

# Economics 633

## Econometrics

### Discrete Choice Models

Adrian Pagan

April 14, 2004

#### 0.1 Origins and Structure

To date we have been generally assuming in regression and other models that the variable  $y_i$  is a continuous random variable. In fact there is nothing in regression that requires this. The variable could be discrete or it could follow counts. But generally with the latter types of random variables linearity of the conditional expectations is unusual. Thus we need to carefully study the type of data that represents some decision. This data generally comes in binary form with a "1" representing a decision to do something and a "0" being a decision not to do something. Thus we have a variable  $y_i$  that takes these values and a set of variables connected with the decision  $x_i$  and we want to model the relation between them.

Suppose however we had what appears an entirely different type of problem in which we have collected data on whether individuals buy a sports car or not and we have various characteristics of the cars and the individuals who buy them. We could approach this problem theoretically by specifying the utility to be gained from purchasing a "family car" as a function of the characteristics of that type of car  $z_0$  (price, durability, size, low insurance rates, good fuel consumption, etc.) and the characteristics of the individual  $w_i$  (age, sex, married, children etc.)

$$\begin{aligned}u_{i0} &= \alpha_0 + z'_{i0}\delta_0 + w'_i\gamma_0 + \epsilon_{i0} \text{ --- utility from a family car} \\u_{i1} &= \alpha_1 + z'_{i1}\delta_1 + w'_i\gamma_1 + \epsilon_{i1} \text{ --- utility from a sports car}\end{aligned}$$

Note the presence of price in the  $z$ 's means that these are indirect utilities while the  $i$  indicates that the value of a sports car characteristics may differ according to individuals, e.g., the insurance rates depend on the individual as well as the type of car. The  $\epsilon$ 's are random variables which could be due to omitted variables. Essentially the idea is that even for individuals with the same measurable characteristics ( $z$ 's and  $w$ 's) different levels of utility will be

gained owing to say “outlook”. We assume that the errors are *i.i.d.*( $0, \sigma^2$ ). This last assumption may not be correct but it is fairly conventional.

Now when would the individual  $i$  purchase a sports car. The answer is he will choose that action which gives highest utility, i.e.,

$$\text{sports car chosen if } u_{i1} - u_{i0} > 0.$$

Since both the  $u$ 's are random we know that the choice is random so we assume that

$$\begin{aligned} Pr[\text{sports car chosen}] &= Pr[u_{i1} - u_{i0} > 0] \\ &= Pr[\alpha_1 + z'_{i1}\delta_1 + w'_i\gamma_1 - \alpha_0 - z'_{i0}\delta_0 - w'_i\gamma_0 + \epsilon_{i1} - \epsilon_{i0} > 0] \\ &= Pr[x'_i\beta - \epsilon_i > 0] \text{ or } Pr[\epsilon_i < x'_i\beta] \end{aligned}$$

$$\text{where } x_i = \begin{bmatrix} 1 \\ z_{i1} \\ z_{i0} \\ w_i \end{bmatrix}, \beta = \begin{bmatrix} \alpha_1 - \alpha_0 \\ \delta_1 \\ -\delta_0 \\ \gamma_1 - \gamma_0 \end{bmatrix} \text{ and } \epsilon_i = \epsilon_{i0} - \epsilon_{i1}.$$

Now let us suppose we have observed the  $z$ 's,  $w$ 's and also the fact whether a sports car has been purchased or not. Let  $y_i$  be the value zero if a family car is chosen and unity if a sports car. Then

$$Pr[\text{sports car chosen}] = Pr[y_i = 1] = Pr[\epsilon_i < x'_i\beta]$$

and our problem is to estimate  $\beta$  given  $z_i$ ,  $w_i$  and  $y_i$ . To do this we need to make some assumption about the  $\epsilon_i$  and we make these normally distributed. Then since  $\epsilon_{i0}$ ,  $\epsilon_{i1}$  are normal so is  $\epsilon_i$ . It is further assumed that the variance of  $\epsilon_i$  is set to unity, as it is impossible to identify it. This can be seen in a number of ways. One can do it formally by observing that the score for  $\sigma^2$  would be zero for any value. Another is to observe that  $Pr[\epsilon_i < x'_i\beta] = Pr[\sigma^{-1}\epsilon_i < x'_i(\sigma^{-1}\beta)]$ , making the  $\beta$  identifiable only up to a factor of proportionality. Intuitively, the problem arises because the numbers in the data are arbitrary i.e. one could have assigned the values of 1 and 2 instead of 0 and 1 to  $y_i$ , so that it would be possible to produce any range of values in  $y_i$ . Hence

$$Prob[\epsilon_i < x'_i\beta] = \Phi(x'_i\beta) = F_i = \int_{-\infty}^{x'_i\beta} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\lambda^2\right\} d\lambda$$

and this model is called the *Probit Model*.

