logistic or discriminant analysis, one factor that distinguishes logistic regression from the other techniques is its use of maximum likelihood as the estimation technique. Maximum likelihood estimation requires larger sample such that all things being equal, logistic regression will require a larger sample size than multiple regression, as of discriminant analysis there are considerations on minimum group size as well.

The another when we go for logistic regression, one point is that you need to have large sample size because it follow maximum likelihood estimate because the value of maximum likelihood estimate is sensitive to the sample size or degrees of freedom.

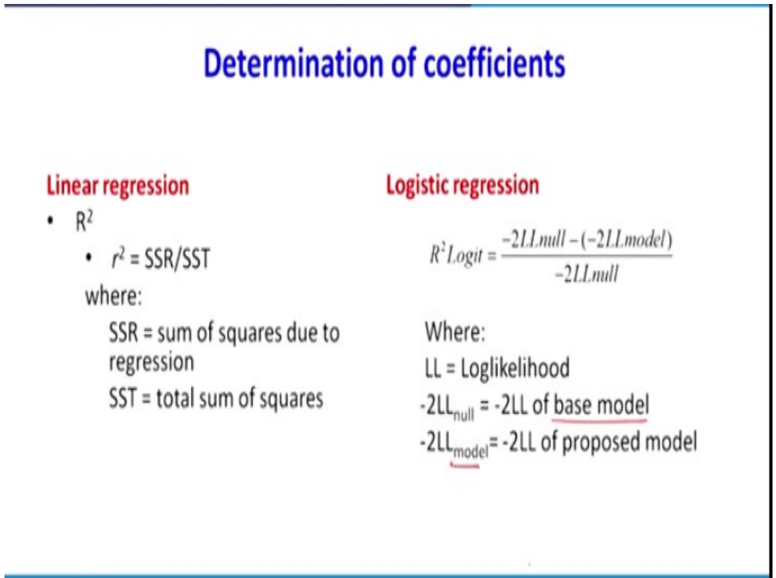
(Refer Slide Time: 13:30)



The recommended sample size for each group is at least 10 observations per estimated parameter, when we go for logistic regression that means, if you are capturing 1 variable, you need to have 20 observations. If you are capturing 3 variables, you have to have 30 observations, this is the thumb rule; this is much greater than multiple regressions which had minimum 5 observations per parameter.

That was for the overall sample not the sample size of each group as seen in the logistic regression, so what the point here is if it is multiple regression for 1 variable, you need to have, you can have 5 respondent rated this regression when you go for logistic regression in to help for one variable in the top 10 respondent where it is regression. When you go for logistic regression, you need to have, for 1 variable, you need to have 10 respondent.

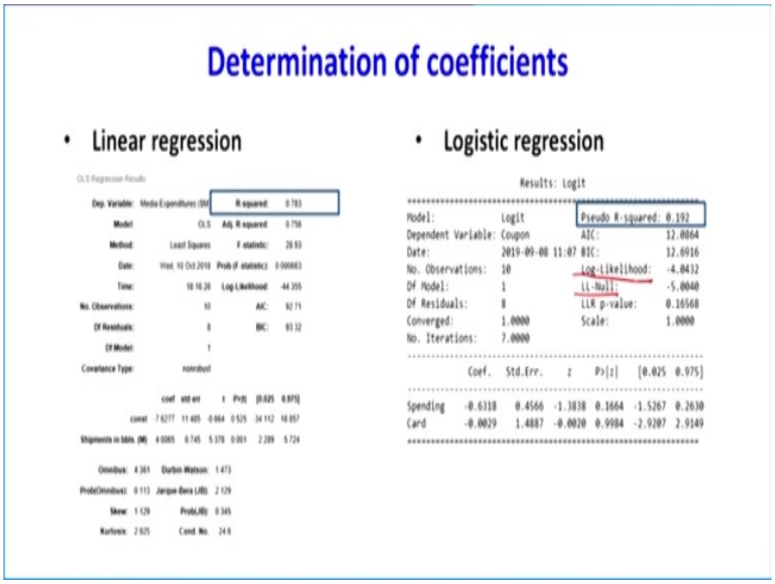
(Refer Slide Time: 14:32)



