Applied Econometrics

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

Lecture - 31
Course outline for Applied Econometrics

Qualitative Response Model-Part III

So, welcome once again to our discussion of Qualitative Response Model that we are discussing in our last class right. So, we will continue again the qualitative response models from today also. So, in our last class if you recall we discussed basically we started with our discussion with linear response linear probability model and we said that the linear probability model that takes this form p i = α+βxi. So, this is what is this p i? p i is basically probability that y i equals to 1 that is p i. And then we said that this linear probability model or in short LPM what is the major limitation of this? Here the probability is modeled as a linear function of x right linear function of x. So, that means if you think about the house ownership problem that we are discussing in our previous class.

So, what happens actually when the individuals income is very low in that range almost all the people they do not have a house actually. So, at lower income people do not have a house almost all of them and at a higher level of income they will almost all of them will have a house, but then once you achieve that level of income then probability of owning a house that does not change actually. For example, when your income is 1.5 lakhs per month then you have a house.

So, that is 1.5 lakhs right and that probability of owning a house at that income range is almost 1, but suppose now income is increasing from 1.5 lakhs to 2 lakhs then once you have the house then you cannot buy you that particular individual would not buy any new house right. So, that means basically it says once you achieve that level of income the probability does not change it will almost 1 and at lower level of income nobody is having a house and at the lower level of income when your income is let us say 5000 per month to 5500, 5600 like that probability does not change that much. So, at the lower end and at the higher end it is constant and it changes in between.

So, that means a linear characterization of probability is a much problematic thing in this context. So, what we actually want if you plot your probability in this way let us say this is 0 and this is p i this is let us say minus infinity this is plus infinity. So, what we want our probability should be like this it should behave in this way and this is what is called a sigmoid s curve type relationship this is a sigmoid s and to capture this type of non-linearity. So, that means in this axis I am measuring let say α+βxi right it ranges from this to this and and this is equals to z i. So, z i basically ranges from minus infinity to plus infinity.

So, this is z i and what we want is the relationship of z i and p i like this at the lower end it will almost 0, but it will never touch 0 here it is 1 actually it will approach towards 1 at higher level of income, but it will never touch 1. So, basically it asymptotically approaches 1 and 0 and after that suppose from this portion it almost constant here also once you achieve here it almost constant it is not changing and it is changing in this particular this region. So, to overcome the problem of linear characterization of probability with z i in logit model what we assume that p i = 1 - p i. And by $1 + e^{-zi}$ and from here you can understand as z i as this model ensure as z i in ranges from minus infinity to plus infinity then your p i will become 0 to 1 that is the advantage of this model that is the advantage of this logit model. Is it clear? So, I will repeat once again this linear probability model it assumes probability is a linear function of x here x is income linear function of x or you can consider α+βxi entire thing is z.

So, it is a linear characterization between p i and z i, but in reality what happens is that probability does not change linearly when income changes from 15000 to 20000 the change in probability is not same when income changes from 1 lakh to 1 lakh 20000. Probably when income changes from 1 lakh to 1 lakh 20000 you will observe either very very insignificant change in probability of owning a house or no change at all. So, it only changes from 20000 to 1 lakh in that range in this range actually probability changes after that it constant. Similarly at the lower end and to overcome that problem we hypothesize a non-linear characterization of probability of owning a house p i with the income x i and that is basically the logit model which is $(1/1 + \mathrm{e})^{-zi}$. And as z i ranges between from minus infinity to plus infinity p i will range between 0 and 1 that is how logit model overcomes the problem of linear probability model.

But then you end up having a non-linear model p i =$(1/1 + \mathrm{e})^{-zi}$, you cannot estimate directly this model applying the linear technique and that is the reason we characterized that means, we transform the apparently looking non-linear model into a linear model by taking log and then we discussed how to estimate that model using the maximum likelihood estimates or MLE where OLS does not work that is how we discussed about the linear probability model and the logit model. Now, today we will discuss another qualitative response model which also characterize non-linear relationship between the probability and x i and this models name as probit model. So, let us try to understand the theoretical structure of this probit model. Now, to understand the theoretical structure of this probit model we will introduce a variable which is called latent variable. Let us say $yi *= \alpha + \beta xi + ei$ here y i star is called a latent variable which is unobserved and then there is a relationship between y i and y i*.

Now, y i equals to 1 when y i* greater than 0. Now, you might be thinking what is this latent variable and how can you get a relationship between y i and y i *. Think about the house owning problem. Given your income each and every individual calculate some amount of utility of satisfaction of buying a house or buying a car or anything and you will observe that individual has actually bought a house when the individual derives a positive amount of utility. Is not it? A positive amount of utility.

If the utility is negative then that means if there is dissatisfaction of owning a house at that level of income then you will see that individual has actually not bought the house. Now, you might be thinking what is the disutility of owning a house? Actually there is no disutility of owning a house as such, but at that level of income when my income level is very less let us say 10,000 and if I buy a house how buying a house is not my priority at that level of income because I have so many other important things to do. So, if I buy a house and then if I start giving EMI for that house probably that will give me a satisfaction. So, each and every individual will calculate the utility at that level of income of owning a house. Depending on the utility household will decide or the individual will decide whether to buy the house or stay in rented apartment.