Let's look at this closely. Although we derived the Probit specification from a utility maximization perspective, we could alternatively have just begun with the proposition that the probability of buying a sports car,  $Pr(y_i = 1) = \Phi(x'_i\beta)$ , is some function of a set of characteristics  $x_i$ . Indeed this is how the earliest specification of these type of models arose. The simplest model was what was referred to as the *linear probability model*, i.e., it just asserted that  $Pr\{y_i = 1\}$  was linearly related to  $x'_i\beta$ , i.e.,

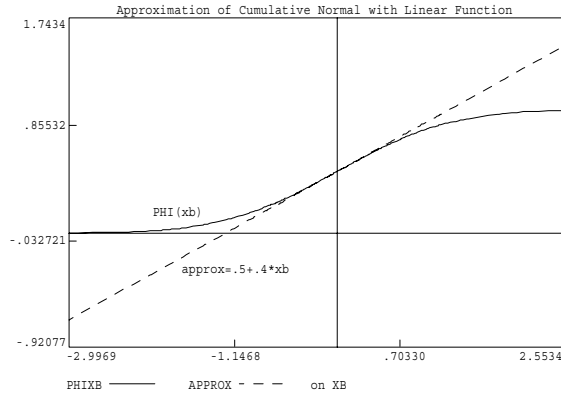


Figure 1:

$$Pr\{y_i = 1\} = x'_i\beta.$$

Obviously one does not get this model from utility maximization very easily, but it is the case that  $\Phi(x'_i\beta)$  and  $x'_i\beta$  are closely related over much of the likely range of  $x'_i\beta$ . The situation can be seen in figure 2. Amemiya notes that the linear probability model  $0.5 + .4(x'_i\beta)$  and  $\Phi(x'_i\beta)$  are reasonably close for values of  $\Phi$  between .3 and .7 so it can be a good representation of  $\Phi$ .

## 1 Estimating Univariate Models

### 1.1 Non-Linear Least Squares

Now let us look a bit closer at models that can be represented in the form

$$Pr\{y_i = 1|x'_i\beta\} = F(x'_i\beta) = F_i$$

forgetting for a moment what  $F_i$  might be i.e. whether it is the form for a logit or probit model or some other discrete choice model (note we have now made explicit the fact that the probability is a function of the scalar or *single index*  $x'_i\beta$ ). As we have already observed  $F_i$  could be  $x'_i\beta$  (linear probability) or  $\Phi$  — the standard cumulative normal distribution function — but we will come across another important choice later. We think of  $y_i$  as a discrete *r.v.* that takes only two values — the value 1 with probability  $F_i$  and the value zero with probability  $1 - F_i$ . Therefore

$$\begin{aligned} E(y_i|x'_i\beta) &= prob(y_i = 0|x'_i\beta)(0) + prob(y_i = 1|x'_i\beta)(1) \\ &= prob(y_i = 1|x'_i\beta) = F_i \end{aligned}$$

Hence we see that the relation

$$\begin{aligned} E(y_i|x'_i\beta) &= F_i \\ \Rightarrow y_i &= F_i + \{y_i - E(y_i)\} = F_i + u_i \end{aligned}$$

and this gives us a non-linear *regression* relation connecting the 0,1 (binary) variable  $y_i$  and the  $F(x'_i\beta)$  term which is the conditional mean, i.e.,  $E(u_i|x'_i\beta) = 0$  as needed for a regression. It is clear therefore why the linear probability model is so popular. If  $F_i = x'_i\beta$  we would have assumed that

$$y_i = x'_i\beta + u_i$$

and we can regress the binary data against the  $x_i$  to get an estimator of  $\beta$  so that the assumption of a linear probability model has great advantages in terms of simple estimation.

What are its drawbacks? First, we are very interested in predicting what an individual with given characteristics ( $z_i, w_i$ ) will do when faced with the family/sports car choice. Since this probability is  $F_i = F(x'_i\beta)$  it is natural to estimate it by  $F(x'_i\hat{\beta})$ . But here is the rub. In the linear probability model there is nothing to guarantee that  $0 < x'_i\hat{\beta} < 1$  and we could therefore get some rather embarrassing probabilities. Note that if  $\hat{\beta}$  were available for the Probit model this could not happen as  $\Phi(x'_i\hat{\beta})$  must lie between 0 and 1 by the definition of the cumulative normal distribution function.

