3 Probit

3.1 Functional Form of Choice Probabilities

The restrictions of the logit model, particularly the IIA property, are due to the assumption that the unobserved components of utility are independently and identically distributed. Let the utility that person n obtains from alternative i, labeled U_{in}, be decomposed into an observed part V_{in} and an unobserved part e_{in} for all i in the choice set J_n. Then, for the logit model, any e_{in} and e_{jn}, $i \neq j$, are assumed to have the same distribution, with the same mean and variance, and also to be uncorrelated. These random variables being uncorrelated means that any factor that the researcher does not observe that affects the utility of alternative i does **not** affect the utility of alternative j. The two terms e_{in} and e_{jn} having the same variance means that the unobserved factors that affect the utility of alternative i have the same variation as the different (due to zero correlation) unobserved factors that affect the utility of alternative j. In the real world, these assumptions will seldom actually hold.

The probit model is derived by relaxing these assumptions about the unobserved components of utility. In particular, these unobserved components are assumed, instead of independent, identical extreme values, to be distributed jointly normal, with a general variance-covariance matrix. The critical change here is not from the extreme value distribution to the normal, since these two distributions for a single random variable are practically the same. The important distinction is that, with the **joint** normal distribution, each e_{in}, for all i in J_n, can have a different variance and can be correlated with other e_{jn}, j in J_n, $j \neq i$.

The probit choice probabilities are derived from the assumption of jointly normal unobserved utility components. As usual, utility is decomposed into observed and unobserved parts:

$$U_{in} = V_{in} + e_{in}, \qquad \text{for all} \quad i \text{ in } J_n.$$

Consider the vector composed of each e_{in} for all i in J_n; label this vector $\tilde{e}_n$. We assume that $\tilde{e}_n$ is distributed normal with a mean vector of zero and variance-covariance matrix denoted Ω_n whose elements are parameters that are either specified a priori or estimated by the researcher. That is, the density function of $\tilde{e}_n$ is

$$\phi(\tilde{e}_n) = (2\pi)^{-\frac{1}{2}m_n} |\Omega_n|^{-\frac{1}{2}} \exp[-\tfrac{1}{2}\tilde{e}_n \Omega_n^{-1} \tilde{e}_n],$$

where $|\Omega_n|$ is the determinate of Ω_n and m_n is the number of alternatives in J_n.

Recall that the probability of choosing alternative i is the probability that the utility associated with alternative i is higher than that of any other alternative:

$$P_{in} = \text{Prob}(V_{in} + e_{in} > V_{jn} + e_{jn}, \text{ for all } j \text{ in } J_n, j \neq i).$$

Rearranging,

$$P_{in} = \text{Prob}(e_{jn} < e_{in} + V_{in} - V_{jn}, \text{ for all } j \text{ in } J_n, j \neq i).$$

To evaluate this expression suppose first that e_{in} is given. Then the right-hand side of this expression is the probability that the random variable e_{jn} is below the known value $e_{in} + V_{in} - V_{jn}$, for all j in $J_n, j \neq i$. That is, for given e_{in}, the expression is simply the cumulative distribution of e_{jn} evaluated at $e_{in} + V_{in} - V_{jn}$, for all j in J_n, j in i. Since a cumulative distribution is the integral of the density function, the probability of choosing alternative i given a particular value of e_{in}, labeled $P_{in}(e_{in})$, is simply the density of the random vector $\tilde{e}_n$ integrated from negative infinity to $e_{in} + V_{in} - V_{jn}$ for each element j in $J_n, j \neq i$:

$$P_{in}(e_{in}) = \int_{e_{1n}=-\infty}^{e_{in}+V_{in}-V_{1n}} \int_{e_{2n}=-\infty}^{e_{in}+V_{in}-V_{2n}} \cdots \int_{e_{m_n n}=-\infty}^{e_{in}+V_{in}-V_{m_n n}} \phi(\tilde{e}_n)\, de_{m_n n} \cdots de_{2n}\, de_{1n}, \tag{3.1}$$

where the "$\cdots$" is over all elements e_{jn} in the vector $\tilde{e}_n$ except e_{in}, which is set equal to its given value.

