2.5 The probit and logit models

An alternative approach, called by Goldberger (1964) the *probit analysis model,* is to assume that there is an underlying response variable y_i^* defined by the regression relationship

$$y_i^* = \beta' x_i + u_i \tag{2.14}$$

In practice, y_i^* is unobservable. What we observe is a dummy variable y defined by

$$\begin{aligned} y &= 1 \quad \text{if } y_i^* > 0 \\ y &= 0 \quad \text{otherwise} \end{aligned} \tag{2.15}$$

In this formulation, $\beta' x_i$ is not $E(y_i \mid x_i)$ as in the linear probability model; it is $E(y_i^* \mid x_i)$.

From the relations (2.14) and (2.15) we get

$$\begin{aligned} \text{Prob}(y_i = 1) &= \text{Prob}(u_i > -\beta' x_i) \\ &= 1 - F(-\beta' x_i) \end{aligned} \tag{2.16}$$

where F is the cumulative distribution function for u.

In this case the observed values of y are just realizations of a binomial process with probabilities given by (2.16) and varying from trial to trial (depending on x_i). Hence, the likelihood function is

$$L = \prod_{y_i=0} F(-\beta' x_i) \prod_{y_i=1} [1 - F(-\beta' x_i)] \tag{2.17}$$

The functional form for F in (2.17) will depend on the assumptions made about u_i in (2.14). If the cumulative distribution of u_i is the logistic, we have the *logit model.* In this case,

$$F(-\beta' x_i) = \frac{\exp(-\beta' x_i)}{1 + \exp(-\beta' x_i)} = \frac{1}{1 + \exp(\beta' x_i)}$$

Hence,

$$1 - F(-\beta' x_i) = \frac{\exp(\beta' x_i)}{1 + \exp(\beta' x_i)} \tag{2.18}$$

In this case we say that there is a closed-form expression for F, because it does not involve integrals explicitly. Not all distributions permit such a closed-form expression. For instance, in the *probit model* (or, more accurately, the *normit* model) we assume that u_i are $IN(0, \sigma^2)$. In this case,

$$F(-\beta' x_i) = \int_{-\infty}^{-\beta' x_i/\sigma} \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{t^2}{2}\right) dt \tag{2.19}$$

It can be easily seen from (2.19) and the likelihood function (2.17) that we can estimate only β/σ, and not β and σ separately. Hence, we might as well assume $\sigma = 1$ to start with.

Because the cumulative normal distribution and the logistic distribution are very close to each other, except at the tails, we are not likely to get very different results using (2.18) or (2.19), that is, the logit or the probit method, unless the samples are large (so that we have enough observations at the tails). However, the estimates of β from the two methods are not directly comparable. Because the logistic distribution has a variation $\pi^2/3$, the estimates of β obtained from the logit model have to be multiplied by $3^{1/2}/\pi$ to be comparable to the estimates obtained from the probit model (where we normalize σ to be equal to 1).

Amemiya (1981) suggested that the logit estimates be multiplied by $1/1.6 = 0.625$, instead of $(3^{1/2}/\pi)$ saying that this transformation produces a closer approximation between the logistic distribution and the distribution function of the standard normal. He also suggested that the coefficients of the linear probability model $\hat{\beta}_{\text{LP}}$ and the coefficients of the logit model $\hat{\beta}_{\text{L}}$ are related by the relationships

$$\begin{aligned} \hat{\beta}_{\text{LP}} &\simeq 0.25\hat{\beta}_{\text{L}} \quad \text{except for the constant term} \\ \hat{\beta}_{\text{LP}} &\simeq 0.25\hat{\beta}_{\text{L}} + 0.5 \quad \text{for the constant term} \end{aligned}$$

Thus, if we need to make $\hat{\beta}_{\text{LP}}$ comparable to the probit coefficients, we need to multiply them by 2.5 and subtract 1.25 from the constant term.

An alternative way of comparing the models would be to (a) calculate the sum of squared deviations from predicted probabilities, (b) compare the percentages correctly predicted, and (c) look at the derivatives of the probabilities with respect to a particular independent variable. Let x_{ik} be the kth element of the vector of explanatory variables x_i, and let β_k be the kth element of β. Then the derivatives for the probabilities given by the linear probability model, probit model, and logit model are, respectively,

$$\frac{\partial}{\partial x_{ik}}(x_i'\beta) = \beta_k$$

$$\frac{\partial}{\partial x_{ik}}\Phi(x_i'\beta) = \phi(x_i'\beta)\beta_k$$

$$\frac{\partial}{\partial x_{ik}}L(x_i'\beta) = \frac{\exp(x_i'\beta)}{[1 + \exp(x_i'\beta)]^2}\beta_k$$

These derivatives will be needed for predicting the effects of changes in one of the independent variables on the probability of belonging to a