A second problem is connected with the regression itself. Let us look at the conditional variance of  $y_i$  i.e.,  $E[(y_i - x_i\beta)^2|x'_i\beta] = E(u_i^2|x'_i\beta)$ .

$$\begin{aligned} Var &= Pr(y_i = 0|x'_i\beta)[(y_i = 0) - E(y_i|x'_i\beta)]^2 \\ + Pr(y_i &= 1|x'_i\beta)[(y_i = 1) - E(y_i|x'_i\beta)]^2 \\ &= (1 - F_i)(-F_i)^2 + F_i(1 - F_i)^2 \\ &= F_i^2 - F_i^3 + F_i(1 - 2F_i + F_i^2) = F_i - F_i^2 = F_i(1 - F_i) \end{aligned}$$

Thus the error term in this non-linear regression does not have a constant variance and therefore the non-linear least squares estimator needs to be applied with some care. It will clearly be the case that this estimator will not be efficient. For the linear probability model  $F_i = x'_i\beta$ , and the variance will be  $x'_i\beta - (x'_i\beta)^2$ , and so the variance changes with the levels of the regressors. Note however that we know the form of the heteroskedasticity, so that conceptually we can devise an efficient estimator by employing weighted least squares, e.g if we divide all data by  $F_i^{1/2}(1 - F_i)^{1/2}$  we get  $F_i^{-1/2}(1 - F_i)^{-1/2}y_i = F_i^{-1/2}(1 - F_i)^{-1/2}x'_i\beta + F_i^{-1/2}(1 - F_i)^{-1/2}u_i$  and it is clear that the error term in this weighted regression is homoskedastic. Its clear that if we had assumed that  $F_i = x_i\beta$  then we might get into trouble with such a strategy as nothing ensures that  $x'_i\beta - (x'_i\beta)^2 > 0$  and if it is not we would be taking the square root of a negative number.

An alternative is to just ignore the heteroskedasticity for the purpose of estimation and to just allow for it in inferences, i.e., one could compute the covariance matrix of the linear probability model estimator allowing for heteroskedasticity of completely unknown form. Certainly if a weighted least squares solution is not possible, and one is not worried by  $F_i$ 's lying outside the permissible region, one should always do this adjustment. Note that the advantage of the linear probability model is that it can be estimated by a regression program and so it will very likely be the first estimator performed on any binary data, just to get a "feel" for the likely relationships. Hence understanding the need to make an allowance for heteroskedasticity is very important.

Now let's say that we are unsatisfied with the linear probability model. Then we might alternatively estimate the Probit model. Note that from our general treatment above

$$y_i = \Phi(x_i'\beta) + u_i$$

when  $\Phi$  is the Probit function. Thus the probit model is really a special type of non-linear regression model – the non-linearity arising because  $F(x_i'\beta)$  is non-linear in  $\beta$  - with the added complication of heteroskedasticity in the errors. But if we think of it like this we might ask why one should choose  $F(\cdot)$  to be the cumulative standard normal; all that is required for our purposes is that  $0 < F < 1$  and many functions satisfy this, some of which are more tractable numerically than others. Going back to our sports car case, it is a question of selecting a distribution function for the error term  $\epsilon_i$  and there is no reason to just select a cumulative normal. For this reason various other distributions have been canvassed, one of the most common being the specification of  $F$  that is associated with the *Logit* model

$$F(x_i'\beta) = e^{x_i'\beta} / (1 + e^{x_i'\beta}).$$

Thus the Logit model estimates

$$y_i = \frac{e^{x_i'\beta}}{1 + e^{x_i'\beta}} + u_i$$

which is also a non-linear regression model. Unlike the  $F_i = \Phi_i$  of the Probit model which requires look-up tables (or routines) to evaluate  $\Phi_i$  for any value of  $\beta$  and  $x_i$ , the logit model does this very easily by just data transformations. So it is very easy to code and estimate  $\beta$  above via a non-linear regression.

