5 Probit

5.1 Choice Probabilities

The logit model is limited in three important ways. It cannot represent random taste variation. It exhibits restrictive substitution patterns due to the IIA property. And it cannot be used with panel data when unobserved factors are correlated over time for each decision maker. GEV models relax the second of these restrictions, but not the other two. Probit models deal with all three. They can handle random taste variation, they allow any pattern of substitution, and they are applicable to panel data with temporally correlated errors.

The only limitation of probit models is that they require normal distributions for all unobserved components of utility. In many, perhaps most situations, normal distributions provide an adequate representation of the random components. However, in some situations, normal distributions are inappropriate and can lead to perverse forecasts. A prominent example relates to price coefficients. For a probit model with random taste variation, the coefficient of price is assumed to be normally distributed in the population. Since the normal distribution has density on both sides of zero, the model necessarily implies that some people have a positive price coefficient. The use of a distribution that has density only on one side of zero, such as the lognormal, is more appropriate and yet cannot be accommodated within probit. Other than this restriction, the probit model is quite general.

The probit model is derived under the assumption of jointly normal unobserved utility components. The first derivation, by Thurstone (1927) for a binary probit, used the terminology of psychological stimuli, which Marschak (1960) translated into economic terms as utility. Hausman and Wise (1978) and Daganzo (1979) elucidated the generality of the specification for representing various aspects of choice behavior. Utility is decomposed into observed and unobserved parts: $U_{nj} = V_{nj} + \varepsilon_{nj} \; \forall j$. Consider the vector composed of each ε_{nj}, labeled $\varepsilon_n' = \langle \varepsilon_{n1}, \ldots, \varepsilon_{nJ} \rangle$.

We assume that ε_n is distributed normal with a mean vector of zero and covariance matrix Ω. The density of ε_n is

$$\phi(\varepsilon_n) = \frac{1}{(2\pi)^{J/2}|\Omega|^{1/2}} e^{-\frac{1}{2}\varepsilon_n' \Omega^{-1} \varepsilon_n}.$$

The covariance Ω can depend on variables faced by decision maker n, so that Ω_n is the more appropriate notation; however, we omit the subscript for the sake of simplicity.

The choice probability is

$$\begin{aligned} P_{ni} &= \text{Prob}(V_{ni} + \varepsilon_{ni} > V_{nj} + \varepsilon_{nj} \ \forall j \neq i) \\ &= \int I(V_{ni} + \varepsilon_{ni} > V_{nj} + \varepsilon_{nj} \ \forall j \neq i) \phi(\varepsilon_n)\, d\varepsilon_n, \end{aligned} \tag{5.1}$$

where $I(\cdot)$ is an indicator of whether the statement in parentheses holds, and the integral is over all values of ε_n. This integral does not have a closed form. It must be evaluated numerically through simulation.

The choice probabilities can be expressed in a couple of other ways that are useful for simulating the integral. Let B_{ni} be the set of error terms ε_n that result in the decision maker choosing alternative i: $B_{ni} = \{\varepsilon_n \text{ s.t. } V_{ni} + \varepsilon_{ni} > V_{nj} + \varepsilon_{nj} \ \forall j \neq i\}$. Then

$$P_{ni} = \int_{\varepsilon_n \in B_{ni}} \phi(\varepsilon_n)\, d\varepsilon_n, \tag{5.2}$$

which is an integral over only some of the values of ε_n rather than all possible values, namely, the ε_n's in B_{ni}.

Expressions (5.1) and (5.2) are J-dimensional integrals over the J errors ε_{nj}, $j = 1, \ldots, J$. Since only differences in utility matter, the choice probabilities can be equivalently expressed as $(J-1)$-dimensional integrals over the differences between the errors. Let us difference against alternative i, the alternative for which we are calculating the probability. Define $\tilde{U}_{nji} = U_{nj} - U_{ni}$, $\tilde{V}_{nji} = V_{nj} - V_{ni}$, and $\tilde{\varepsilon}_{nji} = \varepsilon_{nj} - \varepsilon_{ni}$. Then $P_{ni} = \text{Prob}(\tilde{U}_{nji} < 0 \ \forall j \neq i)$. That is, the probability of choosing alternative i is the probability that all the utility differences, when differenced against i, are negative. Define the vector $\tilde{\varepsilon}_{ni} = \langle \tilde{\varepsilon}_{n1i}, \ldots, \tilde{\varepsilon}_{nJ1} \rangle$ where the "..." is over all alternatives except i, so that $\tilde{\varepsilon}_{ni}$ has dimension $J-1$. Since the difference between two normals is normal, the density of the error differences is

$$\phi(\tilde{\varepsilon}_{ni}) = \frac{1}{(2\pi)^{-\frac{1}{2}(J-1)}|\tilde{\Omega}_i|^{1/2}} e^{-\frac{1}{2}\tilde{\varepsilon}_{ni}' \tilde{\Omega}_i \tilde{\varepsilon}_{ni}},$$

where $\tilde{\Omega}_i$ is the covariance of $\tilde{\varepsilon}_{ni}$, derived from Ω. Then the choice probability expressed in utility differences is

$$P_{ni} = \int I(\tilde{V}_{nji} + \tilde{\varepsilon}_{nji} < 0 \; \forall j \neq i)\phi(\tilde{\varepsilon}_{ni})\, d\tilde{\varepsilon}_{ni}, \tag{5.3}$$

which is a $(J-1)$-dimensional integral over all possible values of the error differences. An equivalent expression is

$$P_{ni} = \int_{\tilde{\varepsilon}_{ni} \in \tilde{B}_{ni}} \phi(\tilde{\varepsilon}_{ni})\, d\tilde{\varepsilon}_{ni}, \tag{5.4}$$

where $\tilde{B}_{ni} = \{\tilde{\varepsilon}_{ni} \text{ s.t. } \tilde{V}_{nji} + \tilde{\varepsilon}_{nji} < 0 \;\; \forall j \neq i\}$, which is a $(J-1)$-dimensional integral over the error differences in $\tilde{B}_{ni}$.

Expressions (5.3) and (5.4) utilize the covariance matrix $\tilde{\Omega}_i$ of the error differences. There is a straightforward way to derive $\tilde{\Omega}_i$ from the covariance of the errors themselves, Ω. Let M_i be the $(J-1)$ identity matrix with an extra column of -1's added as the ith column. The extra column makes the matrix have size $J-1$ by J. For example, with $J=4$ alternatives and $i=3$,

$$M_i = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & -1 & 1 \end{pmatrix}.$$

This matrix can be used to transform the covariance matrix of errors into the covariance matrix of error differences: $\tilde{\Omega}_i = M_i \Omega M_i'$. Note that $\tilde{\Omega}_i$ is $(J-1)\times(J-1)$ while Ω is $J \times J$, since M_i is $(J-1)\times J$. As an illustration, consider a three-alternative situation with errors $\langle \varepsilon_{n1}, \varepsilon_{n2}, \varepsilon_{n3} \rangle$ that have covariance

$$\Omega = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} \end{pmatrix}.$$

Suppose we takes differences against alternative 2. We know from first principles that the error differences $\langle \tilde{\varepsilon}_{n12}, \tilde{\varepsilon}_{n32} \rangle$ have covariance

$$\begin{aligned} \tilde{\Omega}_2 &= \mathrm{Cov}\begin{pmatrix} \varepsilon_{n1} - \varepsilon_{n2} \\ \varepsilon_{n3} - \varepsilon_{n2} \end{pmatrix} \\ &= \begin{pmatrix} \sigma_{11} + \sigma_{22} - 2\sigma_{12} & \sigma_{13} + \sigma_{22} - \sigma_{12} - \sigma_{23} \\ \sigma_{13} + \sigma_{22} - \sigma_{12} - \sigma_{23} & \sigma_{33} + \sigma_{22} - 2\sigma_{23} \end{pmatrix}. \end{aligned}$$