In actuality, the value of e_{in} is not given. Consequently, the probability of choosing alternative i is the probability of choosing it for any given value of e_{in} integrated over all possible values of e_{in}. That is,

$$P_{in} = \int_{e_{in}=-\infty}^{\infty} P_{in}(e_{in})\phi(\tilde{e}_n)\, de_{in}. \tag{3.2}$$

Substituting (3.1) into (3.2) gives

$$P_{in} = \int_{e_{in}=-\infty}^{\infty} \int_{e_{1n}=-\infty}^{e_{in}+V_{in}-V_{1n}} \int_{e_{2n}=-\infty}^{e_{in}+V_{in}-V_{2n}} \cdots \int_{e_{m_n n}=-\infty}^{e_{in}+V_{in}-V_{m_n n}} \phi(\tilde{e}_n)\, de_{m_n n} \cdots de_{2n}\, de_{1n}\, de_{in}, \tag{3.3}$$

where both the parameters entering V_{in} and those entering the variance-

covariance matrix Ω_n are determined in estimation or specified a priori by the researcher.

The probit choice probabilities being in such complex form is the main disadvantage of the model. In particular, estimation of probit models is very expensive because of the complexity of the choice probabilities. To evaluate a log likelihood function (defined in section 2.6) using these choice probabilities, numerous integrations are required for each sampled decision-maker; and to find the value of the parameters that maximizes the function, these numerous integrals must be evaluated numerous times. Several alternative methods of estimating probit models have been proposed, based on Monte Carlo methods and approximations (see section 3.4 for a discussion of these). However, it still remains that estimating a probit model with more than a few alternatives and a few explanatory variables is prohibitively expensive.

There are situations, however, in which the probit model, if the expense of estimation can be borne, is very useful. Two of these are discussed in the following sections.

3.2 Taste Variation

Suppose utility can be decomposed into a linear-in-parameters part that depends only on observed data, plus an unobserved part. Assume further that the parameters are not fixed, but rather vary randomly over decision-makers. This is represented as follows:

$$U_{in} = \beta_n w_{in} + e_{in};$$

$$\beta_n = \bar{\beta} + \tilde{\beta}_n;$$

where

w_{in} is a vector-valued function of observed data,
β_n is a vector of coefficients of w_{in} for person n, unknown to the researcher,
$\bar{\beta}$ is the mean of β_n over all persons, and
$\tilde{\beta}_n$ is the deviation of the coefficient vector of person n from the mean coefficients (i.e., $\tilde{\beta}_n \equiv \beta_n - \bar{\beta}$).

Substituting the equation for β_n,

$$U_{in} = \bar{\beta} w_{in} + \tilde{\beta}_n w_{in} + e_{in}.$$

The last two terms on the right-hand side are both unobserved (since $\tilde{\beta}_n$ is

unobserved); denote their sum as η_{in} to obtain

$$U_{in} = \bar{\beta} w_{in} + \eta_{in}.$$

If both $\tilde{\beta}_n$ and e_{in} are normally distributed, then η_{in} is also normally distributed, and the choice probabilities, stated in terms of $\bar{\beta} w_{in}$, are probit. Estimation of the model provides values for $\bar{\beta}$ and the variance-covariance matrix for η.

For example, consider a two-alternative choice situation in which one explanatory variable enters the representative utility of each alternative. In this case,

$$U_{1n} = \beta_n y_{1n} + e_{1n};$$

$$U_{2n} = \beta_n y_{2n} + e_{2n};$$

where y_{1n} and y_{2n} are the values that the explanatory variable y takes for person n in each of the two alternatives (e.g., y could be the cost of obtaining the alternative). Assume that β_n is normally distributed with mean $\bar{\beta}$ and variance σ_β^2. Assume further that e_{1n} and e_{2n} are independently normally distributed each with zero mean and variance σ_e^2. (The assumption of independence simplifies the example but is not necessary.) With these assumptions, utility can be expressed as

$$U_{1n} = \bar{\beta} y_{1n} + \eta_{1n};$$

$$U_{2n} = \bar{\beta} y_{2n} + \eta_{2n};$$

where η_{1n} and η_{2n} are jointly normally distributed. The η have zero mean:

$$E(\eta_{in}) = E(\tilde{\beta}_n y_{in} + e_{in}) = 0, \qquad i = 1, 2.$$

The variance-covariance matrix for η is determined as follows. The variance of each is

$$\begin{aligned} V(\eta_{in}) &= V(\tilde{\beta}_n y_{in} + e_{in}) \\ &= y_{in}^2 \sigma_\beta^2 + \sigma_e^2, \qquad i = 1, 2, \end{aligned}$$