Table 2.1. *Comparison of the probit, logit, and linear probability models: loan data from South Carolina*

Variable[a]	Linear probability model	Logit model	Probit model
AI	1.489 (4.69)[b]	2.254 (4.60)	2.030 (4.73)
XMD	−1.509 (5.74)	−1.170 (5.57)	−1.773 (5.67)
DF	0.140 (0.78)	0.563 (0.87)	0.206 (0.95)
DR	−0.266 (1.84)	−0.240 (1.60)	−0.279 (1.66)
DS	−0.238 (1.75)	−0.222 (1.51)	−0.274 (1.70)
DA	−1.426 (3.52)	−1.463 (3.34)	−1.570 (3.29)
NNWP	−1.762 (0.74)	−2.028 (0.80)	−2.360 (0.85)
NMFI	0.150 (0.23)	0.149 (0.20)	0.194 (0.25)
NA	−0.393 (1.34)	−0.386 (1.25)	−0.425 (1.26)
Const.	0.501	0.363	0.488

Note: Total number of observations = 750; number of applications rejected = 250; number of applications accepted = 500. To make the coefficients comparable to one another, we have multiplied the logit coefficients by 0.625 and the coefficients of the linear probability model by 2.5 and then subtracted 1.25 from the constant term, as explained in the text.
[a] AI, applicant's + coapplicant's incomes (in 10^5 dollars); XMD, debt minus mortgate payment (in 10^3 dollars); DF, 1 if female; DR, 1 if nonwhite; DS, 1 if single; DA, age of house (in 10^2 years); NNWP, (neighborhood % nonwhite) $\times 10^3$; NMFI, neighborhood mean family income (in 10^5 dollars); NA, neighborhood average age of homes (in 10^2 years).
[b] t ratios in parentheses.

group. In the case of the linear probability model, these derivatives are constant. In the case of the probit and logit models, we need to calculate them at different levels of the explanatory variables to get an idea of the range of variation of the resulting changes in the probabilities.

As an illustration, we consider data on a sample of 750 mortgage loan applications in the Columbia, South Carolina, metropolitan area. There were 500 loan applications accepted and 250 loan applications rejected. Define

$y = 1$ if the loan application was accepted
$y = 0$ if the loan application was rejected

Table 2.1 shows the results for the linear probability model, and logit model, and the probit model. The linear probability model was estimated by ordinary least squares (not weighted least squares). Thus the correction for heteroscedasticity has not been made. The coefficients of the discriminant function can thus be obtained from these estimates by using the formula (2.13).

The likelihood function (2.17) can be written as

$$L = \prod_{i=1}^{n} \left(\frac{1}{1 + \exp(\beta' x_i)} \right)^{1-y_i} \left(\frac{\exp(\beta' x_i)}{1 + \exp(\beta' x_i)} \right)^{y_i} \tag{2.20}$$

$$= \frac{\exp(\beta') \sum_{i=1}^{n} x_i y_i}{\prod_{i=1}^{n} [1 + \exp(\beta' x_i)]} \tag{2.21}$$

Define $t^* = \sum_{i=1}^{n} x_i y_i$. To find the maximum-likelihood (ML) estimate of β, we have

$$\log L = \beta' t^* - \sum_{i=1}^{n} \log[1 + \exp(\beta' x_i)]$$

Hence, $\partial \log L / \partial \beta = 0$ gives

$$S(\beta) = -\sum_{i=1}^{n} \frac{\exp(\beta' x_i)}{1 + \exp(\beta' x_i)} x_i + t^* = 0 \tag{2.22}$$

These equations are nonlinear in β. Hence, we have to use the Newton–Raphson method or the scoring method to solve the equations. The information matrix is

$$I(\beta) = E\left(-\frac{\partial^2 \log L}{\partial \beta \partial \beta'} \right)$$

$$= \sum_{i=1}^{n} \frac{\exp(\beta' x_i)}{[1 + \exp(\beta' x_i)]^2} x_i x_i' \tag{2.23}$$

Starting with some initial value of β, say β_0, we compute the values $S(\beta_0)$ and $I(\beta_0)$. Then the new estimate of β is, by the method of scoring,

$$\beta_1 = \beta_0 + [I(\beta_0)]^{-1} S(\beta_0)$$

In practice, we divide both $I(\beta_0)$ and $S(\beta_0)$ by n, the sample size. This iterative procedure is repeated until convergence. In the present case it is clear that $I(\beta)$ is positive definite at each stage of iteration. Hence, the iterative procedure will converge to a maximum of the likelihood function, no matter what the starting value is. If the final converged estimates are denoted by $\hat{\beta}$, then the asymptotic covariance matrix is estimated by $[I(\hat{\beta})]^{-1}$. These estimated variances and covariances will enable us to test hypotheses about the different elements of $\hat{\beta}$.