This covariance matrix can also be derived by the transformation $\tilde{\Omega}_2 = M_2 \Omega M_2'$:

$$\begin{aligned}
\tilde{\Omega}_n &= \begin{pmatrix} 1 & -1 & 0 \\ 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -1 & -1 \\ 0 & 1 \end{pmatrix} \\
&= \begin{pmatrix} \sigma_{11} - \sigma_{12} & \sigma_{12} - \sigma_{22} & \sigma_{13} - \sigma_{23} \\ -\sigma_{12} + \sigma_{13} & -\sigma_{22} + \sigma_{23} & -\sigma_{23} + \sigma_{33} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -1 & -1 \\ 0 & 1 \end{pmatrix} \\
&= \begin{pmatrix} \sigma_{11} - \sigma_{12} - \sigma_{12} + \sigma_{22} & -\sigma_{12} + \sigma_{22} + \sigma_{13} - \sigma_{23} \\ -\sigma_{12} + \sigma_{13} + \sigma_{22} - \sigma_{23} & \sigma_{22} - \sigma_{23} - \sigma_{23} + \sigma_{33} \end{pmatrix} \\
&= \begin{pmatrix} \sigma_{11} + \sigma_{22} - 2\sigma_{12} & \sigma_{13} + \sigma_{22} - \sigma_{12} - \sigma_{23} \\ \sigma_{13} + \sigma_{22} - \sigma_{12} - \sigma_{23} & \sigma_{33} + \sigma_{22} - 2\sigma_{23} \end{pmatrix}.
\end{aligned}$$

As we will see, this transformation by M_i comes in handy when simulating probit probabilities.

5.2 Identification

As described in Section 2.5, any discrete choice model must be normalized to take account of the fact that the level and scale of utility are irrelevant. The level of utility is immaterial because a constant can be added to the utility of all alternatives without changing which alternative has the highest utility: the alternative with the highest utility before the constant is added still has the highest utility afterward. Similarly, the scale of utility doesn't matter because the utility of each alternative can be multiplied by a (positive) constant without changing which alternative has the highest utility. In logit and nested logit models, the normalization for scale and level occurs automatically with the distributional assumptions that are placed on the error terms. As a result, normalization does not need to be considered explicitly for these models. With probit models, however, normalization for scale and level does not occur automatically. The researcher must normalize the model directly.

Normalization of the model is related to parameter identification. A parameter is *identified* if it can be estimated, and is *unidentified* if it cannot be estimated. An example of an unidentified parameter is k in the utility specification $U_{nj} = V_{nj} + k + \varepsilon_{nj}$. While the researcher might write utility in this way, and might want to estimate k to obtain a measure of the overall level of utility, doing so is impossible. The behavior of the decision maker is unaffected by k, and so the researcher cannot infer its

value from the choices that decision makers have made. Stated directly, parameters that do not affect the behavior of decision makers cannot be estimated. In an unnormalized model, parameters can appear that are not identified; these parameters relate to the scale and level of utility, which do not affect behavior. Once the model is normalized, these parameters disappear. The difficulty arises because it is not always obvious which parameters relate to scale and level. In the preceding example, the fact that k is unidentified is fairly obvious. In many cases, it is not at all obvious which parameters are identified. Bunch and Kitamura (1989) have shown that the probit models in several published articles are not normalized and contain unidentified parameters. The fact that neither the authors nor the reviewers of these articles could tell that the models were unnormalized is testimony to the complexity of the issue.

I provide in the following a procedure that can always be used to normalize a probit model and assure that all parameters are identified. It is not the only procedure that can be used; see, for example, Bunch (1991). In some cases a researcher might find other normalization procedures more convenient. However, the procedure I give can always be used, either by itself or as a check on another procedure.

I describe the procedure in terms of a four-alternative model. Generalization to more alternatives is obvious. As usual, utility is expressed as $U_{nj} = V_{nj} + \varepsilon_{nj}, \ j = 1, \ldots, 4$. The vector of errors is $\varepsilon_n' = \langle \varepsilon_{n1}, \ldots, \varepsilon_{n4} \rangle$. It is normally distributed with zero mean and a covariance matrix that can be expressed explicitly as

$$\Omega = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \cdot & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \cdot & \cdot & \sigma_{33} & \sigma_{34} \\ \cdot & \cdot & \cdot & \sigma_{44} \end{pmatrix}, \tag{5.5}$$

where the dots refer to the corresponding elements on the upper part of the matrix. Note that there are ten elements in this matrix, that is, ten distinct σ's representing the variances and covariances among the four errors. In general, a model with J alternatives has $J(J+1)/2$ distinct elements in the covariance matrix of the errors.

To take account of the fact that the level of utility is irrelevant, we take utility differences. In my procedure, I always take differences with respect to the first alternative, since that simplifies the analysis in a way that we will see. Define error differences as $\tilde{\varepsilon}_{nj1} = \varepsilon_{nj} - \varepsilon_{n1}$ for $j = 2, 3, 4$, and define the vector of error differences as $\tilde{\varepsilon}_{n1} = \langle \tilde{\varepsilon}_{n21}, \tilde{\varepsilon}_{n31}, \tilde{\varepsilon}_{n41} \rangle$. Note that the subscript 1 in $\tilde{\varepsilon}_{n1}$ means that the error differences are against the first alternative, rather than that the errors are for the first alternative.

The covariance matrix for the vector of error differences takes the form

$$\tilde{\Omega}_1 = \begin{pmatrix} \theta_{22} & \theta_{23} & \theta_{24} \\ \cdot & \theta_{33} & \theta_{34} \\ \cdot & \cdot & \theta_{44} \end{pmatrix},$$

where the θ's relate to the original σ's as follows:

$$\begin{aligned}
\theta_{22} &= \sigma_{22} + \sigma_{11} - 2\sigma_{12}, \\
\theta_{33} &= \sigma_{33} + \sigma_{11} - 2\sigma_{13}, \\
\theta_{44} &= \sigma_{44} + \sigma_{11} - 2\sigma_{14}, \\
\theta_{23} &= \sigma_{23} + \sigma_{11} - \sigma_{12} - \sigma_{13}, \\
\theta_{24} &= \sigma_{24} + \sigma_{11} - \sigma_{12} - \sigma_{14}, \\
\theta_{34} &= \sigma_{34} + \sigma_{11} - \sigma_{13} - \sigma_{14}.
\end{aligned}$$

Computationally, this matrix can be obtained using the transformation matrix M_i defined in Section 5.1 as $\tilde{\Omega}_1 = M_1 \Omega M_1'$.