Then the goodness of fit of the both linear and logistic regression, as I told you the linear regression, the R square that is a coefficient of determination is measured by SSR divided by SST, regression sum of square divided by total sum of square. The equivalent term in logistic regression is pseudo R square, otherwise R square logit equal to – 2LL for null model, - 2 log likelihood minus of – 2LL log likelihood for model divided by - 2 log likelihood for null.

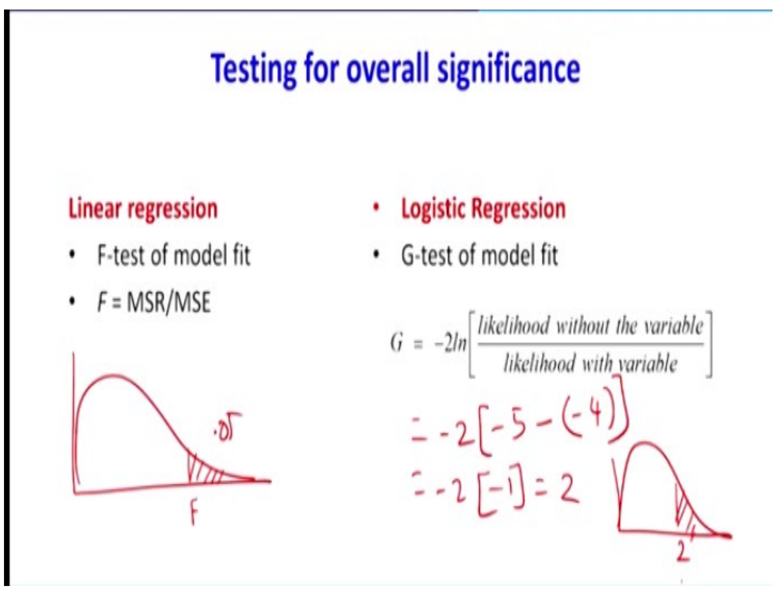
This value I will show what is the null LL and model LL, so LL is likelihood, if I say – 2LL of log likelihood of base model without any independent variable, - 2LL model is meaning here is for model means, it is a proposed model, when you bring a new dependent variable into the logistic regression model that time what was the corresponding log likelihood value that is called this model value for log likelihood.

(Refer Slide Time: 15:54)



As I told you, you see that when you go for I have run this one previously, see here it is R square 0.783, here it is pseudo R square and as I told you in the previous class, you see that here there is a LL null, when there is no independent variable corresponding the value of likelihood is this much, this is null model, this is the base model, so this is your model likelihood, so we have to find the difference of these 2 to get the R square.

(Refer Slide Time: 16:29)