given that $\tilde{\beta}_n$ and e_{in} are uncorrelated. Their covariance is

$$\begin{aligned} E(\eta_{1n} \cdot \eta_{2n}) &= E((\tilde{\beta}_n y_{1n} + e_{1n})(\tilde{\beta}_n y_{2n} + e_{2n})) \\ &= E(\tilde{\beta}_n^2 y_{1n} y_{2n} + e_{1n} e_{2n} + e_{1n} \tilde{\beta}_n y_{1n} + e_{2n} \tilde{\beta}_n y_{2n}) \\ &= y_{1n} y_{2n} \sigma_\beta^2, \end{aligned}$$

since e_{1n} and e_{2n} are uncorrelated and $\tilde{\beta}_n$ is uncorrelated with either e. Therefore, in this example

$$\Omega_n = \begin{bmatrix} y_{1n}^2\sigma_\beta^2 + \sigma_e^2 & y_{1n}y_{2n}\sigma_\beta^2 \\ y_{1n}y_{2n}\sigma_\beta^2 & y_{2n}^2\sigma_\beta^2 + \sigma_e^2 \end{bmatrix}$$

$$= \sigma_\beta^2 \begin{bmatrix} y_{1n}^2 & y_{1n}y_{2n} \\ y_{1n}y_{2n} & y_{2n}^2 \end{bmatrix} + \sigma_e^2 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

One last step is required for estimation. Note that decisionmakers' choices are not affected by a multiplicative transformation of utility: U_{in} is larger than U_{jn} for all $j \neq i$ if and only if U_{in}/λ is larger than U_{jn}/λ for all $j \neq i$. Consequently, the model $U_{in} = \bar{\beta}y_{in} + \eta_{in}$, where $\text{Var}(\eta_{in}) = y_{in}^2\sigma_\beta^2 + \sigma_e^2$, is equivalent to the model $U_{in}^* = (\bar{\beta}/\sigma_e)y_{in} + \eta_{in}^*$, where $\text{Var}(\eta_{in}^*) = y_{in}^2(\sigma_\beta/\sigma_e)^2 + 1$. Since any set of parameters $\bar{\beta}$, σ_β^2, and σ_e^2 that have the same ratios result in the same utility specification, a normalization is applied for estimation. A convenient normalization for this case is $\sigma_e^2 = 1$. Under this normalization,

$$\Omega_n = \sigma_\beta^2 \begin{bmatrix} y_{1n}^2 & y_{1n}y_{2n} \\ y_{1n}y_{2n} & y_{2n}^2 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

The values of y_{1n} and y_{2n} are observed by the researcher. The parameters σ_β^2 and $\bar{\beta}$ are determined through estimation of the model on a sample of decisionmakers. Thus, the researcher learns both the mean ($\bar{\beta}$) and the variance (σ_β^2) of the random coefficients of the observed variables entering utility.

3.3 Nonindependence from Irrelevant Alternatives

Independence or nonindependence from irrelevant alternative only becomes an issue in situations of three or more alternatives (since with only two alternatives there is no other alternative for the ratio of the two probabilities to be independent or nonindependent from). Consider a simple three-alternative case in which a home buyer can choose among purchase-money mortgages offered by three different lending institutions. One of the mortgages has a fixed interest rate, while the other two have variable rates. In this situation, it is unrealistic to expect the choice probabilities to exhibit IIA. Improving the characteristics of one variable rate loan (i.e., decreasing its initial interest rate) would be expected to reduce the probability of the other variable rate loan much more (proportionately)

than the probability of the fixed rate loan, since the (unobserved) concern about risk that is associated with variable rate loans must be overcome in switching from a fixed to a variable rate loan but not (or not to the degree) in switching between two kinds of variable rate loans.

This situation can be modeled by probit with the source of non-IIA explicitly incorporated. Label the fixed rate loan as F and the two variable rate loans as VA and VB. Suppose the utility of homebuyer n associated with each loan depends on the initial interest rate of the loan (I_{in}, which is different for each of the three loans and varies over homebuyers on the basis of their credit worthiness) and the maximum possible increase in the interest rate (M_{in}, which is zero for the fixed rate loan and positive but perhaps different for each of the two variable rate loans). In addition, assume that utility depends on two unobserved factors: the homebuyer's perception of, and concern about, the degree of risk associated with the possibility of increased mortgage payments (labeled R_{in}, which is zero for the fixed rate loan and varies randomly for each of the two variable rate loans), and the homebuyer's perception of the ease of dealing with each institution (labeled η_i and depending on the location, reputation, and so on of each institution). With linear-in-parameters utility and suppressing alternative-specific constants for notational simplicity, we have

$$U_{in} = \alpha I_{in} + \beta M_{in} + e_{in}, \qquad i = \text{F, VA, VB,}$$

where $e_{in} = -R_{in} + \eta_{in}$ and the negative sign before R_{in} reflects the fact that risk is undesirable.