To set the scale of utility, one of the diagonal elements is normalized. I set the top-left element of $\tilde{\Omega}_1$, which is the variance of $\tilde{\varepsilon}_{n21}$, to 1. This normalization for scale gives us the following covariance matrix:

$$\tilde{\Omega}_1^* = \begin{pmatrix} 1 & \theta_{23}^* & \theta_{24}^* \\ \cdot & \theta_{33}^* & \theta_{34}^* \\ \cdot & \cdot & \theta_{44}^* \end{pmatrix}. \tag{5.6}$$

The θ^*'s relate to the original σ's as follows:

$$\begin{aligned}
\theta_{33}^* &= \frac{\sigma_{33} + \sigma_{11} - 2\sigma_{13}}{\sigma_{22} + \sigma_{11} - 2\sigma_{12}}, \\
\theta_{44}^* &= \frac{\sigma_{44} + \sigma_{11} - 2\sigma_{14}}{\sigma_{22} + \sigma_{11} - 2\sigma_{12}}, \\
\theta_{23}^* &= \frac{\sigma_{23} + \sigma_{11} - \sigma_{12} - \sigma_{13}}{\sigma_{22} + \sigma_{11} - 2\sigma_{12}}, \\
\theta_{24}^* &= \frac{\sigma_{24} + \sigma_{11} - \sigma_{12} - \sigma_{14}}{\sigma_{22} + \sigma_{11} - 2\sigma_{12}}, \\
\theta_{34}^* &= \frac{\sigma_{34} + \sigma_{11} - \sigma_{13} - \sigma_{14}}{\sigma_{22} + \sigma_{11} - 2\sigma_{12}}.
\end{aligned}$$

There are five elements in $\tilde{\Omega}_1^*$. These are the only identified parameters in the model. This number is less than the ten elements that enter Ω. Each θ^* is a function of the σ's. Since there are five θ^*'s and ten σ's, it is not

possible to solve for all the σ's from estimated values of the θ^*'s. It is therefore not possible to obtain estimates of all the σ's.

In general, a model with J alternatives and an unrestricted covariance matrix will have $[(J-1)J/2]-1$ covariance parameters when normalized, compared to the $J(J+1)/2$ parameters when unnormalized. Only $[(J-1)J/2]-1$ parameters are identified. This reduction in the number of parameters is *not* a restriction. The reduction in the number of parameters is a normalization that simply eliminates irrelevant aspects of the original covariance matrix, namely the scale and level of utility. The ten elements in Ω allow for variance and covariance that is due simply to scale and level, which has no relevance for behavior. Only the five elements in $\tilde{\Omega}_1^*$ contain information about the variance and covariance of errors independent of scale and level. In this sense, only the five parameters have economic content, and only the five parameters can be estimated.

Suppose now that the researcher imposes structure on the covariance matrix. That is, instead of allowing a full covariance matrix for the errors, the researcher believes that the errors follow a pattern that implies particular values for, or relations among, the elements in the covariance matrix. The researcher restricts the covariance matrix to incorporate this pattern.

The structure can take various forms, depending on the application. Yai *et al.* (1997) estimate a probit model of route choice where the covariance between any two routes depends only on the length of shared route segments; this structure reduces the number of covariance parameters to only one, which captures the relation of the covariance to shared length. Bolduc *et al.* (1996) estimate a model of physicians' choice of location where the covariance among locations is a function of their proximity to one another, using what Bolduc (1992) has called a "generalized autoregressive" structure. Haaijer *et al.* (1998) impose a factor-analytic structure that arises from random coefficients of explanatory variables; this type of structure is described in detail in Section 5.3. Elrod and Keane (1995) impose a factor-analytic structure, but one that arises from error components rather than random coefficients *per se*.

Often the structure that is imposed will be sufficient to normalize the model. That is, the restrictions that the researcher imposes on the covariance matrix to fit her beliefs about the way the errors relate to each other will also serve to normalize the model. However, this is not always the case. The examples cited by Bunch and Kitamura (1989) are cases where the restrictions that the researcher placed on the covariance matrix seemed sufficient to normalize the model but actually were not.

The procedure that I give in the preceding text can be used to determine whether the restrictions on the covariance matrix are sufficient to normalize the model. The researcher specifies Ω with her restrictions on its elements. Then the stated procedure is used to derive $\tilde{\Omega}_1^*$, which is normalized for scale and level. We know that each element of $\tilde{\Omega}_1^*$ is identified. If each of the restricted elements of Ω can be calculated from the elements of $\tilde{\Omega}_1^*$, then the restrictions are sufficient to normalize the model. In this case, each parameter in the restricted Ω is identified. On the other hand, if the elements of Ω cannot be calculated from the elements of $\tilde{\Omega}_1^*$, then the restrictions are not sufficient to normalize the model and the parameters in Ω are not identified.

To illustrate this approach, suppose the researcher is estimating a four-alternative model and assumes that the covariance matrix for the errors has the following form:

$$\Omega = \begin{pmatrix} 1+\rho & \rho & 0 & 0 \\ \cdot & 1+\rho & 0 & 0 \\ \cdot & \cdot & 1+\rho & \rho \\ \cdot & \cdot & \cdot & 1+\rho \end{pmatrix}.$$

This covariance matrix allows the first and second errors to be correlated, the same as the third and fourth alternatives, but allows no other correlation. The correlation between the appropriate pairs is $\rho/(1+\rho)$. Note that by specifying the diagonal elements as $1+\rho$, the researcher assures that the correlation is between -1 and 1 for any value of ρ, as required for a correlation. Is this model, as specified, normalized for scale and level? To answer the question, we apply the described procedure. First, we take differences with respect to the first alternative. The covariance matrix of error differences is

$$\tilde{\Omega}_1 = \begin{pmatrix} \theta_{22} & \theta_{23} & \theta_{24} \\ \cdot & \theta_{33} & \theta_{34} \\ \cdot & \cdot & \theta_{44} \end{pmatrix},$$

where the θ's relate to the original σ's as follows:

$$\begin{aligned} \theta_{22} &= 2, \\ \theta_{33} &= 2+2\rho, \\ \theta_{44} &= 2+2\rho, \\ \theta_{23} &= 1, \\ \theta_{24} &= 1, \\ \theta_{34} &= 1+2\rho. \end{aligned}$$

We then normalize for scale by setting the top-left element to 1. The normalized covariance matrix is

$$\tilde{\Omega}_1^* = \begin{pmatrix} 1 & \theta_{23}^* & \theta_{24}^* \\ \cdot & \theta_{33}^* & \theta_{34}^* \\ \cdot & \cdot & \theta_{44}^* \end{pmatrix},$$

where the θ^*'s relate to the original σ's as follows:

$$\begin{aligned} \theta_{33}^* &= 1 + \rho, \\ \theta_{44}^* &= 1 + \rho, \\ \theta_{23}^* &= \tfrac{1}{2}, \\ \theta_{24}^* &= \tfrac{1}{2}, \\ \theta_{34}^* &= \tfrac{1}{2} + \rho. \end{aligned}$$

Note that $\theta_{33}^* = \theta_{44}^* = \theta_{34}^* - \frac{1}{2}$ and that the other θ^*'s have fixed values. There is one parameter in $\tilde{\Omega}_1^*$, as there is in Ω. Define $\theta = 1 + \rho$. Then $\tilde{\Omega}_1^*$ is

$$\tilde{\Omega}_1^* = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{2} \\ \cdot & \theta & \theta - \frac{1}{2} \\ \cdot & \cdot & \theta \end{pmatrix}.$$

The original ρ can be calculated directly from θ. For example, if θ is estimated to be 2.4, then the estimate of ρ is $\theta - 1 = 1.4$ and the correlation is $1.4/2.4 = .58$. The fact that the parameters that enter Ω can be calculated from the parameters that enter the normalized covariance matrix $\tilde{\Omega}_1^*$ means that the original model is normalized for scale and level. That is, the restrictions that the researcher placed on Ω also provided the needed normalization.