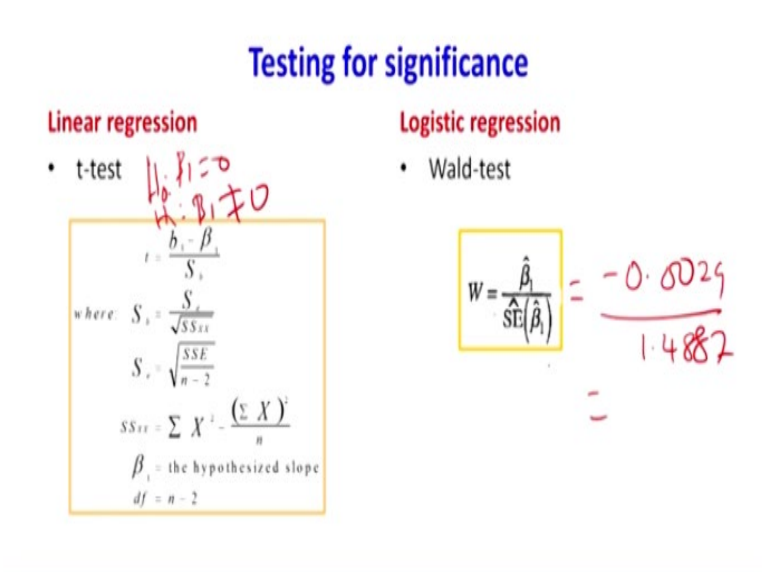


Now, we will see how to test the overall significance of linear regression and logistic regression, to test the overall significance of a linear regression we know that the F value is nothing but MSR divided by MSE, mean regression sum of square divided by mean error sum of square, then what we will do; we will, the error of distribution will follow like this, we will find out what is the say, suppose this is 0.05, we will get corresponding F value, this is our calculated F value.

With the calculated F value is lying on that side will reject null hypothesis, if it is lying on acceptance say, you will accept it but for logistic regression, the formula is – 2LL likelihood without the variable divided by likelihood with variable, this value can be find out, see – 2, when we say log value, it is division is nothing but subtraction, so likelihood without the variable is - 5 minus; this minus for the log of division, likelihood with variable is your – 4. When we simplify - 5 + 4, so we will get – 1, so + 2.

This G also follow chi square distribution; chis square is a right skewed distribution, this is your G value, G value is 2 for the degrees of freedom is number of independent variable in a logistical regression, that number of independent variable is nothing but the degrees of freedom for the G value, this value you can get it from this is output of Python of logistic regression and the linear regression, this was the comparison.

(Refer Slide Time: 18:40)

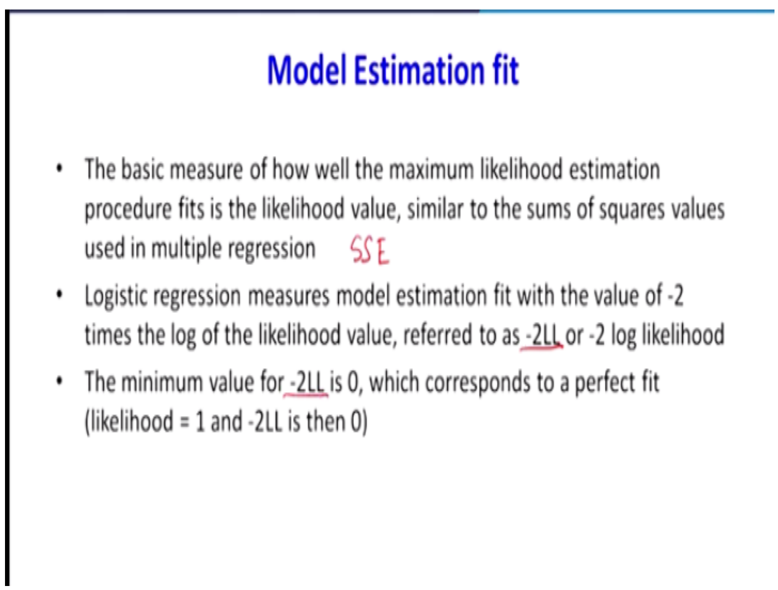


The another point here is to test the significance of each independent variable in a linear regression model, will use t test, the t test; the calculated t test is b1 – beta 1, here the assumption was beta 1 equal to 0 that was our null hypothesis, H1 is beta 1 not equal to 0, then we will find out t value, then we look for see n - 2 degrees of freedom, n – k – 2 degrees of freedom, then we will compare it, whether it lying on the acceptance in the rejection site.

The equivalent test in the logistic regression to test the individual significance of each independent variable we should go for this test called Wald test, this Wald test is nothing but estimated beta 1 divided by standard error of a beta 1, so here you go back, see here the for example, the card; the estimated beta 1 is - 0.0029, the standard error is this value is standard error 1.4887, so this is equivalent to your z value.

