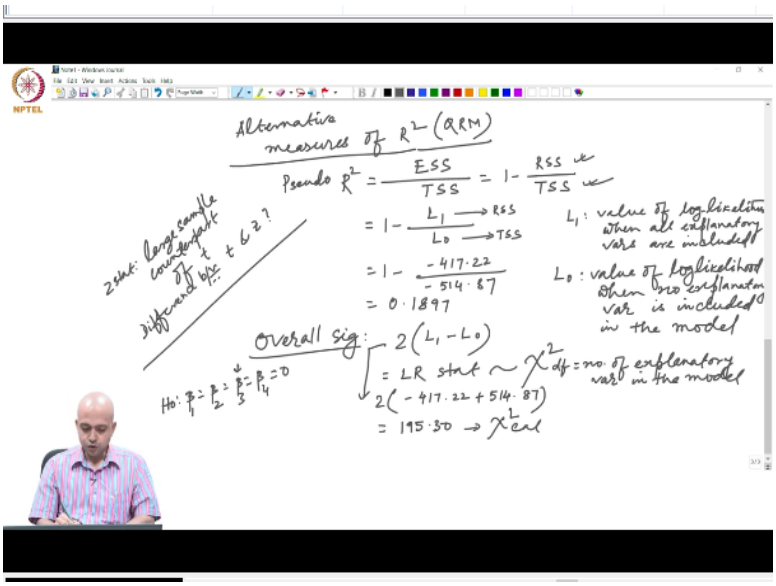
Applied Econometrics
Prof. Sabuj Kumar Mandal
Department of Humanities and Social Sciences
Indian Institute of Technology-Madras

Lecture-33
Qualitative Response Model-Part VI

Let us see what type of alternative model we can derive.

(Refer Slide Time: 00:19)



So, alternative measures of R^2 in the context of qualitative response model. So, how will you modify this? Let us say that this is

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

Now in the context of QRM we need to get some kind of equivalent measure of RSS and TSS. And econometrician they say that if it is qualitative response model then your RSS means basically what 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 what the log likelihood method they do? 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. **(Video Starts: 02:24)**

Look at how stata is estimating the model 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. What is the 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 variable in the model.

Similarly iteration 1 means one explanatory variable, 2 means two expanded variable, three, four, five so on. This is your initial log likelihood value, this is your the final log likelihood value reported here -417.2214. So, out of these five explanatories, five levels of iteration and five log likelihood value something represent your RSS, something represented 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 0 in stata is reported as L_0, and the final one is called L_1. **(Video Ends: 04:13)**

$$R^2 = \frac{L_1}{L_0}$$

So, this is defined as L_0 and L1, when you have when you have no explanatory variable that is L_0. That means that is your that time it is L_0 and this is L1. So, that means this is your RSS, this is TSS. L_0 means 0th level of iteration in terms of stata's outcome if you look at what is L_0? -514.87 = 514.87 or something divided by I made a mistake I think, what is L_1? L_1 is this.

Since both are minus, minus, minus will get canceled, 417.22 and then this is 514.87. So, this would become your R^2. So, if you calculate this then you will get your $R^2$and that $R^2$is basically the R^2reported here look at this, this is called pseudo R^2 which is 0.1897, which is called pseudo R^2. So, since the standard measure of $R^2$is not applicable in the context of qualitative response model.

Econometrician they have derived an alternative measure of R^2which is this ESS by 1 - L_1 divided by L_0, where L_1 is basically what is L_1? I will specifically write here L_1 is the value of log likelihood when all explanatory variables are included and L_0 is value of log likelihood when no explanatory variables are included in the model. This is one thing you have to keep in mind which is called pseudo R^2.

Then another important measure what we need how do you check overall significance of the model? So, in standard econometric model what we used to do? We used to do by F-stat, F statistics but here we get a different measure, if you estimate the model and then as I said you will get L_1 and L_0. And L_0 is the initial value of likelihood; L_1 is the final value of likelihood when you include all the variables in your model.

And if you calculate this

$$LR\ Stat = 2(L1 - L0)$$

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 a chi-square distribution where degrees of freedom equal to number of explanatory variable in the model.

From here what we can calculate is 2 into what is your L_1? L_1 is -417.22 - L_0 and L_0 is what? L_0 is -414. So, this you can put on minus. So, minus and minus this should become +514.87. If you; calculate this then that would become your calculated value of chi square which is nothing but 195.30. This is your chi square calculated 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 lowered 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

$$\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

This is the null. So, overall significance of the model you will get by LR chi square statistic given by stata.

Look at now this is LR chi square 195.30 with 4 degrees of freedom and stata is also reporting the corresponding P value which is 0.000. That means which means LR chi square is highly significant that 1 percent level. And this shows that your model is overall significant and another difference between the standard model and qualitative response model you look at your how do you measure the significance of individual variable instead of T here we are getting Z.