Naturally one might well ask whether one would get greatly different answers if one uses Logit instead of Probit as we have already commented on the close correspondence of the linear probability model with the Probit model over most of the likely range of variation. There seems to be no obvious correspondence, but in fact it is fairly close. Amemiya has found by trial and error that  $e^{\lambda x_i'\beta} / (1 + e^{\lambda x_i'\beta})$  with  $\lambda = 1.6$  gives an excellent correspondence ( $F(\cdot)$  below is for the logit model)

$x'_i\beta$	0	.1	.2	.3	.4	.5	.6	1.0	3.0
$\Phi(x'_i\beta)$	.5	.5398	.5793	.6179	.6554	.6915	.7257	.8413	.9987
$F(\lambda x'_i\beta)$	.5	.5399	.5793	.6177	.6548	.6900	.7231	.8320	.9918

Since  $\lambda = 1.6$  is just a scaling parameter it is clear that, unless there is a lot of the data giving very high probabilities that  $y_i = 1$  (i.e.,  $\Phi \simeq 1$  and hence a lot of  $y_i = 1$  in the data), there is very little difference between the fit of a Logit and a Probit model, and we can establish (numerically) the following approximations:  $\hat{\beta}_L \approx 1.6\hat{\beta}_P$ ,  $\hat{\beta}_{LP} \approx .25\hat{\beta}_L + .5 \approx .4\hat{\beta}_P + .5$ ,  $\hat{\beta}_{LP} \approx .25\hat{\beta}_L \approx .4\hat{\beta}_P$ . Thus it is generally believed that when one only has binary data it is very hard to distinguish between the Logit and Probit fits. Given this fact there is a lot to be said for taking the representation which is the easiest to work with, and in many respects this means the logit model.

As mentioned above, estimating the Logit and Probit models on the same set of data will produce different estimates of  $\beta$ , but this does not mean that estimates of  $E(y_i|x'_i\beta)$  would be different, and it is the latter which is of primary interest. Indeed one needs to think carefully about what we normally want to measure when we estimate these models. Generally we want to measure how  $E(y_i|x'_i\beta)$  varies as  $x_i$  changes. Normally this means we want to evaluate

$$\begin{aligned}
\frac{\partial E(y_i|x'_i\beta)}{\partial x_{ij}} &= \frac{\partial E(y_i|x'_i\beta)}{\partial (x'_{ij}\beta)} \times \frac{\partial (x'_i\beta)}{\partial x_{ij}} \\
&= \frac{\partial E(y_i|x'_i\beta)}{\partial (x'_{ij}\beta)} \times \beta_j \\
&= \frac{\partial F(x'_i\beta)}{\partial (x'_i\beta)} \times \beta_j \\
&= f(x'_i\beta)\beta_j
\end{aligned}$$

where the latter follows from and so, unlike linear regression, we cannot infer the effect of a change in  $x_i$  upon  $\Pr(y_i = 1|x'_i\beta) = E(y_i|x'_i\beta)$  from the coefficients  $\beta$ . Clearly the answer to the question depends upon the level of  $x'_i\beta$  and so we generally need to ask such questions in the context of a "control group" which specifies a given value for  $x_i$  e.g. we might use the average values of  $x_i$ . The situation is made even more complex when the changes in  $x_i$  are not small in which case we really need to evaluate the change as  $F(x_i^{**'}\beta) - F(x_i^{*'}\beta)$ , where  $x_i^{**}$  is the changed value

Let's go back to estimation. Now let us look at using the Gauss-Newton algorithm to solve the non-linear regression problem. Starting with initial values  $\beta_{(0)}$  we linearize  $F_i$  around this value to give

$$\begin{aligned}
F_i &\simeq F(x'_i\beta_{(0)}) + \frac{\partial F}{\partial \beta}(\beta - \beta_{(0)}) \\
&= F(x'_i\beta_{(0)}) + \psi_i(\beta - \beta_{(0)}) \\
\therefore y_i &\simeq F(x'_i\beta_{(0)}) + \psi_i(\beta - \beta_{(0)}) + \text{error}
\end{aligned}$$

and  $y_i - F(x'_i\beta_{(0)})$  as dependent,  $\psi_i$  as independent enables  $\Delta_\beta = \beta - \beta_{(0)}$  to be determined by a regression program. This gives a new value of  $\beta$ ,  $\beta_{(1)} = \beta_{(0)} + \hat{\Delta}_\beta$ , and this is the value linearized around again to give a new  $\beta_{(2)}$  etc.

Now when we look at the error term it is clear that it will be heteroskedastic of the form  $F_i(1 - F_i)$  and therefore *weighted* non-linear regression would seem to be the better estimator, i.e.,  $F_i^{-1/2}(1 - F_i)^{-1/2}(y_i - F(x'_i\beta_{(0)}))$  would be the dependent variable and  $F_i^{-1/2}(1 - F_i)^{-1/2}\psi_i$  would be the independent variable. Clearly this latter estimator will be more efficient (note that  $F_i$  in the first stage will be estimated by  $F(x'_i\beta_{(0)})$ ).

of  $x_i$  and  $x_i^*$  is the original value.