So, this z value will be used to test whether the model is the individual independent variable is significant or not, this z value. This z value you got it, this dividing – 0.0027 by 1.480, this value is nothing but your Wald statistic.

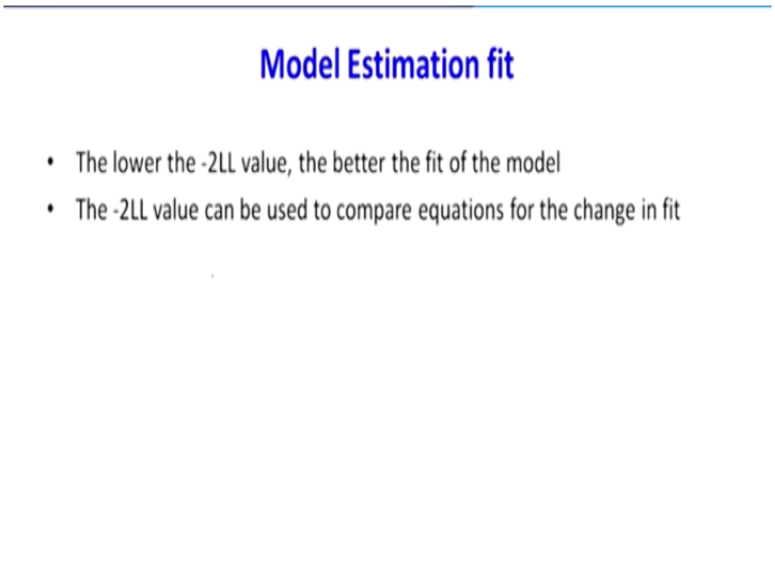
(Refer Slide Time: 20:36)



Model estimation fit; the basic measure of how well the maximum likelihood estimation procedure fits is the likelihood value, similar to the sum of square value used in the multiple regressions, it is equivalent to your SSE. What will happen in multiple regression if the value of SSE is low, it is a good model, the same way logistic regression measures model estimation fit with the value of – 2 times log of likelihood value referred to as – 2LL.

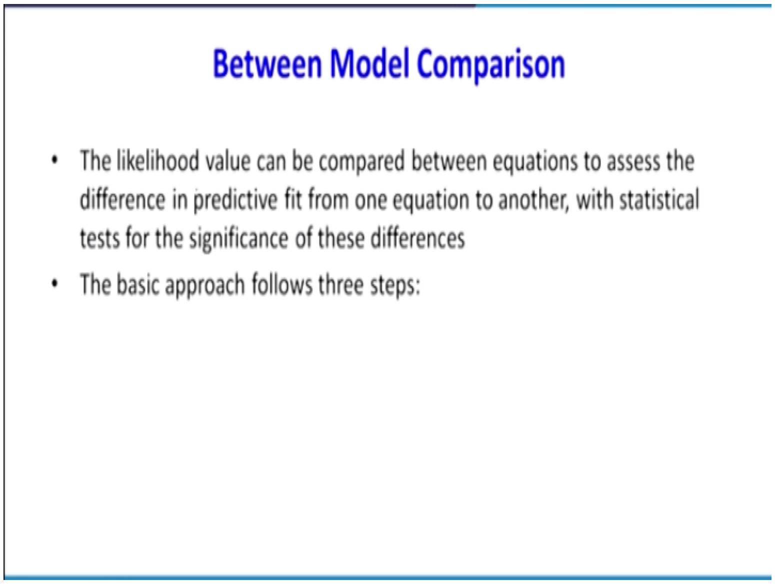
The minimum value of – 2LL is 0, the similar to the SSE equal to 0, which corresponds to a perfect fit, so in the linear regression it is SSE, in the logistic regression it is – 2LL, always we prefer lower is better.

