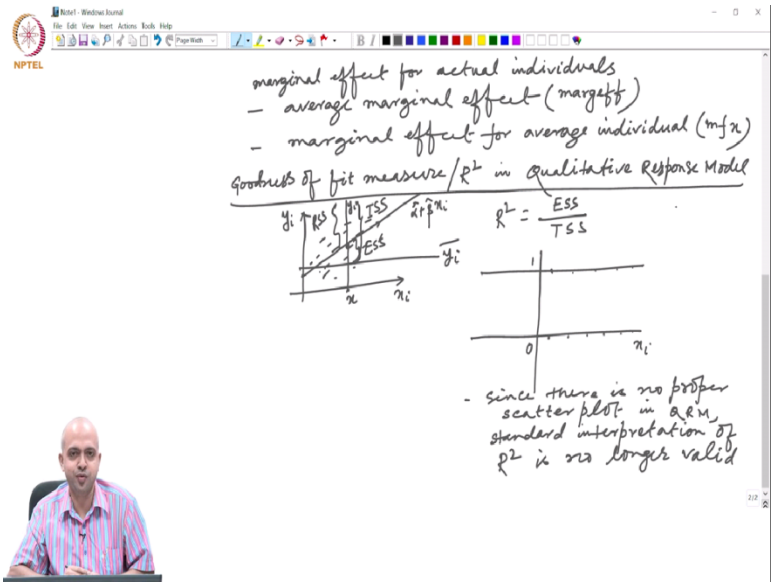
Introduction to Econometrics
Professor. Sabuj Kumar Mandal
Department of Humanities and Social Sciences
Indian Institute of Technology, Madras
Qualitative Response Models- Probit and Tobit Models Part - 1

(Refer Slide Time: 00:15)



So, let us discuss something about goodness of fit measure or R square in qualitative response model. Now, you have to keep in mind that the standard interpretation of R square which we derived earlier in case of linear regression model is no longer applicable in qualitative response model. So, standard R square makes little sense in the context of qualitative response model.

Why this is so? Because, in the case of linear regression model, what is the R square measure? Let us say this is your xi, this is your yi and let us say you have this type of scatter plot and the regression line alpha hat plus beta hat xi is basically a representation of your data. That means, this is some kind of average line which represents these data points.

And what is your R square? Your R squared is for any given value of xi, what is your predicted value? What is your observed value? Let us say this is your observed and let us say this is my yi bar. So, this is my total variation TSS, and out of which, I have explained this much, this is y I had, this is ESS and this is RSS and your R square you have defined as ESS by TSS.

So, out of your total variation in y which is measured by yi summation yi by y whole square total variation and how much you are explained, summation yi hat minus yi bar and then your R

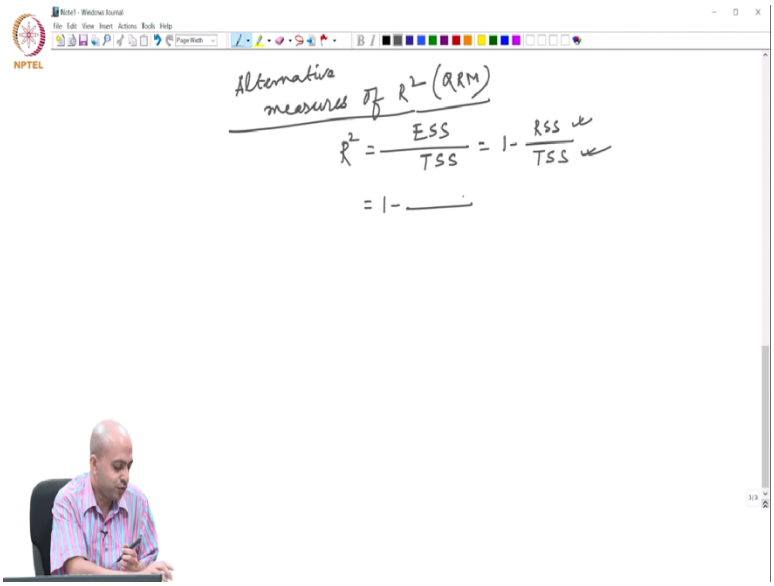
squared is ESS by TSS. So, when you have a scatter plot like, so that means given xi your yi can take any value and you have this proper scatter plot and you have a regression line, which basically represents the scatter plot. And what is goodness of fit? How good my regression line is in representing this scatter plot, which is nothing but raw data.