## 1.2 Maximum Likelihood Estimation

We might however wonder if weighted non-linear regression is the most efficient estimator possible, particularly when we look at the error term  $u_i$  and observe that it is certainly not normal. Obviously, the non-linear regression estimator is the quasi-MLE. For this reason one might prefer to go for MLE, and in fact this is what one normally gets from canned programs that estimate the Probit and Logit models. Let's look at this then. We have two types of observations on  $y_i$ . First,  $y_i = 0$  with probability  $1 - F_i$ . Second,  $y_i = 1$  with probability  $F_i$ . Hence, if we let  $I_0$  represent all those values of  $i$  such that  $y_i = 0$ ,  $I_1$  such that  $y_i = 1$ , the likelihood must be

$$\prod_{i \in I_0} (1 - F_i) \prod_{i \in I_1} F_i$$

and the log likelihood is

$$\begin{aligned}
L &= \sum_{i \in I_0} \log(1 - F_i) + \sum_{i \in I_1} \log F_i \\
&= \sum_{i=1}^T (1 - y_i) \log(1 - F_i) + \sum_{i=1}^T y_i \log F_i
\end{aligned}$$

from the values of  $y_i$  in the different sets [Note when  $i$  is such that  $y_i = 1$  the first term disappears, while, if  $y_i = 0$ , the second term drops out.]

$$\begin{aligned}
\therefore L_\beta &= - \sum_{i \in I_0} (1 - F_i)^{-1} \frac{\partial F_i}{\partial \beta} + \sum_{i \in I_1} F_i^{-1} \frac{\partial F_i}{\partial \beta} \\
&= - \sum_{i=1}^n (1 - y_i)(1 - F_i)^{-1} \frac{\partial F_i}{\partial \beta} + \sum_{i=1}^n y_i F_i^{-1} \frac{\partial F_i}{\partial \beta} \\
&= - \sum (1 - F_i)^{-1} \frac{\partial F_i}{\partial \beta} + \sum y_i \left[ (1 - F_i)^{-1} \frac{\partial F_i}{\partial \beta} + F_i^{-1} \frac{\partial F_i}{\partial \beta} \right] \\
&= - \sum (1 - F_i)^{-1} \frac{\partial F_i}{\partial \beta} + \sum y_i [(1 - F_i)^{-1} + F_i^{-1}] \frac{\partial F_i}{\partial \beta}
\end{aligned}$$

The first term can be written as  $-\sum (1 - F_i)^{-1} F_i^{-1} F_i \frac{\partial F_i}{\partial \beta}$  while the second term

$$= \sum y_i \left\{ \frac{F_i + 1 - F_i}{F_i(1 - F_i)} \right\} \frac{\partial F_i}{\partial \beta} = \sum y_i F_i^{-1} (1 - F_i)^{-1} \frac{\partial F_i}{\partial \beta}$$

$$\begin{aligned}
\therefore L_\beta &= - \sum (1 - F_i)^{-1} F_i^{-1} F_i \frac{\partial F_i}{\partial \beta} + \sum y_i (1 - F_i)^{-1} F_i^{-1} \frac{\partial F_i}{\partial \beta} \\
&= \sum (y_i - F_i) (1 - F_i)^{-1} F_i^{-1} \frac{\partial F_i}{\partial \beta} \\
&= \sum \left[ \frac{y_i - F_i}{(1 - F_i)^{1/2} F_i^{1/2}} \right] \left[ \frac{\partial F_i / \partial \beta}{(1 - F_i)^{1/2} F_i^{1/2}} \right] \\
&= \sum w_i v_i = w'v.
\end{aligned}$$

where

$$w_i = \frac{y_i - F_i}{(1 - F_i)^{1/2} F_i^{1/2}}, \quad v_i = \frac{\partial F_i / \partial \beta}{(1 - F_i)^{1/2} F_i^{1/2}}$$

The MLE works with the moment conditions  $E(w_i v_i) = 0$  which is identical to those employed by the weighted least squares estimator. Hence we conclude that the final estimate obtained from the non-linear regression is *identical* to the MLE. Finally to evaluate  $\frac{\partial F}{\partial \beta} = \frac{\partial \int_{-\infty}^{x_i^\beta} f(u) du}{\partial \beta}$  we need to use Leibniz' rule

$$\frac{\partial}{\partial x} \int_{-\infty}^{g(x)} \psi(x, y) dy = \int_{-\infty}^{g(x)} \frac{\partial \psi}{\partial x} + \frac{\partial g}{\partial x} \psi(x, g(x)).$$

Applying this we get

$$\frac{\partial F_i}{\partial \beta} = x'_i f(x'_i \beta)$$



and so the score is

$$\begin{aligned} L_\beta &= \sum_{i=1}^N x'_i f(x'_i \beta) F_i^{-1} (1 - F_i)^{-1} (y_i - F_i) \\ &= \sum_{i=1}^N L_{\beta i}. \end{aligned}$$

We can then use this to get the MLE's of  $\beta$  and also to perform specification tests.