(Refer Slide Time: 21:28)



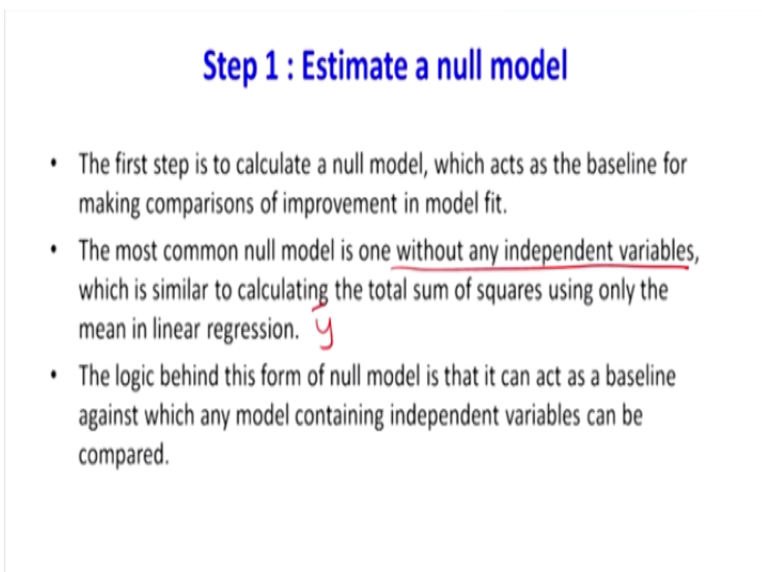
The lower the – 2LL value, the better fit the model is, the – 2LL value can be used to compare equations for change in fit. So, what will happen; first we have to run this model without any independent variable, we have to get what is – 2LL, then we have to introduce another independent variable, then we have to compare how much error term is there. If it is lesser, then the variable which have included is explaining the model in better way that is a meaning of comparing equations for change in fit.

(Refer Slide Time: 22:07)



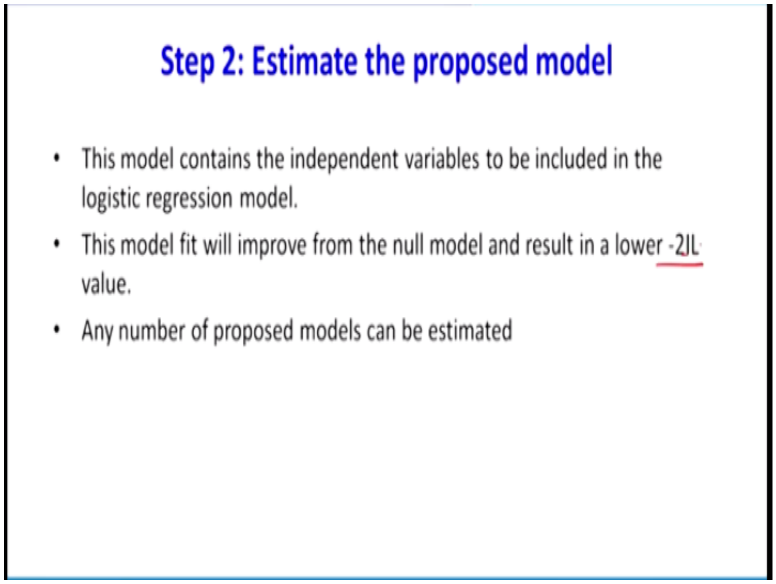
As I told you between model comparison, the likelihood value can be compared between equations to assess the difference in predictive fit from one equation to another with a statistical test for significance of these differences. There are 3 step is there that to assess whether the model is fit or not after introducing a new variable.