But suppose we are in a context where this scatter plot itself does not exist. Then what you will do? That means this regression line itself makes little sense. we cannot get this type of regression line which we call average line or average representation of the scatter plot, then how will you take this RSS, ESS by TSS, it does not make any sense. And that is what is happening in the context of qualitative response model. How? Let us see here.

In this axis, let us say I am measuring xi and this is yi and this is 0, this is 1. So, that means for any value of x, you are either here or here that means your probability is here or here, either the event will happen or not happen. So, either the event will happen or not happen, depending on your x value, let us say that is income. So, that means there is no scatter plot and if that is the case, if the scatter plot does not exist in the context of qualitative response model, standard R square interpretation also is no longer valid.

So, what I am saying, since there is no proper scatter plot in qualitative response model, standard interpretation of R square is no longer valid because I cannot get a line like this. And unless I get that type of regression line, I cannot apply the ESS by TSS formula to get the R square. So, what we need to know, we need to derive alternative methods of R square and let us see what type of alternative model we can derive.

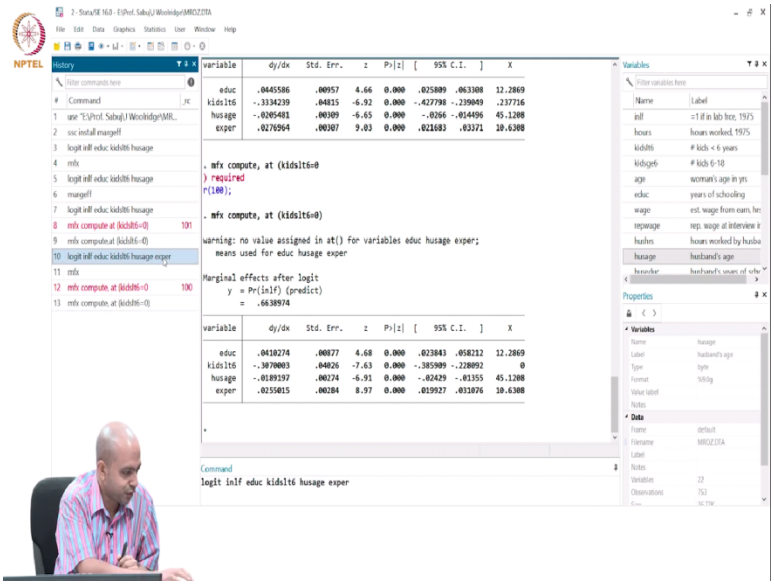
(Refer Slide Time: 06:40)