Sometimes restrictions on the original covariance matrix can appear to be sufficient to normalize the model when in fact they do not. Applying our procedure will determine whether this is the case. Consider the same model, but now suppose that the researcher allows a different correlation between the first and second errors than between the third and fourth errors. The covariance matrix of errors is specified to be

$$\Omega = \begin{pmatrix} 1+\rho_1 & \rho_1 & 0 & 0 \\ \cdot & 1+\rho_1 & 0 & 0 \\ \cdot & \cdot & 1+\rho_2 & \rho_2 \\ \cdot & \cdot & \cdot & 1+\rho_2 \end{pmatrix}.$$

The correlation between the first and second errors is $\rho_1/(1+\rho_1)$, and the correlation between the third and fourth errors is $\rho_2/(1+\rho_2)$. We can derive $\tilde{\Omega}_1$ for error differences and then derive $\tilde{\Omega}_1^*$ by setting the top-left element of $\tilde{\Omega}_1$ to 1. The resulting matrix is

$$\tilde{\Omega}_1^* = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{2} \\ \cdot & \theta & \theta - \frac{1}{2} \\ \cdot & \cdot & \theta \end{pmatrix},$$

where now $\theta = 1 + (\rho_1 + \rho_2)/2$. The values of ρ_1 and ρ_2 cannot be calculated from a value of θ. The original model is therefore not normalized for scale and level, and the parameters ρ_1 and ρ_2 are not identified. This fact is somewhat surprising, since only two parameters enter the original covariance matrix Ω. It would seem, unless the researcher explicitly tested in the manner we have just done, that restricting the covariance matrix to consist of only two elements would be sufficient to normalize the model. In this case, however, it is not.

In the normalized model, only the average of the ρ's appears: $(\rho_1 + \rho_2)/2$. It is possible to calculate the average ρ from θ, simply as $\theta - 1$. This means that the average ρ is identified, but not the individual values. When $\rho_1 = \rho_2$, as in the previous example, the model is normalized because each ρ is equal to the average ρ. However, as we now see, any model with the same average ρ's is equivalent, after normalizing for scale and level. Hence, assuming that $\rho_1 = \rho_2$ is no different than assuming that $\rho_1 = 3\rho_2$, or any other relation. All that matters for behavior is the average of these parameters, not their values relative to each other. This fact is fairly surprising and would be hard to realize without using our procedure for normalization.

Now that we know how to assure that a probit model is normalized for level and scale, and hence contains only economically meaningful information, we can examine how the probit model is used to represent various types of choice situations. We look at three situations in which logit models are limited and show how the limitation is overcome with probit. These situations are taste variation, substitution patterns, and repeated choices over time.

5.3 Taste Variation

Probit is particularly well suited for incorporating random coefficients, provided that the coefficients are normally distributed. Hausman and Wise (1978) were the first, to my knowledge, to give this derivation. Haaijer *et al.* (1998) provide a compelling application. Assume that representative utility is linear in parameters and that the coefficients

vary randomly over decision makers instead of being fixed as we have assumed so far in this book. The utility is $U_{nj} = \beta_n' x_{nj} + \varepsilon_{nj}$, where β_n is the vector of coefficients for decision maker n representing that person's tastes. Suppose the β_n is normally distributed in the population with mean b and covariance W: $\beta_n \sim N(b, W)$. The goal of the research is to estimate the parameters b and W.

The utility can be rewritten with β_n decomposed into its mean and deviations from its mean: $U_{nj} = b' x_{nj} + \tilde{\beta}_n' x_{nj} + \varepsilon_{nj}$, where $\tilde{\beta}_n = b - \beta_n$. The last two terms in the utility are random; denote their sum as η_{nj} to obtain $U_{nj} = b' x_{nj} + \eta_{nj}$. The covariance of the η_{nj}'s depends on W as well as the x_{nj}'s, so that the covariance differs over decision makers.

The covariance of the η_{nj}'s can be described easily for a two-alternative model with one explanatory variable. In this case, the utility is

$$U_{n1} = \beta_n x_{n1} + \varepsilon_{n1},$$
$$U_{n2} = \beta_n x_{n2} + \varepsilon_{n2}.$$

Assume that β_n is normally distributed with mean b and variance σ_β. Assume that ε_{n1} and ε_{n2} are identically normally distributed with variance σ_ε. The assumption of independence is for this example and is not needed in general. The utility is then rewritten as

$$U_{n1} = b x_{n1} + \eta_{n1},$$
$$U_{n2} = b x_{n2} + \eta_{n2},$$

where η_{n1} and η_{n2} are jointly normally distributed. Each has zero mean: $E(\eta_{nj}) = E(\tilde{\beta}_n x_{nj} + \varepsilon_{nj}) = 0$. The covariance is determined as follows. The variance of each is $V(\eta_{nj}) = V(\tilde{\beta}_n x_{nj} + \varepsilon_{nj}) = x_{nj}^2 \sigma_\beta + \sigma_\varepsilon$. Their covariance is

$$\begin{aligned}\mathrm{Cov}(\eta_{n1}, \eta_{n2}) &= E[(\tilde{\beta}_n x_{n1} + \varepsilon_{n1})(\tilde{\beta}_n x_{n2} + \varepsilon_{n2})] \\ &= E\left(\tilde{\beta}_n^2 x_{n1} x_{n2} + \varepsilon_{n1}\varepsilon_{n2} + \varepsilon_{n1}\tilde{\beta}_b x_{n2} + \varepsilon_{n2}\tilde{\beta}_n x_{n1}\right) \\ &= x_{n1} x_{n2} \sigma_\beta.\end{aligned}$$

The covariance matrix is

$$\begin{aligned}\Omega &= \begin{pmatrix} x_{n1}^2\sigma_\beta + \sigma_\varepsilon & x_{n1}x_{n2}\sigma_\beta \\ x_{n1}x_{n2}\sigma_\beta & x_{n2}^2\sigma_\beta + \sigma_\varepsilon \end{pmatrix} \\ &= \sigma_\beta \begin{pmatrix} x_{n1}^2 & x_{n1}x_{n2} \\ x_{n1}x_{n2} & x_{n2}^2 \end{pmatrix} + \sigma_\varepsilon \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.\end{aligned}$$

One last step is required for estimation. Recall that behavior is not affected by a multiplicative transformation of utility. We therefore need to set the scale of utility. A convenient normalization for this case is

$\sigma_\varepsilon = 1$. Under this normalization,

$$\Omega = \sigma_\beta \begin{pmatrix} x_{n1}^2 & x_{n1}x_{n2} \\ x_{n1}x_{n2} & x_{n2}^2 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

The values of x_{n1} and x_{n2} are observed by the researcher, and the parameters b and σ_β are estimated. Thus, the researcher learns both the mean and the variance of the random coefficient in the population. Generalization to more than one explanatory variable and more than two alternatives is straightforward.

5.4 Substitution Patterns and Failure of IIA

Probit can represent any substitution pattern. The probit probabilities do not exhibit the IIA property that gives rise to the proportional substitution of logit. Different covariance matrices Ω provide different substitution patterns, and by estimating the covariance matrix, the researcher determines the substitution pattern that is most appropriate for the data.

A full covariance matrix can be estimated, or the researcher can impose structure on the covariance matrix to represent particular sources of nonindependence. This structure usually reduces the number of the parameters and facilitates their interpretation. We consider first the situation where the researcher estimates a full covariance matrix, and then turn to a situation where the researcher imposes structure on the covariance matrix.