(Refer Slide Time: 22:32)



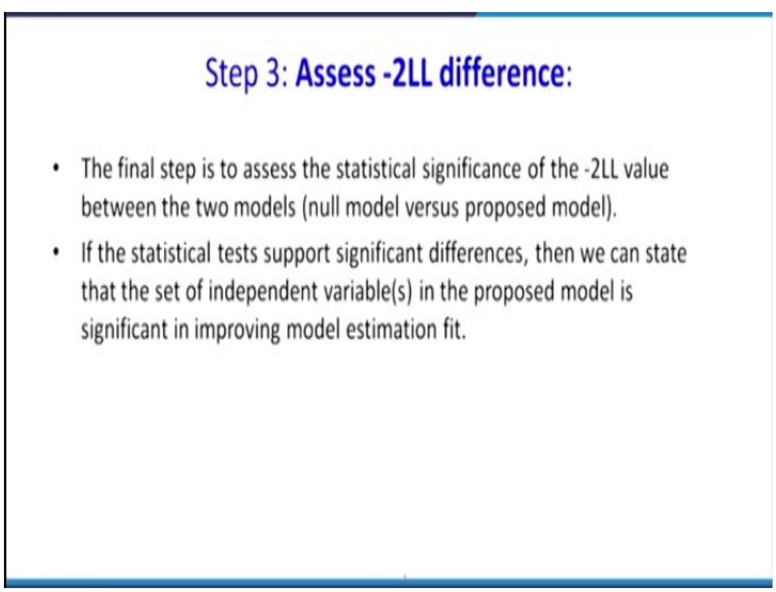
The first step is; we have to estimate the null model, what is null model? The first step is to calculate a null model, which act as a baseline for making comparison of improvement in the model fit. The most common null model is one without any independent variables which is similar to calculating the total sum of square using only the mean linear regression, it is like you know y bar. The logic behind this form of null model is that it can act as a baseline against which any model containing independent variable can be compared.

(Refer Slide Time: 23:16)



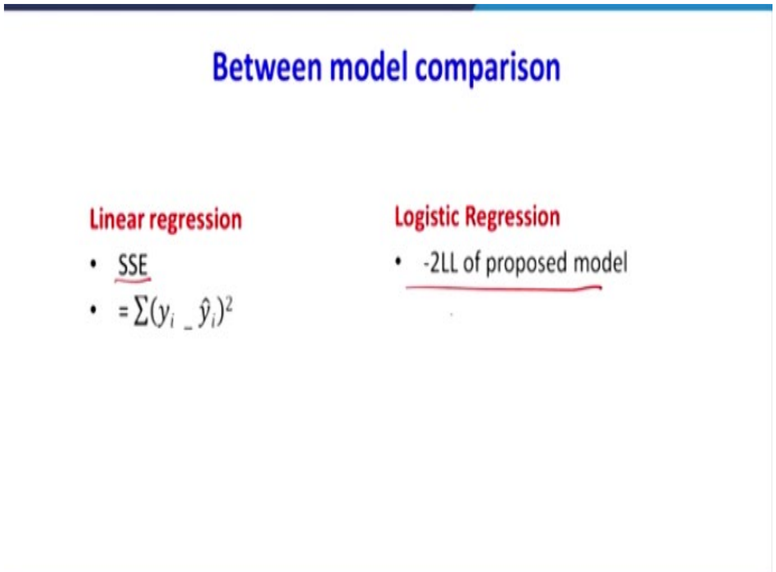
Step 2 is estimate the proposed model, this model contains the independent variables to be included in the logistic regression model, this model fit will improve from the null model and result in lower – 2LL value, if the after including a new independent variable, if the value of – 2LL low, then the model is good model, any number of proposed model can be estimated this way.

(Refer Slide Time: 23:47)