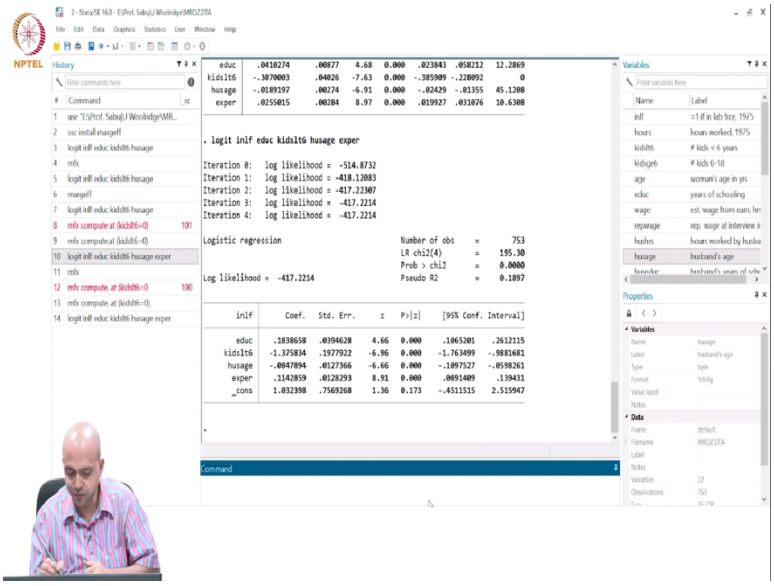


So, alternative measures of R square in the context of qualitative response model. So, how will you modify this? Let us say that this is R square equals to ESS by TSS and that we can write as 1 minus RSS by TSS. Now, in the context of QRM, we need to get some kind of equivalent measure of RSS and TSS. And econometrician say that it is qualitative response model, then you have incorporated certain explanatory variable and your model is explaining some part of the total variation in y and the remaining portion is RSS.

In qualitative response model when you are estimating your coefficient using log likelihood method they go for several rounds of iteration to maximize different log likelihood functions. So, the first step in that iteration is when you have not incorporated any explanatory variable and if you do not incorporate any explanatory variable, then you see if you go to Stata's output, look at how Stata is estimating the model.

(Refer Slide Time: 08:47)



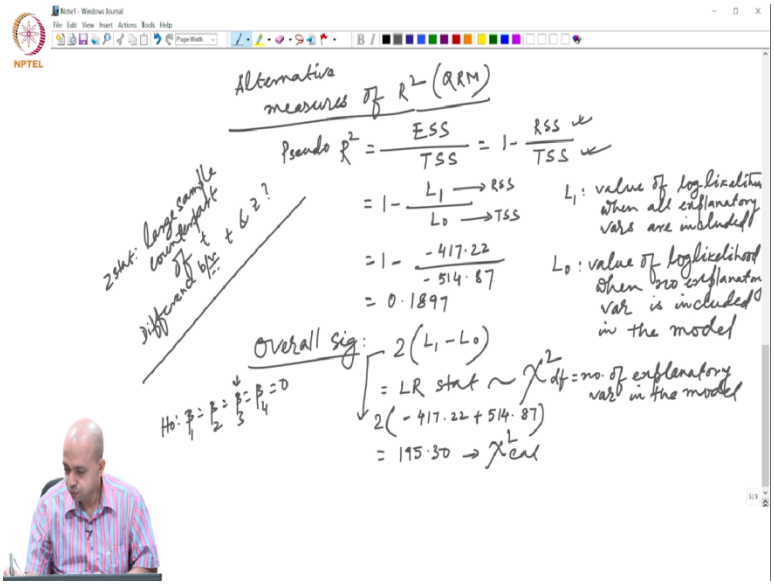


If you put the Logit command, let us say this is my model. Now, see, this is iteration 0, 1, 2, 3, 4 and this is the final. That means iteration 0 means, in your log likelihood function, you have no explanatory variable only intercept and that is your log likelihood value. And what is iteration 4, you have included all your explanatory variables in the model. Similarly, iteration 1 means 1 explanatory variable, 2 means 2 explanatory variable, 3 explanatory variable.

So, this is your initial log likelihood value, this is your final log likelihood value reported here, -417.2214. So, out of these 4, 5 levels of iteration and 5 log likelihood value something represent

your RSS, something represent your TSS and we need to identify which one should be my TSS and which is my RSS. Obviously, when you have no explanatory variable in the model there is no question of explaining anything, there is no question of remaining anything that is the reason the log likelihood corresponding to iteration zero in Stata is reported as L0 and the final one is called L1.

(Refer Slide Time: 10:37)



So, this is defined as L0 and L1, when you have no explanatory variable that is L0, that means, that time it is L0 and this is L1. So, that means, this is your RSS; this is TSS. L0 means zero-ith level of iteration in terms of Stata's outcome and L0 is -514.87. minus 514 equals to minus 514.87 or something divided by, I made a mistake I think, this is, what is L1? L1 is this. Since both are minus, minus, minus will get cancelled. 417.22 and then this is 514.87. So, this would become your R square.