### 1.3 Issues of Testing

There is one further difference between the standard regression set up and the analysis of binary data — the size of samples. It is rare in time series analysis to have more than 100 observations, but it is common to see 5000 or so with binary data. We therefore have to ask if some of our familiar attitudes have to be changed by this fact, and the answer is at least yes to one of them. Consider the  $t$ -statistic on  $x_{2t}$  in the regression of  $y_i$  against  $x_{1i}$  and  $x_{2i}$ . If  $x_{2i}$  should be in this regression, as the sample size ( $n$ ) grows so will the  $t$ -value. i.e. as  $n \rightarrow \infty, t \rightarrow \infty$ . Thus, if we compare it to (say) a critical value of 2, eventually we must always reject the smaller model in favor of the (correct) larger one. This is the property of a *consistent test statistic* i.e. it results in a rejection of a false null with probability one as  $n \rightarrow \infty$ . But let's look at what happens if  $x_{2i}$  should not be in the model i.e. its coefficient is zero. Then in large samples the estimated coefficient of  $x_{2i}$  would be normally distributed around zero. But now notice what happens. There is always some probability that we can get a  $t$ -value  $> 2$  purely by chance. Precisely, as  $n \rightarrow \infty$  there is a 5% chance that  $|t| > 1.96$ . Thus there is a 5% chance that we would select the larger model when it is actually incorrect. The combination of these two tendencies produces a bias towards selecting a larger model. Clearly this is a bit odd. We seem more concerned with avoiding a type II error (accepting a false model) than with Type I (rejecting a true model) as  $n \rightarrow \infty$ . Just why one should have such an asymmetry is not at all clear, and it seems more reasonable that one would want to keep a constant balance between the two types of errors as the sample size grows. How do we do that? It is clear that the problem arises from keeping the critical value constant as  $n \rightarrow \infty$ ; instead we should be making it *larger*. Exactly how it should vary with  $n$  is of course impossible to resolve unless one could specify a loss function that shows the trade-off between Type I and Type II errors, and various suggestions have been made. One of these — known as the Schwarz criterion — produces the following critical values as  $n \rightarrow \infty$ . Clearly for  $n \simeq 100$  the value of 2 is quite reasonable, but as the sample size grows one should be increasing the critical value. Hence with large sample sizes it does not make a lot of sense to conclude that a variable has a “significant” impact unless the  $t$ -value is around 4. So one needs to take great care in assessing the results of Probit/Tobit/Logit (and regressions for that matter) fitted to large numbers of data points.

sig level \ n	<u>5</u>	<u>10</u>	<u>50</u>	<u>100</u>	<u>1000</u>	<u>10000</u>	<u>100000</u>
.05	2.57	2.23	2.01	1.98	1.96	1.96	1.96
.01	4.03	3.17	2.67	2.60	2.58	2.58	2.58
Schwartz	1.32	1.56	2.00	2.16	2.63	3.04	3.39

## 1.4 Multivariate Discrete Choice Models

Now we have to allow for the fact that many of the situations with discrete responses are more complex than the binary case looked at above. Sometimes one sees such extensions categorized in the following way. First, we could have instances where there are *multiple choices* to be made e.g. one might have the two possibilities of “travel to work in rush hour,” and “travel to work out of rush hour” as well as the choice of bus or car. Second, we might have to make a single choice out of more than two alternatives. For example, one might have data upon electoral choices and be interested in explaining the vote for a particular party. Whilst convenient for discussion, the distinction should not be exaggerated, since we could always enumerate the travel-time, travel-mode choice combinations and then treat the problem as making a single decision amongst the four alternatives, although, to some extent, one has lost some of the structure to the choices.

Let’s consider the single choice/ multiple alternatives case. One could have collected data upon votes for three candidates  $A$ ,  $B$ ,  $C$  and analyzed them by giving a 1 for an “ $A$ ” Vote, zero for the rest, fitting a binary model to explain the “ $A$ ” vote, and then do the same for “ $B$ ” and “ $C$ ”. But this is troublesome as the probabilities of each action should sum to one, but there is no restriction being imposed that would ensure this. The situation is the same as that for portfolio models in time series, where the sum of asset demands must sum to total wealth. These restrictions should not be ignored if sensible results are to be had.