Full Covariance: Unrestricted Substitution Patterns

For notational simplicity, consider a probit model with four alternatives. A full covariance matrix for the unobserved components of utility takes the form of Ω in (5.5). When normalized for scale and level, the covariance matrix becomes $\tilde{\Omega}_1^*$ in (5.6). The elements of $\tilde{\Omega}_1^*$ are estimated. The estimated values can represent any substitution pattern; importantly, the normalization for scale and level does not restrict the substitution patterns. The normalization only eliminates aspects of Ω that are irrelevant to behavior.

Note, however, that the estimated values of the θ^*'s provide essentially no interpretable information in themselves (Horowitz, 1991). For example, suppose θ_{33}^* is estimated to be larger than θ_{44}^*. It might be tempting to interpret this result as indicating that the variance in unobserved utility of the third alternative is greater than that for the fourth alternative; that is, that $\sigma_{33} > \sigma_{44}$. However, this interpretation is incorrect. It is quite possible that $\theta_{33}^* > \theta_{44}^*$ and yet $\sigma_{44} > \sigma_{33}$, if the covariance σ_{13} is

sufficiently greater than σ_{14}. Similarly, suppose that θ_{23} is estimated to be negative. This does not mean that unobserved utility for the second alternative is negatively correlated with unobserved utility for the third alternative (that is, $\sigma_{23} < 0$). It is possible that σ_{23} is positive and yet σ_{12} and σ_{13} are sufficiently large to make θ^*_{23} negative. The point here is that estimating a full covariance matrix allows the model to represent any substitution pattern, but renders the estimated parameters essentially uninterpretable.

Structured Covariance: Restricted Substitution Patterns

By imposing structure on the covariance matrix, the estimated parameters usually become more interpretable. The structure is a restriction on the covariance matrix and, as such, reduces the ability of the model to represent various substitution patterns. However, if the structure is correct (that is, actually represents the behavior of the decision makers), then the true substitution pattern will be able to be represented by the restricted covariance matrix.

Structure is necessarily situation-dependent: an appropriate structure for a covariance matrix depends on the specifics of the situation being modeled. Several studies using different kinds of structure were described in Section 5.2. As an example of how structure can be imposed on the covariance matrix and hence substitution patterns, consider a homebuyer's choice among purchase-money mortgages. Suppose four mortgages are available to the homebuyer from four different institutions: one with a fixed rate, and three with variable rates. Suppose the unobserved portion of utility consists of two parts: the homebuyer's concern about the risk of rising interest rates, labeled r_n, which is common to all the variable-rate loans; and all other unobserved factors, labeled collectively η_{nj}. The unobserved component of utility is then

$$\varepsilon_{nj} = -r_n d_j + \eta_{nj},$$

where $d_j = 1$ for the variable-rate loans and 0 for the fixed-rate loan, and the negative sign indicates that utility decreases as concern about risk rises. Assume that r_n is normally distributed over homebuyers with variance σ, and that $\eta_{nj}\ \forall j$ is iid normal with zero mean and variance ω. Then the covariance matrix for $\varepsilon_n = \langle \varepsilon_{n1}, \ldots, \varepsilon_{n4} \rangle$ is

$$\Omega = \begin{pmatrix} 0 & 0 & 0 & 0 \\ \cdot & \sigma & \sigma & \sigma \\ \cdot & \cdot & \sigma & \sigma \\ \cdot & \cdot & \cdot & \sigma \end{pmatrix} + \omega \begin{pmatrix} 1 & 0 & 0 & 0 \\ \cdot & 1 & 0 & 0 \\ \cdot & \cdot & 1 & 0 \\ \cdot & \cdot & \cdot & 1 \end{pmatrix}.$$

The model needs to normalized for scale but, as we will see, is already normalized for level. The covariance of error differences is

$$\tilde{\Omega}_1 = \begin{pmatrix} \sigma & \sigma & \sigma \\ \cdot & \sigma & \sigma \\ \cdot & \cdot & \sigma \end{pmatrix} + \omega \begin{pmatrix} 2 & 1 & 1 \\ \cdot & 2 & 1 \\ \cdot & \cdot & 2 \end{pmatrix}.$$

This matrix has no fewer parameters than Ω. That is to say, the model was already normalized for level. To normalize for scale, set $\sigma + 2\omega = 1$. Then the covariance matrix becomes

$$\tilde{\Omega}_1^* = \begin{pmatrix} 1 & \theta & \theta \\ \cdot & 1 & \theta \\ \cdot & \cdot & 1 \end{pmatrix},$$

where $\theta = (\sigma + \omega)/(\sigma + 2\omega)$. The values of σ and ω cannot be calculated from θ. However, the parameter θ provides information about the variance in utility due to concern about risk relative to that due to all other unobserved factors. For example, suppose θ is estimated to be 0.75. This estimate can be intrepreted as indicating that the variance in utility attributable to concern about risk is twice as large as the variance in utility attributable to all other factors:

$$\begin{aligned} \theta &= 0.75, \\ \frac{\sigma + \omega}{\sigma + 2\omega} &= 0.75, \\ \sigma + \omega &= 0.75\sigma + 1.5\omega, \\ 0.25\sigma &= 0.5\omega, \\ \sigma &= 2\omega. \end{aligned}$$

Stated equivalently, $\hat{\theta} = 0.75$ means that concern about risk accounts for two-thirds of the variance in the unobserved component of utility.

Since the original model was already normalized for level, the model could be estimated without reexpressing the covariance matrix in terms of error differences. The normalization for scale could be accomplished simply by setting $\omega = 1$ in the original Ω. Under this procedure, the parameter σ is estimated directly. Its value relative to 1 indicates the variance due to concern about risk relative to the variance due to perceptions about ease of dealing with each institution. An estimate $\hat{\theta} = 0.75$ corresponds to an estimate $\hat{\sigma} = 2$.

5.5 Panel Data

Probit with repeated choices is similar to probit on one choice per decision maker. The only difference is that the dimension of the covariance

matrix of the errors is expanded. Consider a decision maker who faces a choice among J alternatives in each of T time periods or choices situations. The alternatives can change over time, and J and T can differ for different decision makers; however, we suppress the notation for these possibilities. The utility that decision maker n obtains from alternative j in period t is $U_{njt} = V_{njt} + \varepsilon_{njt}$. In general, one would expect ε_{njt} to be correlated over time as well as over alternatives, since factors that are not observed by the researcher can persist over time. Denote the vector of errors for all alternatives in all time periods as $\varepsilon_n = \langle \varepsilon_{n11}, \ldots, \varepsilon_{nJ1}, \varepsilon_{n12}, \ldots, \varepsilon_{nJ2}, \ldots, \varepsilon_{n1T}, \ldots, \varepsilon_{nJT} \rangle$. The covariance matrix of this vector is denoted Ω, which has dimension $JT \times JT$.

Consider a sequence of alternatives, one for each time period, $\mathbf{i} = \{i_1, \ldots, i_T\}$. The probability that the decision maker makes this sequence of choices is

$$\begin{aligned} P_{n\mathbf{i}} &= \text{Prob}(U_{ni_t t} > U_{njt} \; \forall j \neq i_t, \forall t) \\ &= \text{Prob}(V_{ni_t t} + \varepsilon_{ni_t t} > V_{njt} + \varepsilon_{njt} \; \forall j \neq i_t, \forall t) \\ &= \int_{\varepsilon_n \in B_n} \phi(\varepsilon_n)\, d\varepsilon_n. \end{aligned}$$

where $B_n = \{\varepsilon_n \text{ s.t. } V_{ni_t t} + \varepsilon_{ni_t t} > V_{njt} + \varepsilon_{njt} \; \forall j \neq i_t, \forall t\}$ and $\phi(\varepsilon_n)$ is the joint normal density with zero mean and covariance Ω. Compared to the probit probability for one choice situation, the integral is simply expanded to be over JT dimensions rather than J.