But utility is something you cannot observe. What you observe is actually the decision. And what is the decision? Whether I have bought or not that is the realization. So, that is why you cannot observe the utility, but you can observe the decision. Here y i is basically the decision, the ultimate realization whether the event has happened or not.

But in between how and what amount of utility the individual has derived that you cannot observe and that unobserved utility let us say we defined as y i*. It depends on your income, but then there is some amount of error term also which makes the utility unpredictable unobserved. So, when y* > 0 you derive a positive amount of utility and then y i = 1, 0 otherwise. This is the structure of the probit model that y i is related to an unobserved variable y i* which is called latent variable. Now, once you hypothesize that type of relationship between y i and y i* then what you have to do basically when you are calculating probability y i = 1 that means you are saying in turn it is nothing but probability y i* > 0 because then only y i = 1.

Now, from the relationship you can easily understand when can you get y i* > 0. So, from this relationship I can easily understand that y i * will = 0 greater positive when your e i is actually > than –(α+ βxi). From this relationship it is very easy to understand y i* will become greater than 0 when e i is actually > –(α+ βxi). And if you recall the definition of probability density function from the properties of probability density function we can write when e i is actually a random variable and this is less than which is greater tha α+ βxi. then we can say that this is nothing but 1 – (f (α+βxi)) that is how you can that is how you can derive this one. So, what is this (f (α+βxi)) ? This is actually I will say that is cumulative distribution function.

Now depending on what type of specific cumulative distribution function this f (α+βxi)) will take, you will get either linear probability model, logit model or probit model. What I am saying? This f (α+βxi)) can take three different values. It can be a cumulative linear distribution function which is that means I can say that f (α+βxi)) , can be simply be f (α+βxi)) = $(1/1+\mathrm{e}).^{-\alpha+\beta xi}$. And then you will get the logit model. And in the context of probit this f (α+βxi)) takes this type of form equals to and this is called this is actually cumulative cdf of a logistic distribution function.

So, this is basically this is actually f (α+βxi)), I will say that cumulative distribution function or cdf. So, in the context of logit this is cdf of a logistic distribution function,. And in the context of probit, this cumulative distribution function in the context of t, in the context of probit this f (α+βxi)) is actually the cumulative distribution function of a normal distribution. So, that means, this is normal cdf, cdf that means in the context of probit what I can write is that this p i = f (α+βxi)) $\int_{-\infty}^{\alpha+\beta xi} fzdz$. And what is this f z? f z is basically a normal probability density function and I can write that where

f z = (1/$\sqrt{2}$) / 2piσ^{2}*$e^{-zi\hat{}2/2}$.

And what is z i? z i is actually how it is defined? z i is defined in this way (z i – mu/σ) (z i – mu/σ) ^{2} but a standard normal variable. Is this clear? So, that means, here in the context of probit only difference that it makes is f (α+βxi) takes the cumulative since I am taking the integration of this f z which is basically a normal distribution function. I am taking when I am taking integration that becomes the cumulative density function or cdf. So, this is the cdf of a normal distribution function where f z = (1/$\sqrt{2}$) / 2piσ^{2}*$e^{-zi\hat{}2/2}$ and how z i is defined? z i is defined as z i small z i bar minus mu divided by sigma whole square that means z i is basically a standard normal variable. So, if p i equals to this then from here you can say that means α+βxi = $f^{-1}pi$ that is how you can get.

Now if you recall the log likelihood function what we got in the context of logit same type of log likelihood function you will get in the context of probit also that means your log L $\sum yi$* log p i + $\sum(1-yi)*\log(1-pi)$ and that you are trying to maximize with respect to α and β. And this p i what is this p i? p i = $\sum_{i=ni+1}^{n} yi$ summation y i,and then this is log of what is p i? p i is basically f (α+βxi) $\sum(1-yi)$ = log (1 –f(α+βxi)). So, this is your log likelihood function in the context of probit in the context of probit and that you maximize once again with respect to α and β and then you will get your α* and β*. You will get α* and β* by maximizing this. So, that means this is an alternative derivation of the probit model and if you follow then you can derive the logit model also in this way because up to this when p i = f of f (α+βxi) that is same and depending on which particular cumulative density function you will get it will define whether it is a logit model or probit model or linear probability model.

So, f (α+βxi) = α+βxi in the context of linear model, linear probability model f(α+βxi)) = $(1/1+\mathrm{e})$.$^{-\alpha+\beta xi}$ in the context of logit. That means it assumes cumulative density function of a logistic distribution and here it is the cumulative density function of a normal distribution function, where f z = f z = (1/$\sqrt{2}$) / 2pi$\sigma^{2}$$*e^{-zi\text{^}2/2}$$1$ z i is actually standard normal variable defined as and you know the standard normal variable it has 0 mean and 1 = σ^2, . Thank you.