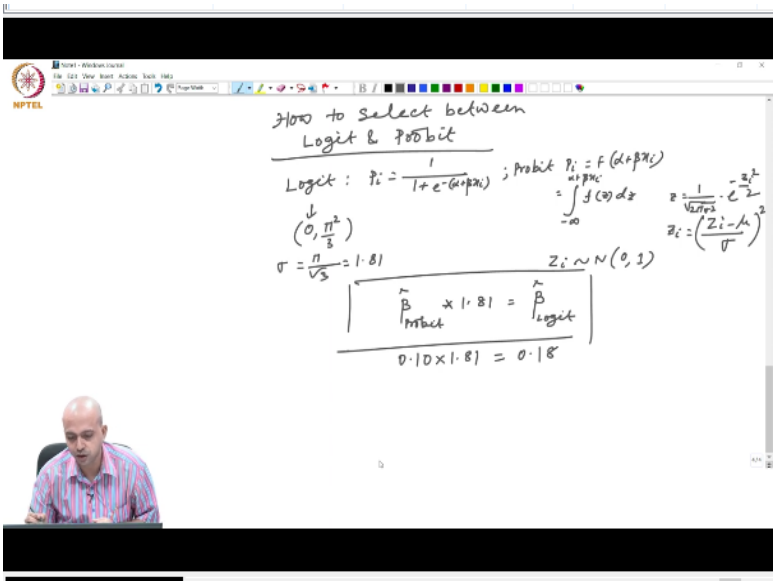
What is Z basically? Z you have to remember that Z stat is the large sample counterpart of T. As I told you earlier that in the context of qualitative response model in the logit model the estimation technique is MLE maximum likelihood estimation which requires large sample because in the large sample only all these desirable properties that means estimates are asymptotically efficient, they are consistent and efficient at the large sample.

That is why MLA 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 what is the difference between T and Z. 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 actually compute estimate the logit model and you can interpret. **(Video Starts: 15:05)**

Same way you can estimate the model using another qualitative response model that we have discussed earlier which is called probit. 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. **(Video Ends: 16:10)** But the empirical question; how to choose between logit and probit? How to select between logit and probit?

(Refer Slide Time: 16:27)



So, in logit once again if you look at what we assume

logit

$$P_i = \frac{1}{1 + e^{-(\alpha+\beta x_i)}}$$

And in the context of probit

$$P_i = f(\alpha + \beta x_i)$$

assumes the cumulative distribution function of the logistic distribution.

$$= \int_{-\infty}^{\alpha+\beta x_i} f(z)dz$$

and $$z = \frac{1}{\sqrt{2\pi\sigma^2}} . e^{-\frac{z_i^2}{2}}$$

And

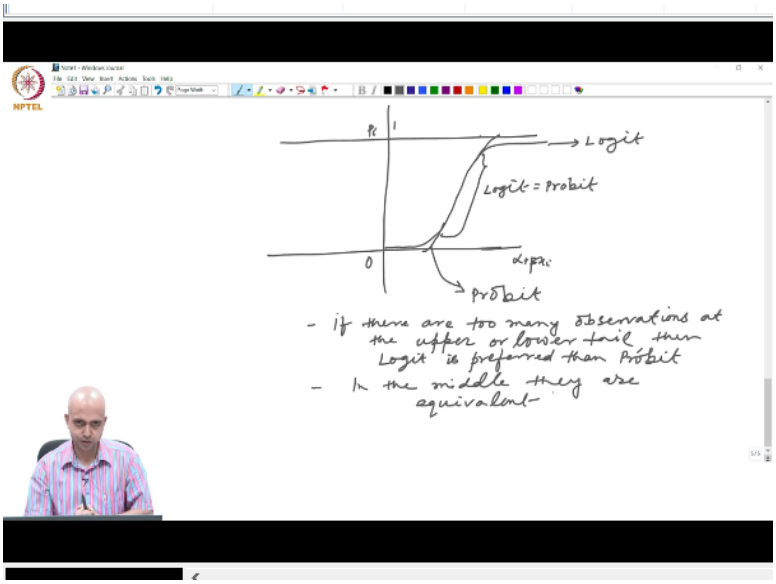
$$Z_i = \left(\frac{Zi - \mu}{\sigma}\right)^2$$

Now both this normal and the logistic distribution they have 0 mean because here we are assuming standard normal variable. So, mean is 0, but there are variance and standard deviation they are different. So, here the mean is 0 in the context of standard normal variable zi.

That means here zi follows a normal distribution with 0 mean and 1 variance and here this zi follows in the context of logistic distribution the mean is again 0, the variance actually $\frac{\pi^2}{3}$, it is not 1. Because of the difference in their variance here it is 1, here it is $\frac{\pi^2}{3}$, where π = 22/7. So, that means here it is $\sigma = \frac{\pi}{\sqrt{3}}$ = 1.81 or something. So, if you multiply the probit equation by 1.81.

So, that means what I can say that $\hat{\beta}$ probit multiplied by 1.81 = $\hat{\beta}$ logit. This first of all you have to check. **(Video Starts: 20:26)** If you look at your model what is the probit estimate is for a particular variable 0.10 and, here it is 0.81. So, probit estimate equals to 0.10 multiplied by 1.81 equals to 0.18. So, probit estimate is 0.10 and if you multiply that by 0.181 then you will get the logit estimate that is the idea. **(Video Ends: 21:24)** Now the question is how do you select between probit and logit model?

(Refer Slide Time: 21:42)



Now for most of the cases this simple diagram will explain which particular this is $\alpha + \beta x_i$, here I am measuring P_i and this is 0, this is 1 and it says this is let us say CDF of logit and this is CDF of probit, this is CDF of normal, this is logistic or I will say that instead of writing all this I will say that let us say this is my logit and this is probit. So, that means compared to probate logit has a flatter tail, what does it mean?

That means if you have too many observations at the upper or lower tail then from the diagram itself it is very clear logit is a better representation of that sigmoid or non-linear relationship than the probit because probit is little short. And in the middle range both logit and probit are same, logit equals to probit.

Then logit is preferred than probit. Many times, we use logit and probit alternatively, 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 owning or not relationship if you sample says 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 you 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 0s and 1s and then we can say that logit is equivalent to probit. Most of the cases we are actually in the middle range.

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 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 that is what we say. In the middle they are equivalent. So, with this we are basically closing our discussion on logit and probit. In our next class or you will discuss another qualitative response model.s