It is often more convenient to work in utility differences. The probability of sequence $\mathbf{i}$ is the probability that the utility differences are negative for each alternative in each time period, when the differences in each time period are taken against the alternative identified by $\mathbf{i}$ for that time period:

$$\begin{aligned} P_{n\mathbf{i}} &= \text{Prob}(\tilde{U}_{nji_t t} < 0 \; \forall j \neq i_t, \forall t) \\ &= \int_{\tilde{\varepsilon}_n \in \tilde{B}_n} \phi(\tilde{\varepsilon}_n)\, d\tilde{\varepsilon}_n, \end{aligned}$$

where $\tilde{U}_{nji_t t} = U_{njt} - U_{ni_t t}$; $\tilde{\varepsilon}_n' = \langle (\varepsilon_{n11} - \varepsilon_{ni_1 1}), \ldots, (\varepsilon_{nJ1} - \varepsilon_{ni_1 1}), \ldots, (\varepsilon_{n1T} - \varepsilon_{ni_T T}), \ldots, (\varepsilon_{nJT} - \varepsilon_{ni_T T}) \rangle$ with each $\ldots$ being over all alternatives except i_t, and the matrix $\tilde{B}_n$ is the set of $\tilde{\varepsilon}_n$'s for which $\tilde{U}_{nji_t t} < 0 \; \forall j \neq i_t, \; \forall t$. This is a $(J-1)T$-dimensional integral. The density $\phi(\tilde{\varepsilon}_n)$ is joint normal with covariance matrix derived from Ω. The simulation of the choice probability is the same as for situations with one choice per decision maker, which we describe in Section 5.6, but with a larger dimension for the covariance matrix and integral. Borsch-Supan *et al.* (1991) provide an example of a multinomial probit on panel data that allows covariance over time and over alternatives.

For binary choices, such as whether a person buys a particular product in each time period or works at a paid job each month, the probit model simplifies considerably (Gourieroux and Monfort, 1993). The net utility of taking the action (e.g., working) in period t is $U_{nt} = V_{nt} + \varepsilon_{nt}$, and the person takes the action if $U_{nt} > 0$. This utility is called *net* utility because it is the difference between the utility of taking the action and that of not taking the action. As such, it is already expressed in difference terms. The errors are correlated over time, and the covariance matrix for $\varepsilon_{n1}, \ldots, \varepsilon_{nT}$ is Ω, which is $T \times T$.

A sequence of binary choices is most easily represented by a set of T dummy variables: $d_{nt} = 1$ if person n took the action in period t, and $d_{nt} = -1$ otherwise. The probability of the sequence of choices $d_n = d_{n1}, \ldots, d_{nT}$ is

$$\begin{aligned} P_{nd_n} &= \text{Prob}(U_{nt}d_{nt} > 0 \; \forall t) \\ &= \text{Prob}(V_{nt}d_{nt} + \varepsilon_{nt}d_{nt} > 0 \; \forall t) \\ &= \int_{\varepsilon_n \in B_n} \phi(\varepsilon_n)\, d\varepsilon_n, \end{aligned}$$

where B_n is the set of ε_n's for which $V_{nt}d_{nt} + \varepsilon_{nt}d_{nt} > 0 \; \forall t$, and $\phi(\varepsilon_n)$ is the joint normal density with covariance Ω.

Structure can be placed on the covariance of the errors over time. Suppose in the binary case, for example, that the error consists of a portion that is specific to the decision maker, reflecting his proclivity to take the action, and a part that varies over time for each decision maker: $\varepsilon_{nt} = \eta_n + \mu_{nt}$, where μ_{nt} is iid over time and people with a standard normal density, and η_n is iid over people with a normal density with zero mean and variance σ. The variance of the error in each period is $V(\varepsilon_{nt}) = V(\eta_n + \mu_{nt}) = \sigma + 1$. The covariance between the errors in two different periods t and s is $\text{Cov}(\varepsilon_{nt}, \varepsilon_{ns}) = E(\eta_n + \mu_{nt})(\eta_n + \mu_{ns}) = \sigma$. The covariance matrix therefore takes the form

$$\Omega = \begin{pmatrix} \sigma+1 & \sigma & \cdots & \cdots & \sigma \\ \sigma & \sigma+1 & \sigma & \cdots & \sigma \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \sigma & \cdots & \cdots & \sigma & \sigma+1 \end{pmatrix}.$$

Only one parameter, σ, enters the covariance matrix. Its value indicates the variance in unobserved utility across individuals (the variance of η_n) relative to the variance across time for each individual (the variance of μ_{nt}). It is often called the *cross-subject variance relative to the within-subject variance.*

The choice probabilities under this structure on the errors can be easily simulated using the concepts of convenient error partitioning from Section 1.2. Conditional on η_n, the probability of *not* taking the action in period t is $\text{Prob}(V_{nt} + \eta_n + \mu_{nt} < 0) = \text{Prob}(\mu_{nt} < -(V_{nt} + \eta_n)) = \Phi(-(V_{nt} + \eta_n))$, where $\Phi(\cdot)$ is the cumulative standard normal function. Most software packages include routines to calculate this function. The probability of taking the action, conditional on η_n, is then $1 - \Phi(-(V_{nt} + \eta_n)) = \Phi(V_{nt} + \eta_n)$. The probability of the sequence of choices d_n, conditional on η_n, is therefore $\prod_t \Phi((V_{nt} + \eta_n)d_{nt})$, which we can label $H_{nd_n}(\eta_n)$.

So far we have conditioned on η_n, when in fact η_n is random. The *unconditional* probability is the integral of the conditional probability $H_{nd_n}(\eta_n)$ over all possible values of η_n:

$$P_{nd_n} = \int H_{nd_n}(\eta_n)\phi(\eta_n)\, d\eta_n$$

where $\phi(\eta_n)$ is the normal density with zero mean and variance σ. This probability can be simulated very simply as follows:

1. Take a draw from a standard normal density using a random number generator. Multiply the draw by $\sqrt{\sigma}$, so that it becomes a draw of η_n from a normal density with variance σ.
2. For this draw of η_n, calculate $H_{nd_n}(\eta_n)$.
3. Repeat steps 1–2 many times, and average the results. This average is a simulated approximation to P_{nd_n}.

This simulator is much easier to calculate than the general probit simulators described in the next section. The ability to use it arises from the structure that we imposed on the model, namely, that the time dependence of the unobserved factors is captured entirely by a random component η_n that remains constant over time for each person. Gourieroux and Monfort (1993) provide an example of the use of this simulator with a probit model of this form.

The representative utility in one time period can include exogenous variables for other time periods, the same as we discussed with respect to logit models on panel data (Section 3.3.3). That is, V_{nt} can include exogenous variables that relate to periods other than t. For example, a lagged response to price changes can be represented by including prices from previous periods in the current period's V. Anticipatory behavior (by which, for example, a person buys a product now because he correctly anticipates that the price will rise in the future) can be represented by including prices in future periods in the current period's V.