So, if you calculate this, then you will get your R square. And that R square is basically the R square reported here, look at this, this is called pseudo R square which is 0.1897. So, since a standard measure of R square is not applicable in the context of qualitative response model, econometricians have derived an alternative measure of R square, which is this ESS by 1 minus l1 divided by L0.

L1 is the value of log likelihood when all explanatory variables are included and L0 is the value of log likelihood when no explanatory variables are included in the model. This is one thing you have to keep in mind.

Then how do you check overall significance of the model? So, in standard econometric model we used to do by F statistic. But here, we get a different measure. If you estimate the model and then as I said you will get L1 and L0 and L0 is the initial value of log likelihood and L1 is the final value of log likelihood when you include all the variables in your model.

And if you calculate this (2 *(L1-L0)), then you will get an equivalent measure of F statistic in the context of qualitative response model that is called log likelihood ratio, LR chi square or likelihood ratio statistic. And that actually follows as chi square distribution where degrees of freedom equals to number of explanatory variables in the model. So, this is also known as LR chi square. So, log likelihood ratio statistic.

So, that means calculated value of chi square is nothing but 190.30. And again, you have to compare this calculated chi square with the tabulated chi square at a specific level of significance and depending on whether this calculated value is lower or higher than the tabulated value, you have to reject or do not reject your null hypothesis.

And what is your null hypothesis here? In the context of overall significance as we all know, this is basically beta 1 equals to beta 2 equals to beta 3 equals to beta 4 equals to 0. This is the null. So overall significance of the model you will get by LR chi square statistic given by Stata. Look at now, LR chi square is 195.30 with 4 degrees of freedom and Stata is also reporting the corresponding v value which is 0.000 which means LR chi square is highly significant at 1 percent level. And this shows that your model is overall significant.

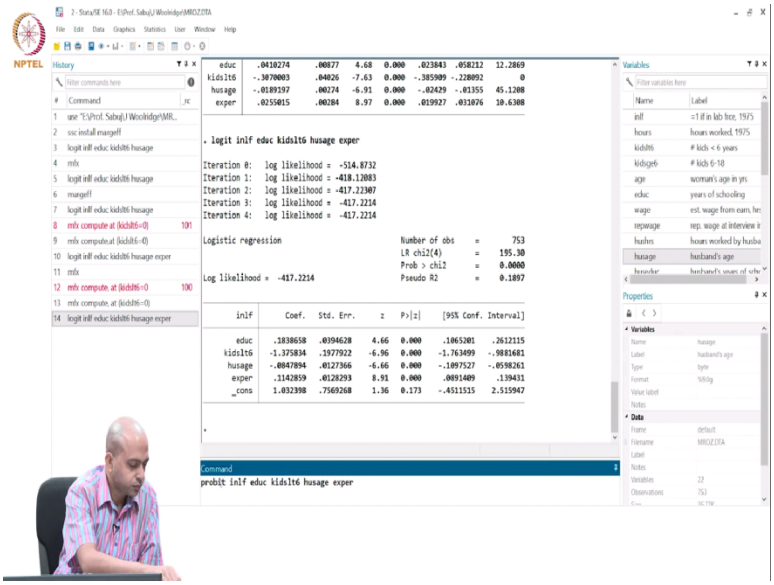
And another difference between the standard model and qualitative response model is you measure the significance of individual variable instead of t here we are getting z. What is z basically? z stat is the large sample counterpart of t. As I told you earlier that in the context of qualitative response model and the Logit model, the estimation technique is MLE, maximum likelihood estimation, which requires large sample.