One would expect R_{in} to be correlated over the two variable rate loans: if the homebuyer thinks interest rates will rise and is concerned about the ability to keep up payments with an increased loan rate, then the concern would be applicable for both the variable rate loans.[1] Thus, even if η_{in} is independent across alternatives, the entire unobserved component of utility, e_{in}, is correlated.

Let η_{in} be distributed independently identically normal with zero mean and variance ω^2 and not correlated with R_{in}. Also let R_{in} be normally distributed for each of the two variable rate loans, with zero mean, variance σ^2 for each loan, and a covariance across the variable rate loans of σ^2_{AB}. (R_{in} for the fixed rate loan is nonstochastically zero.) Then the unobserved component of utility is also normally distributed with zero mean and variance-covariance

$$\Omega_n = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \sigma^2 & \sigma^2_{AB} \\ 0 & \sigma^2_{AB} & \sigma^2 \end{bmatrix} + \omega \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Estimation of probit choice probabilities provides values of the coefficients α and β in the observed component of utility as well as the variance and covariance of the perceived risk associated with each variable rate loan.[2] The more general case in which the variance of e_{in} and R_{in} is different for each loan, and in which η_{in} is correlated over alternatives, can also be specified.

3.4 Estimation

The most straightforward (at least theoretically) way to estimate parameters in a probit model is through maximum likelihood techniques. The log likelihood function, defined in section 2.6 for logit models, is

$$\mathrm{LL} = \sum_{n \in N} \sum_{i \in J_n} \delta_{in} \log P_{in}, \tag{3.4}$$

where δ_{in} equals one if decisionmaker n chose alternative i and zero otherwise, and N is the total number of decisionmakers in the sample. Substitution of the formula for probit choice probabilities, i.e., expression (3.3) into (3.4), gives LL as an explicit function of the parameter vector β entering representative utility, and the parameters entering the variance-covariance matrix, Ω_n, of the unobserved component of utility. The values of these two sets of parameters that maximize LL are, under fairly general conditions, consistent and efficient.

As mentioned, however, calculating probit probabilities for any given parameters involves numerous integrations; and these integrations must be performed numerous times in the search for the maximizing parameter values. Consequently, estimation of probit models with more than just a few alternatives and few explanatory variables is extremely expensive with standard maximum likelihood methods.

For this reason, alternative estimation methods have been developed. Two are particularly prominent: (1) a method based on an approximation by C. Clark and (2) a Monte Carlo method that utilizes randomly generated values for the unobserved component of utility. Each of the methods will now be discussed.

Estimation with the Clark Approximation

Clark (1961) demonstrated that the maximum of two normally distributed variables is distributed approximately normal. As will be shown, using this approximation reduces the number of integrals that must be evaluated in the calculation of probit choice probabilities to only one. Since the exact formula for choice probabilities in a situation of K alternatives involes K integrals, this approximation can considerably reduce the cost of estimating probit models.

Consider a choice situation with three alternatives labeled 1, 2, and 3. Denote the vector of utilities associated with these alternatives (U_1, U_2, U_3); assumed to be distributed jointly normal with mean vector (V_1, V_2, V_3) and variance-covariance matrix Ω, where the subscript denoting decisionmaker is suppressed for simplicity. Consider the choice probability for alternative 3. By equation (3.3), the formula for P_3 involves three integrals. However, using Clark's approximation, P_3 can be approximated by a formula with only one integral.

Define $z = \max(U_1, U_2)$. Since U_1 and U_2 are normal, it is possible to derive the mean and variance of z and the covariance of z with U_3. Label these variables as follows:

$E(z) = V_z$;

$\text{Var}(z) = \sigma_z^2$;

$\text{cov}(z, U_3) = \sigma_{z3}^2$.