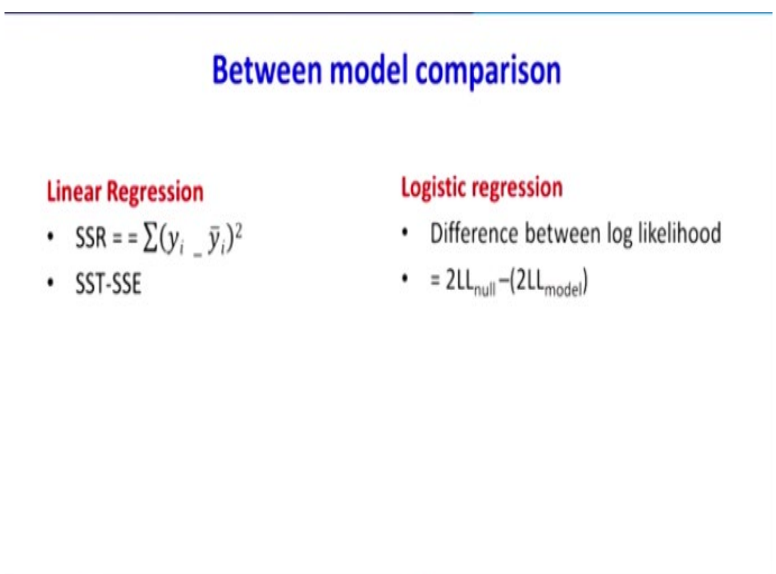
The third step is assessing the – 2LL difference, the final step is to assess the statistical significance of the – 2LL value between 2 models that is null model versus proposed model, if the statistical tests support significant differences, then we can state that the set of independent variables in the proposed model is significant in improving the model estimation fit.

(Refer Slide Time: 24:20)



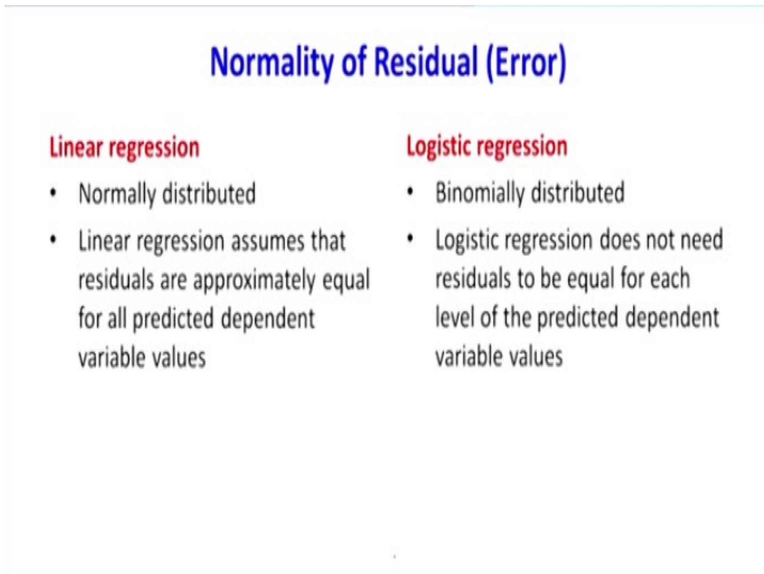
Another difference between logistic and linear regression is SSE, this also I have explained in my previous slide. In linear regression we say SSE, in logistic regression we say – 2LL of a proposed model, if it is lesser then model is good.

(Refer Slide Time: 24:40)



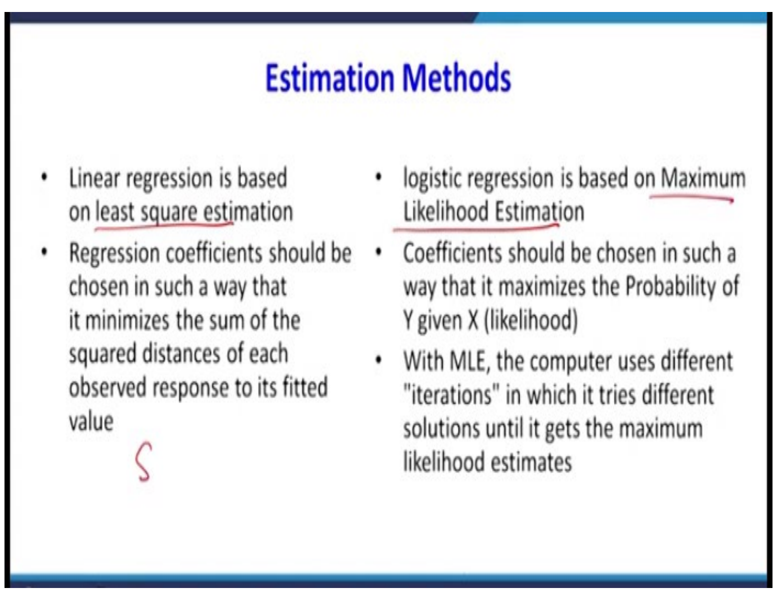
In the linear regression we say SSR; in the logistic regression we say difference between log likelihood of null model and the model after introducing the 1 independent variable.