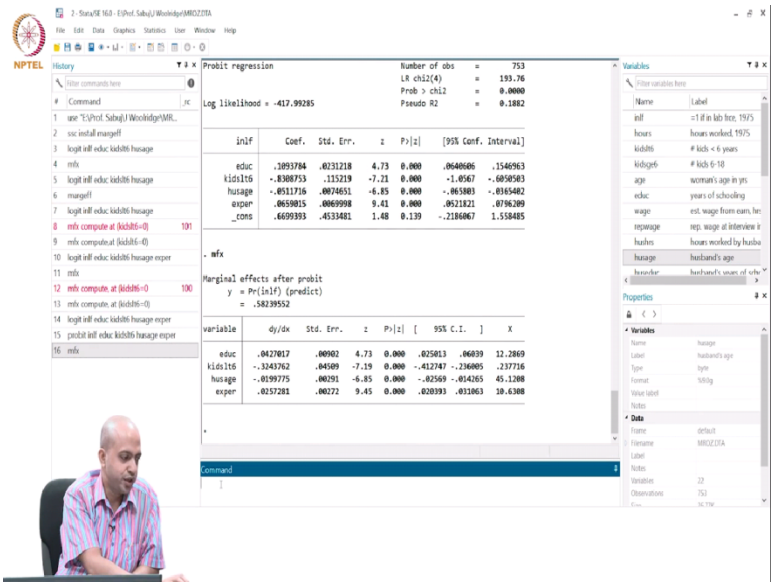
Because in the large sample only all these desirable properties that means, estimates are asymptotically efficient, they are consistent and efficient at the large sample that is why MLE

requires a large sample and the large sample counterpart of t is actually the z statistic. There are other differences of t and z statistic which probably you can look at.

So, what Stata is reporting is the large sample counterpart of the t to test the individual level significance of the explanatory variable. So, this is how you can estimate the Logit model and you can interpret. Same way you can estimate the model using another qualitative response model that we have discussed earlier, which is called Probit.

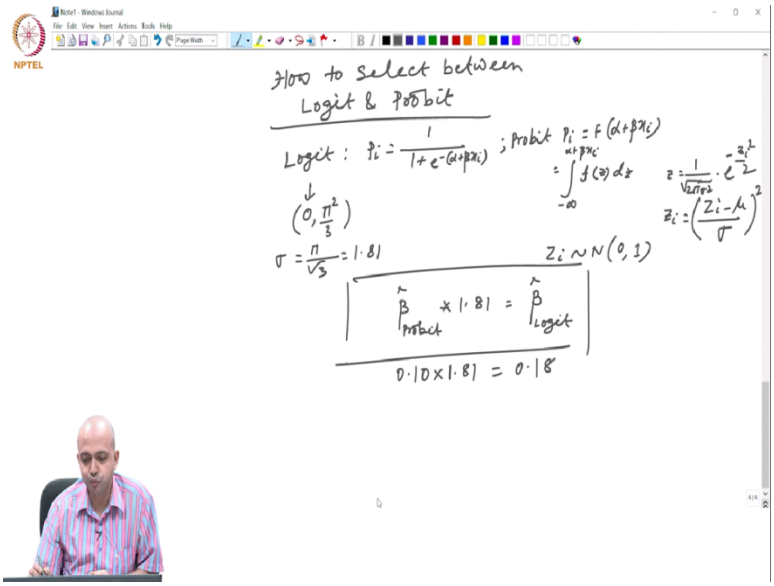
(Refer Slide Time: 21:35)





So, that means, instead of Logit you can put Probit also and you will get this. Again mfx command if you put, you will get the estimated value. So, now, if you compare the Logit and Probit model, look at the coefficient here, education coefficient is 0.10 in the context of Probit. But that was 0.18 in the context of Logit. So, that means, the coefficients of Probit model and the coefficients of Logit model they are not directly comparable. But the question is how to choose between Logit and Probit?