Entering a lagged dependent variable is possible, but introduces two difficulties that the researcher must address. First, since the errors are correlated over time, the choice in one period is correlated with the errors in subsequent periods. As a result, inclusion of a lagged dependent variable without adjusting the estimation procedure appropriately results in inconsistent estimates. This issue is analogous to regression analysis, where the ordinary least squares estimator is inconsistent when a lagged dependent variable is included and the errors are serially correlated. To estimate a probit consistently in this situation, the researcher must determine the distribution of each ε_{nt} conditional on the value of the lagged dependent variables. The choice probability is then based on this conditional distribution instead of the unconditional distribution $\phi(\cdot)$ that we used earlier. Second, often the researcher does not observe the decision makers' choices from the very first choice that was available to them. For example, a researcher studying employment patterns will perhaps observe a person's employment status over a period of time (e.g., 1998–2001), but usually will not observe the person's employment status starting with the very first time the person could have taken a job (which might precede 1998 by many years). In this case, the probability for the first period that the researcher observes depends on the choices of the person in the earlier periods that the researcher does not observe. The researcher must determine a way to represent the first choice probability that allows for consistent estimation in the face of missing data on earlier choices. This is called the *initial conditions problem* of dynamic choice models. Both of these issues, as well as potential approaches to dealing with them, are addressed by Heckman (1981b, 1981a) and Heckman and Singer (1986). Due to their complexity, I do not describe the procedures here and refer interested and brave readers to these articles.

Papatla and Krishnamurthi (1992) avoid these issues in their probit model with lagged dependent variables by assuming that the unobserved factors are independent over time. As we discussed in relation to logit on panel data (Section 3.3.3), lagged dependent variables are not correlated with the current errors when the errors are independent over time, and they can therefore be entered without inducing inconsistency. Of course, this procedure is only appropriate if the assumption of errors being independent over time is true in reality, rather than just by assumption.

5.6 Simulation of the Choice Probabilities

The probit probabilities do not have a closed-form expression and must be approximated numerically. Several nonsimulation procedures have been used and can be effective in certain circumstances.

Quadrature methods approximate the integral by a weighted function of specially chosen evaluation points. A good explanation for these procedures is provided by Geweke (1996). Examples of their use for probit include Butler and Moffitt (1982) and Guilkey and Murphy (1993). Quadrature operates effectively when the dimension of the integral is small, but not with higher dimensions. It can be used for probit if the number of alternatives (or, with panel data, the number of alternatives times the number of time periods) is no more than four or five. It can also be used if the researcher has specified an error-component structure with no more than four or five terms. However, it is not effective for general probit models. And even with low-dimensional integration, simulation is often easier.

Another nonsimulation procedure that has been suggested is the Clark algorithm, introduced by Daganzo *et al.* (1977). This algorithm utilizes the fact, shown by Clark (1961), that the maximum of several normally distributed variables is itself approximately normally distributed. Unfortunately, the approximation can be highly inaccurate in some situations (as shown by Horowitz *et al.*, 1982), and the degree of accuracy is difficult to assess in any given setting.

Simulation has proven to be very general and useful for approximating probit probabilities. Numerous simulators have been proposed for probit models; a summary is given by Hajivassiliou *et al.* (1996). In the preceding section, I described a simulator that is appropriate for a probit model that has a particularly convenient structure, namely a binary probit on panel data where the time dependence is captured by one random factor. In the current section, I describe three simulators that are applicable for probits of any form: accept–reject, smoothed accept–reject, and GHK. The GHK simulator is by far the most widely used probit simulator, for reasons that we discuss. The other two methods are valuable pedagogically. They also have relevance beyond probit and can be applied in practically any situation. They can be very useful when the researcher is developing her own models rather than using probit or any other model in this book.

5.6.1. Accept–Reject Simulator

The accept–reject (AR) is the most straightforward simulator. Consider simulating P_{ni}. Draws of the random terms are taken from their distributions. For each draw, the researcher determines whether those values of the errors, when combined with the observed variables as faced by person n, would result in alternative i being chosen. If so, the draw is called an *accept*. If the draw would result in some other

alternative being chosen, the draw is a *reject*. The simulated probability is the proportion of draws that are accepts. This procedure can be applied to any choice model with any distribution for the random terms. It was originally proposed for probits (Manski and Lerman, 1981), and we give the details of the approach in terms of the probit model. Its use for other models is obvious.

We use expression (5.1) for the probit probabilities:

$$P_{ni} = \int I(V_{ni} + \varepsilon_{ni} > V_{nj} + \varepsilon_{nj} \ \forall j \neq i)\phi(\varepsilon_n)\, d\varepsilon_n,$$

where $I(\cdot)$ is an indicator of whether the statement in parentheses holds, and $\phi(\varepsilon_n)$ is the joint normal density with zero mean and covariance Ω. The AR simulator of this integral is calculated as follows:

1. Draw a value of the J-dimensional vector of errors, ε_n, from a normal density with zero mean and covariance Ω. Label the draw ε_n^r with $r = 1$, and the elements of the draw as $\varepsilon_{n1}^r, \ldots, \varepsilon_{nJ}^r$.
2. Using these values of the errors, calculate the utility that each alternative obtains with these errors. That is, calculate $U_{nj}^r = V_{nj} + \varepsilon_{nj}^r \ \forall j$.
3. Determine whether the utility of alternative i is greater than that for all other alternatives. That is, calculate $I^r = 1$ if $U_{ni}^r > U_{nj}^r$, indicating an accept, and $I^r = 0$ otherwise, indicating a reject.
4. Repeat steps 1–3 many times. Label the number of repetitions (including the first) as R, so that r takes values of 1 through R.
5. The simulated probability is the proportion of draws that are accepts: $\check{P}_{ni} = \frac{1}{R}\sum_{r=1}^{R} I^r$.

The integral $\int I(\cdot)\phi(\varepsilon_n)\, d\varepsilon$ is approximated by the average $\frac{1}{R}\sum I^r(\cdot)$ for draws from $\phi(\cdot)$. Obviously, $\check{P}_{ni}$ is unbiased for P_{ni}: $E(\check{P}_{ni}) = \frac{1}{R}\sum E[I^r(\cdot)] = \frac{1}{R}\sum P_{ni} = P_{ni}$, where the expectation is over different sets of R draws. The variance of $\check{P}_{ni}$ over different sets of draws diminishes as the number of draws rises. The simulator is often called the "crude frequency simulator," since it is the frequency of times that draws of the errors result in the specified alternative being chosen. The word "crude" distinguishes it from the *smoothed* frequency simulator that we describe in the next section.

The first step of the AR simulator for a probit model is to take a draw from a joint normal density. The question arises: how are such draws obtained? The most straightforward procedure is that described in Section 9.2.5, which uses the Choleski factor. The covariance matrix for the errors is Ω. A Choleski factor of Ω is a lower-triangular matrix L such that $LL' = \Omega$. It is sometimes called the generalized square root of

Ω. Most statistical software packages contain routines to calculate the Choleski factor of any symmetric matrix. Now suppose that η is a vector of J iid standard normal deviates such that $\eta \sim N(0, I)$, where I is the identity matrix. This vector can be obtained by taking J draws from a random number generator for the standard normal and stacking them into a vector. We can construct a vector ε that is distributed $N(O, \Omega)$ by using the Choleski factor to tranform η. In particular, calculate $\varepsilon = L\eta$. Since the sum of normals is normal, ε is normally distributed. Since η has zero mean, so does ε. The covariance of ε is $\text{Cov}(\varepsilon) = E(\varepsilon\varepsilon') = E(L\eta(L\eta)') = E(L\eta\eta'L') = LE(\eta\eta')L' = LIL' = LL' = \Omega$.

Using the Choleski factor L of Ω, the first step of the AR simulator becomes two substeps:

1A. Draw J values from a standard normal density, using a random number generator. Stack these values into a vector, and label the vector η^r.
1B. Calculate $\varepsilon_n^r = L\eta^r$.