(Refer Slide Time: 24:54)



Another important assumption between linear and logistic regression is we can say difference with respect to error is; a linear regression model the error term follow normal distribution but in a logistic regression, the error term follow binomial distribution. Linear regression assumes that the residuals are approximately equal for all predicted dependent variable values. Logistic regression does not need residuals to be equal for each level of predicted dependent variables.

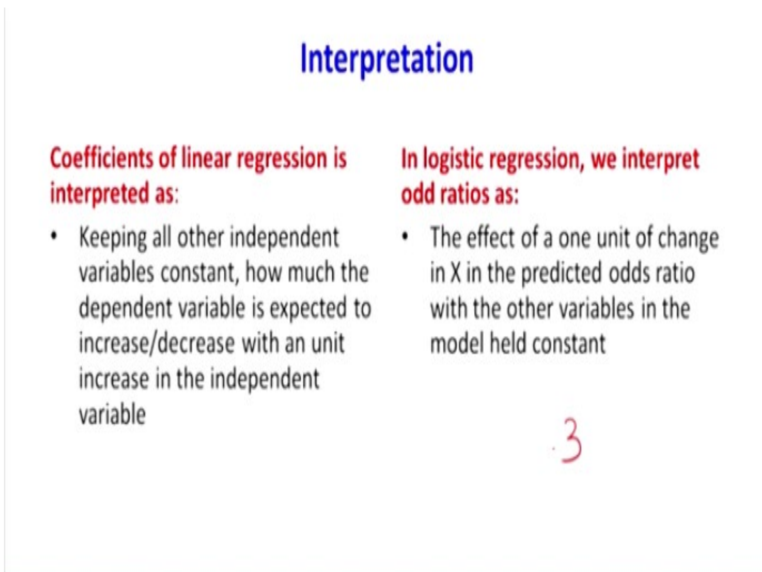
(Refer Slide Time: 25:26)



Another important difference is linear regression is based on the least square estimation via less but the logistic regression is based on the maximum likelihood estimation, this should be our first point. Regression coefficient should be chosen in such a way that it minimises the sum of the square distance of each observed responses to its fitted value, nothing but the error sum of square has to be minimised.

But here, the coefficient should be chosen in such a way that it maximises the probability of y given x, with the maximum likelihood estimation, the computer uses different iterations in which it tries to different solutions until it gets the maximum likelihood estimations, that is how many time solving logistic regression with the help of hand is very difficult.

(Refer Slide Time: 26:25)



Another difference between logistic and linear regression is the way we interpret the coefficient; this is a very important difference. In logistic regression keeping all other independent variable constant, how much the dependent variable is expected to increase or decrease with an unit increase in the independent variable, this is the way we interpret the meaning of coefficient of each independent variable in a linear regression model.

But in a logistic regression model, the effect of 1 unit change in the X in the predicted odd ratio with respect to other variables in the model held constant, the point here is that the coefficient in the logistic regression is explained with the help of odd ratio, suppose the odd ratio is 3, if the odd ratio is 3, if there is a 1 unit increase in the independent variable, there is a 3 times more chance, the probability will be increased.

This is the way to interpret the coefficient of logistic regression, in this class I have explained the difference between logistic regression and linear regression model, there are many parameters I have compared the equivalent term for logistic regression, equivalent means with respect to linear regression. If you are so thorough on interpreting the linear regression model after listening to this lecture, you can interpret the logistic regression model in a very easy manner. Thank you very much.