Before we look at this situation we should first provide another derivation of the Probit model which will turn out to be very useful. This emphasises that the discrete choice model can be thought of as having an underlying latent variable form where  $y_i$  is observed,  $y_i^*$  is not and  $x_i$  is assumed to be weakly exogenous

$$y_i^* = x_i' \beta + e_i^*,$$

where we observe  $y_i = 1$  if  $y_i^* > 0$  and zero otherwise i.e.  $y_i = 1(y_i^* > 0)$  where  $1(A)$  is the indicator function taking the value unity if the event  $A$  is true and zero otherwise.

If we assume that  $e_i^*$  is *i.i.d.*  $N(0,1)$  then the model we are looking at is the *Probit* model. The data is  $y_i, x_i$  so that we need  $\Pr(y_i = 1|x_i)$  and we get this

from

$$\begin{aligned}
\Pr(y_i = 1|x_i) &= \Pr(y_i^* > 0|x_i) \\
&= \Pr(x_i'\beta + e_i^* > 0|x_i) \\
&= \Pr(e_i^* > -x_i'\beta|x_i) \\
&= \Pr(e_i^* < x_i'\beta|x_i) \\
&= \int_{-\infty}^{x_i'\beta} \phi(u)du = \Phi(x_i'\beta)
\end{aligned}$$

using the properties of the normal density and using the standard symbols that  $\phi$  is the standard normal density and  $\Phi$  is the cumulative normal.

Now there is nothing conceptually difficult about moving from a binary to a multi-response framework, but the numerical difficulties can be horrendous. Thus most of this literature seeks ways to structure the problem so as to cut down the computational load as well as just choosing the simplest possible statistical framework. Some of the problems become evident when we try to extend the Probit model. When faced with the travel time/mode example above it is natural to formulate this as a system of equations with two latent variables  $y_{1i}^*$  and  $y_{2i}^*$  and their associated observed variables  $y_{1i}$  and  $y_{2i}$  which take the values zero and unity depending on which choice is made in each category. Then we have the two equation system

$$y_{1i}^* = x_{1i}'\beta_1 + u_{1i}^* \quad (1)$$

$$y_{2i}^* = x_{2i}'\beta_2 + u_{2i}^* \quad (2)$$

and  $y_{1i} = 1(y_{1i}^* > 0)$ ,  $y_{2i} = 1(y_{2i}^* > 0)$ . Following the same logic as for the single equation case

$$\begin{aligned}
\therefore \Pr\{y_{2i} = 1, y_{1i} = 1\} &= \Pr\{u_{1i}^* < x_{1i}'\beta_1, u_{2i}^* < x_{2i}'\beta_2\} \\
&= \int_{-\infty}^{x_{1i}'\beta_1} \int_{-\infty}^{x_{2i}'\beta_2} f(u_1^*, u_2^*) du_2^* du_1^*
\end{aligned}$$

Thus a bivariate numerical integration is required and this will need to be done every time the likelihood is evaluated, making the extension of the binary Probit model to a multivariate one computationally very expensive. A good deal of time has been spent in recent years working out computer intensive methods of overcoming this problem.

Thinking of the problem as making a single choice from a set of alternatives can sometimes be more useful. In particular, the alternative estimator for binary data, the Logit model, has a nice generalization to the multi response set-up. With  $K$  choices it defines the probabilities of choosing the  $j$ 'th alternative as being given by

$$Prob(z_{ji} = 1) = P_{ji} = \frac{e^{x'_{ji}\beta_j}}{\sum_{\ell=1}^K e^{x'_{i\ell}\beta_\ell}}.$$

Thus in the case of two choices we have

$$\begin{aligned} P_{1i} &= \frac{e^{x'_{1i}\beta}}{e^{x'_{1i}\beta} + e^{x'_{2i}\beta}} \\ P_{2i} &= \frac{e^{x'_{2i}\beta}}{e^{x'_{1i}\beta} + e^{x'_{2i}\beta}}. \end{aligned}$$

Now immediately we can see a potential problem. Suppose we changed the value of  $\beta$  to  $\beta^* = \beta + \delta$ . Then

$$P_{1i} = \frac{e^{x'_{1i}(\beta+\delta)}}{e^{x'_{1i}(\beta+\delta)} + e^{x'_{2i}(\beta+\delta)}} = \left[ \frac{e^{x'_{1i}\delta} e^{x'_{1i}\beta}}{e^{x'_{1i}\delta} e^{x'_{1i}\beta} + e^{x'_{2i}\delta} e^{x'_{2i}\beta}} \right]$$

and, if  $x_{i1} = x_{i2} = x_i$ , we see that  $P_{ji}$  does not depend upon  $\delta$  i.e. we could not distinguish between the values  $\beta$  and  $\beta^*$  since the two are observationally equivalent as they generate the same probabilities for the choices. Some restriction needs to be placed upon the problem in order to estimate the  $\beta$  i.e. to ensure that it is identified. One solution is to maintain that the  $x_{jt}$  are not identical but many problems exist in which one wants to make them identical. Another is to allow the parameter values to vary with  $j$  i.e.