(Refer Slide Time: 22:47)



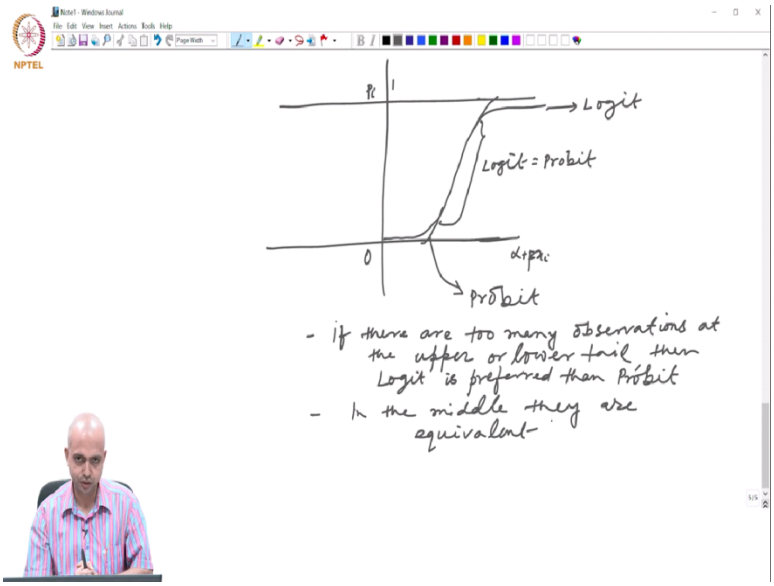
How to select between Logit and Probit? So, in Logit, once again, if you look at what we assume, Pi equals to 1 by 1 plus e to the power minus alpha plus beta xi. And in the context of Probit, Probit Pi equals to f(alpha plus beta xi). And this f(alpha plus beta xi) assumes the cumulative distribution function of the logistic distribution, that means fi equals to integration minus infinity to alpha plus beta xi and then, alpha plus beta xi fz.dz and z equals to 1 by root 2 pi sigma square into e to the power minus zi square by 2 and this zi equals to zi minus mu by sigma whole square, the standard normal variable.

Now both normal and the logistic distribution have zero mean because here we are assuming standard normal variable. So, mean is 0 but their variance and standard deviation are different. So zi follows a normal distribution with 0 mean and variance 1. And here, the zi follows in the context of logistic distribution, the mean is again 0 the variance actually pi square by 3, it is not 1.

Because of the difference in their variance, here it is 1, here it is pi square by 3, where pi equals 22 by 7, so that means here it is sigma equals to pi by root 3 equals to 1.81 or something. So, if you multiply the Probit equation by 1.81, that means beta hat Probit multiplied by 1.81 equals to beta hat Logit. If you look at your model what is the Probit estimate? Probit estimate is for a particular variable 0.10 and here, it is 0.10 here it is 0.18.

So, Probit estimate is point 0.10 multiplied that by 1.81 equals to 0.18. So, Probit estimate is 0.10 and if you multiply that by 1.81 then you will get the Logit estimate and that is the idea. Now the question is how do you select between Probit and Logit model?

(Refer Slide Time: 27:59)



Compared to Probit, Logit has a flatter tail. What does it mean? That means if there are too many observations at the upper or lower tail then from the diagram itself it is very clear that Logit is a better representation of that sigmoid ace or non-linear relationship than the Probit because Logit it is a little sour. And in the middle range Logit equals to Probit, then Logit is preferred than Probit.

Many times, we use Logit and Probit exchangeabley. And we think that we may use either Logit or Probit it does not matter. But the question is what is your data set? If in your sample you see for example, in the car owner or house owner relationship, say that there are 90 percent individuals who are actually having car that means you are at the upper end or there are only 10 percent people who are having car are at the lower end, that situation is better captured by Logit model not the Probit one.

But if your data says around 50 percent, 50-50, 40-60 or 45-55 something like that, that means actually you are in the middle range, you have equal number of zeros and ones. And then, we can say that Logit is equivalent to Probit. Most of the cases we are actually in the middle range. That is why we use Logit or Probit alternatively. But actually, that is not the case.

If your sample says that you have too many observations in the tail either upper or lower, then you must use Logit model and not the Probit model. That is why we need to know this type of

diagrammatic representation to know which one to select between Logit and Probit. Logit is preferred than Probit and in the middle they are equivalent.

So, with this, we are basically closing our discussion on Logit and Probit. In our next class for you we will discuss another qualitative response model. Thank you.