After estimating β, we can get estimated values of the probability that the ith observation is equal to 1. Denoting these estimated values by $\hat{p}_i$, we have

$$\hat{p}_i = \frac{\exp(\hat{\beta}' x_i)}{1 + \exp(\hat{\beta}' x_i)}$$

Equation (2.22) shows that

$$\sum \hat{p}_i x_i = \sum y_i x_i \tag{2.24}$$

Thus, if x_i includes a constant term, then the sum of the estimated probabilities is equal to $\sum y_i$ or the number of observations in the sample for which $y_i=1$. In other words, the predicted frequency is equal to the actual frequency. Similarly, if x_i includes a dummy variable, say 1 for female, 0 for male, then the predicted frequency will be equal to the actual frequency for each sex group. Similar conclusions follow for the linear probability model by virtue of the fact that equations (2.24) are the least-squares normal equations in that case.

In any case, after estimating $\hat{\beta}$ and then $\hat{p}_i$ by the logit model, it is always good practice to check whether or not equations (2.24) are satisfied.

For the probit model, we substitute expression (2.19) in equation (2.17).

Let us denote by $\phi(\cdot)$ and $\Phi(\cdot)$ the density function and the distribution function, respectively, of the standard normal. Then for the probit model the likelihood function corresponding to (2.20) is

$$L = \prod_{i=1}^{n} [\Phi(\beta' x_i)]^{y_i} [1 - \Phi(\beta' x_i)]^{1-y_i}$$

and the log-likelihood is

$$\log L = \sum_{i=1}^{n} y_i \log \Phi(\beta' x_i) + \sum_{i=1}^{n} (1 - y_i) \log[1 - \Phi(\beta' x_i)]$$

Differentiating $\log L$ with respect to β yields

$$S(\beta) = \frac{\partial \log L}{\partial \beta} = \sum_{i=1}^{n} \frac{[y_i - \Phi(\beta' x_i)]}{\Phi(\beta' x_i)[1 - \Phi(\beta' x_i)]} \phi(\beta' x_i) x_i$$

The ML estimator $\hat{\beta}_{\mathrm{ML}}$ can be obtained as a solution of the equations $S(\beta)=0$.

These equations are nonlinear in β, and thus we have to solve them by an iterative procedure. The information matrix is

$$I(\beta) = E\left(-\frac{\partial^2 \log L}{\partial \beta \partial \beta'}\right)$$

$$= \sum_{i=1}^{n} \frac{[\phi(\beta' x_i)]^2}{\Phi(\beta' x_i)[1 - \Phi(\beta' x_i)]} x_i x_i'$$

As with the logit model, we start with an initial value of β, say β_0, and compute the values $S(\beta_0)$ and $I(\beta_0)$. Then the new estimate of β is, by the method of scoring,

$$\beta_1 = \beta_0 + [I(\beta_0)]^{-1} S(\beta_0)$$

Note that $I(\beta)$ is positive definite at each stage of the iteration. Hence, the iterative procedure will converge to a maximum of the likelihood function no matter what the starting value is. If the final converged estimates are denoted by $\hat{\beta}$, then the asymptotic covariance matrix is estimated by $[I(\hat{\beta})]^{-1}$. These can be used to conduct any tests of significance.

2.6 Comparison of the logit model and normal discriminant analysis

There have been many studies on the relative performances of the logit model and discriminant analysis in analyzing models with dichotomous dependent variables. The reason for this interest is that ordinary least-squares procedures can be used to estimate the coefficients of the linear discriminant function, whereas maximum-likelihood methods are required for estimation of the logit model. Given the high-speed computers now available, computational simplicity is no longer an adequate criterion.

If the independent variables are normally distributed, the discriminant-analysis estimator is the true maximum-likelihood estimator and therefore is asymptotically more efficient than the logit maximum-likelihood estimator (MLE). However, if the independent variables are not normal, the discriminant-analysis estimator is not even consistent, whereas the logit MLE is consistent and therefore more robust (see section 4.4 for some examples of this). Press and Wilson (1978) calculated the probability of correct classification for the two estimators in two empirical examples in which the independent variables were dummy variables, and thus the assumption of normality was violated. In both examples, the logit MLE did slightly better than the discriminant-analysis estimator. The criterion of the goodness of prediction in their study was the probability of correct classification defined by

$$P(\beta' x_i \geqslant 0 \mid y_i = 1) Q + P(\beta' x_i < 0 \mid y_i = 0)(1 - Q)$$

where $Q=\mathrm{Prob}(y_i=1)$.

The close relationship between the logit model and the discriminant function discussed here does not hold good except under very special circumstances in the case of McFadden's (1974) conditional logit model. This model is discussed in section 2.12 and Chapter 3. The relationship between this logit model and discriminant analysis has been discussed by McFadden (1976*a*).