Though the maximum of two normally distributed variables is not itself normally distributed, Clark showed that treating the maximum as if it is normally distributed does not introduce substantial error. That is, $z \dot{\sim} N(V_z, \sigma_z^2)$ with covariance of σ_{z3}^2 with U_3.

By definition, $P_3 = \text{Prob}(U_3 > z)$. That is, the probability of choosing alternative 3 is the probability that U_3 is greater than the maximum of U_1 and U_2, and hence is greater than both. Rearranging, $P_3 = \text{Prob}(z - U_3 < 0)$. Since U_3 is normally distributed and z is approximately so, $z - U_3$ is also approximately normally distributed, with mean $V_z - V_3$ and variance $\sigma_z^2 + \sigma_3^2 - 2\sigma_{z3}^2$, where σ_3^2 is the variance of U_3. Therefore,

$$P_3 = \text{Prob}(z - U_3 < 0) = \int_{s=-\infty}^{0} \phi\left(\frac{s - (V_z + V_3)}{\sqrt{\sigma_z^2 + \sigma_3^2 - 2\sigma_{z3}^2}}\right) ds,$$

where ϕ is the standard normal density. Approximated in this way, P_3 involves only one integral.

The procedure can be applied recursively when more than three alternatives are involved. Consider a situation with four alternatives labeled 1, 2, 3, and 4. Define $z = \max(U_1, U_2)$ and $y = \max(U_1, U_2, U_3)$. By definition $y = \max(z, U_3)$. Since U_1 and U_2 are normal, z is approximately normal; then, since U_3 is normal and z is approximately normal, y is also approximately normal. The probability of choosing alternative 4 is $P_4 = \text{Prob}(U_4 > y) = \text{Prob}(y - U_4 < 0)$. Since y is approximately normal and U_4 is normal, their difference is approximately normal, and P_4 is simply the density of this approximately normal variable integrated from negative infinity to zero. Instead of performing four integrations as required by expression (3.3), P_4 can be approximated by a formula with only one integral. Situations with more alternatives are handled analogously. For a more detailed discussion, see Daganzo, Bouthelier, and Sheffi (1977).

Monte Carlo Method

The Monte Carlo method approximates the probit choice probabilities by simulating the choices of each decisionmaker under numerous, randomly generated values for unobserved utility.

The process begins by the researcher specifying particular values of the parameters entering V_{in} and the variance-covariance matrix of the vector $\tilde{e}_n$ (consisting of elements e_{in} for all i in J_n). Given the joint distribution of the vector $\tilde{e}_n$ (including values for the parameters entering this matrix), a random number generator produces a realization of this vector. Adding this realization to the observed component of utility (calculated at given values of the parameters entering V_{in}) gives total utility. Comparing total utility across alternatives identifies the alternative that has the highest total utility.

Choice probabilities are then approximated by repeating this process numerous time with the parameters held constant. In each randomly generated realization of unobserved utility, one of the alternative has highest utility. The proportion of times alternative i has the highest utility is an estimate of P_{in}. Obviously, as the number of repetitions increases, this proportion can be expected to become arbitrarily close to P_{in}. Since the distribution of $\tilde{e}_n$ depends on the parameters entering Ω_n and the value of representative utility depends on the parameters entering V_{in}, the Monte Carlo estimate of P_{in} depends on these parameters.

Choice probabilities calculated in this way for the given parameter values

are then entered into the log likelihood function to determine the value of the function at the given values of the parameters. The entire process is then repeated for various different parameter values specified by the researcher. The parameters that result in the highest value of the log likelihood function are taken as the parameter estimates.

While both the Monte Carlo method and that based on the Clark approximation are less expensive than the standard maximum likelihood estimation of probit models, they do not completely solve the problem of probit estimation. The Clark approximation has been found in some situations to be very inaccurate. Especially bothersome is the fact that, in most cases, the researcher does not know the degree of inaccuracy unless standard maximum likelihood estimation is performed for comparison, in which case the approximation method is redundant. The Monte Carlo method does not necessarily entail the accuracy problems of the Clark approximation method, since the true probabilities can be approximated to any degree of accuracy by simply generating a sufficiently large number of realizations of unobserved utility for each sampled decisionmaker. Unfortunately, in most cases, when the number of repetitions is increased sufficiently to assure accuracy, the expense of the Monte Carlo method is not appreciably lower than that of the standard maximum likelihood method. For more details on estimation of probit models see Daganzo (1979) and Lerman and Manski (1981).