$$Prob(z_{ji} = 1) = P_{ji} = \frac{e^{x'_{ji}\beta_j}}{\sum_{\ell=1}^K e^{x'_{i\ell}\beta_\ell}}$$

An extreme form of this assumption that guarantees identification is that  $\beta_1 = 0$  so that

$$\begin{aligned} Prob(z_{1i} = 1) = P_{1i} &= \frac{1}{1 + \sum_{\ell=2}^K e^{x'_{i\ell}\beta_\ell}} \\ Prob(z_{ji} = 1) = P_{ji} &= \frac{e^{x'_{i1}\beta_j}}{1 + \sum_{\ell=2}^K e^{x'_{i\ell}\beta_\ell}}, \quad j = 2, \dots, K, \end{aligned}$$

and this produces the *multinomial Logit model*. It represents a very simple extension of the binary one, and for this reason it has probably become the preferred option in dealing with multi-response data. Notice that the normalization is like that of a dummy variable i.e. things are being measured relative to the first choice.

Estimation is almost invariably with MLE. Let  $I_j$  be the observations on individuals who choose response  $j$ . Then the likelihood is

$$\prod_{i \in I_1} P_{1i} \prod_{i \in I_2} P_{2i} \dots \prod_{i \in I_K} P_{Ki}$$

giving the log likelihood

$$\sum_{i \in I_1} \log P_{1i} + \sum_{i \in I_2} \log P_{2i} + \dots + \sum_{i \in I_K} \log P_{Ki}$$

or

$$\begin{aligned} & \sum_{i \in I_1} \log \left( \frac{1}{1 + e^{x'_{2i}\beta_2} + \dots + e^{x'_{Ki}\beta_K}} \right) + \sum_{i \in I_2} \log \left( \frac{e^{x'_{2i}\beta_2}}{1 + e^{x'_{2i}\beta_2} + \dots + e^{x'_{Ki}\beta_K}} \right) \\ & + \dots + \sum_{i \in I_K} \log \left( \frac{e^{x'_{Ki}\beta_K}}{1 + e^{x'_{2i}\beta_2} + \dots + e^{x'_{Ki}\beta_K}} \right) \\ = & \sum_{i \in I_2} x'_{2i}\beta_2 + \dots + \sum_{i \in I_K} x'_{Ki}\beta_K - \dots - \sum_{i=1}^N \log(1 + e^{x'_{2i}\beta_2} + \dots + e^{x'_{Ki}\beta_K}) \end{aligned}$$

This is a fairly easy function to maximize and one gets standard MLE properties out of it. Unfortunately the multi-response logit model has one weakness, known as the independence of irrelevant alternatives assumption (IIA). What this assumption implies is that the choice between any two alternatives does not depend upon a third one, i.e., the relative probabilities of choosing 1 vs. 2 would be

$$\frac{P_{i1}}{P_{i2}} = \frac{e^{x'_{i1}\beta_1}}{e^{x'_{i2}\beta_2}}$$

and does not depend upon  $x_{i3}$  at all. This is both a strength and weakness as it enables one to ignore third alternatives in estimation. But the fact that it may clash with the data requires careful thought and a good deal of energy has gone into constructing tests of the IIA in recent years. By far the most successful of these has been that in Hausman/McFadden(1983, *Econometrica*) which compares the estimates of  $\beta$  when the third alternative is dropped from the data and when it is in. This is a form of encompassing test. If IIA does not hold then the two should be quite different since there would be inconsistent estimators of  $\beta$  in both cases if IIA is not true, but they will converge to different values. Computing the point estimates provides an informal check but Hausman/McFadden formalize this into a formal specification test.

**Notes:** (i) We have ignored many other types of models used in discrete choice analysis e.g. if the responses can be *ordered* or *nested* in some way. Greene has more on these issues.

(ii) Perhaps the most interesting work in the estimation of discrete choice model in the past decade has been the use of computer intensive methods to

do MLE on the multivariate Probit model. One advantage of that model is that it can avoid the IIA assumption by making the errors  $u_{ji}^*$  correlated. The computer is basically used to evaluate the multivariate integrals needed to form the likelihood.