Then, using ε_n^r, calculate the utility of each alternative and see whether alternative i has the highest utility.

The procedure that we have described operates on utilities and expression (5.1), which is a J-dimensional integral. The procedure can be applied analogously to utility differences, which reduces the dimension of the integral to $J - 1$. As given in (5.3), the choice probabilities can be expressed in terms of utility differences:

$$P_{ni} = \int I(\tilde{V}_{nji} + \tilde{\varepsilon}_{nji} < 0 \; \forall j \neq i)\phi(\tilde{\varepsilon}_{ni})\, d\tilde{\varepsilon}_{ni},$$

where $\phi(\tilde{\varepsilon}_{ni})$ is the joint normal density with zero mean and covariance $\tilde{\Omega}_i = M_i \Omega M_i'$. This integral can be simulated with AR methods through the following steps:

1. Draw $\tilde{\varepsilon}_{ni}^r = L_i \eta^r$ as follows:
   (a) Draw $J - 1$ values from a standard normal density using a random number generator. Stack these values into a vector, and label the vector η^r.
   (b) Calculate $\tilde{\varepsilon}_{ni}^r = L_i \eta^r$, where L_i is the Choleski factor of $\tilde{\Omega}_i$.
2. Using these values of the errors, calculate the utility difference for each alternative, differenced against the utility of alternative i. That is, calculate $\tilde{U}_{nji}^r = V_{nj} - V_{ni} + \tilde{\varepsilon}_{nji}^r \; \forall j \neq i$.
3. Determine whether each utility difference is negative. That is, calculate $I^r = 1$ if $U_{nji}^r < 0 \; \forall j \neq i$, indicating an accept, and $I^r = 0$ otherwise, indicating a reject.

4. Repeat steps 1–3 R times.
5. The simulated probability is the number of accepts divided by the number of repetitions: $\check{P}_{ni} = \frac{1}{R}\sum_{r=1}^{R} I^r$.

Using utility differences is slightly faster computationally than using the utilities themselves, since one dimension is eliminated. However, it is often easier conceptually to remain with utilities.

As just stated, the AR simulator is very general. It can be applied to any model for which draws can be obtained for the random terms and the behavior that the decision maker would exhibit with these draws can be determined. It is also very intuitive, which is an advantage from a programming perspective, since debugging becomes comparatively easy. However, the AR simulator has several disadvantages, particularly when used in the context of maximum likelihood estimation.

Recall that the log-likelihood function is $\text{LL} = \sum_n \sum_j d_{nj} \log P_{nj}$, where $d_{nj} = 1$ if n chose j and 0 otherwise. When the probabilities cannot be calculated exactly, as in the case of probit, the simulated log-likelihood function is used instead, with the true probabilities replaced with the simulated probabilities: $\text{SLL} = \sum_n \sum_j d_{nj} \log \check{P}_{nj}$. The value of the parameters that maximizes SLL is called the maximum simulated likelihood estimator (MSLE). It is by far the most widely used simulation-based estimation procedure. Its properties are described in Chapter 8. Unfortunately, using the AR simulator in SLL can be problematic.

There are two issues. First, $\check{P}_{ni}$ can be zero for any finite number of draws R. That is, it is possible that each of the R draws of the error terms result in a reject, so that the simulated probability is zero. Zero values for $\check{P}_{ni}$ are problematic because the log of $\check{P}_{ni}$ is taken when it enters the log-likelihood function and the log of zero is undefined. SLL cannot be calculated if the simulated probability is zero for any decision maker in the sample.

The occurrence of a zero simulated probability is particularly likely when the true probability is low. Often at least one decision maker in a sample will have made a choice that has a low probability. With numerous alternatives (such as thousands of makes and models for the choice of car), each alternative has a low probability. With repeated choices, the probability for any sequence of choices can be extremely small; for example, if the probability of choosing an alternative is 0.25 in each of 10 time periods, the probability of the sequence is $(0.25)^{10}$, which is less than 0.000001.

Furthermore, SLL needs to be calculated at each step in the search for its maximum. Some of the parameter values at which SLL is

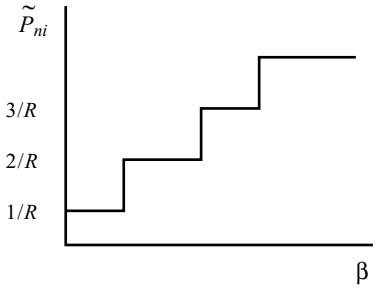

Figure 5.1. The AR simulator is a step function in parameters.

calculated can be far from the true values. Low probabilities can occur at these parameter values even when they do not occur at the maximizing values.

Nonzero simulated probabilities can always be obtained by taking enough draws. However, if the researcher continues taking draws until at least one accept is obtained for each decision maker, then the number of draws becomes a function of the probabilities. The simulation process is then not independent of the choice process that is being modeled, and the properties of the estimator become more complex.

There is a second difficulty with the AR simulator for MSLE. The simulated probabilities are not smooth in the parameters; that is, they are not twice differentiable. As explained in Chapter 8, the numerical procedures that are used to locate the maximum of the log-likelihood function rely on the first derivatives, and sometimes the second derivatives, of the choice probabilities. If these derivatives do not exist, or do not point toward the maximum, then the numerical procedure will not perform effectively.

The AR simulated probability is a step function, as depicted in Figure 5.1. $\check{P}_{ni}$ is the proportion of draws for which alternative i has the highest utility. An infinitesimally small change in a parameter will usually not change any draw from a reject to an accept or vice versa. If U_{ni}^r is below U_{nj}^r for some j at a given level of the parameters, then it will also be so for an infinitesimally small change in any parameter. So, usually, $\check{P}_{nj}$ is constant with respect to small changes in the parameters. Its derivatives with respect to the parameters are zero in this range. If the parameters change in such a way that a reject becomes an accept, then $\check{P}_{nj}$ rises by a discrete amount, from M/R to $(M+1)/R$, where M is the number of accepts at the original parameter values. $\check{P}_{nj}$ is constant (zero slope) until an accept becomes a reject or vice versa, at which point $\check{P}_{nj}$ jumps by $1/R$. Its slope at this point is undefined. The first derivative of $\check{P}_{nj}$ with respect to the parameters is either zero or undefined.

This fact hinders the numerical procedures that are used to locate the maximum of SLL. As discussed in Chapter 8, the maximization procedures use the gradient at trial parameter values to determine the direction to move to find parameters with higher SLL. With the slope $\check{P}_{nj}$ for each n either zero or undefined, the gradient of SLL is either zero or undefined. This gradient provides no help in finding the maximum.

This problem is not actually as drastic as it seems. The gradient of SLL can be approximated as the change in SLL for a non-infinitesimally small change in the parameters. The parameters are changed by an amount that is large enough to switch accepts to rejects and vice versa for at least some of the observations. The approximate gradient, which can be called an arc gradient, is calculated as the amount that SLL changes divided by the change in the parameters. To be precise: for parameter vector β of length K, the derivate of SLL with respect to the kth parameter is calculated as $(\text{SLL}^1 - \text{SLL}^0)/(\beta_k^1 - \beta_k^0)$, where SLL^0 is calculated at the original β with kth element β_k^0 and SLL^1 is calculated at β_k^1 with all the other parameters remaining at their original values. The arc gradient calculated in this way is not zero or undefined, and provides information on the direction of rise. Nevertheless, experience indicates that the AR simulated probability is still difficult to use.

5.6.2. Smoothed AR Simulators

One way to mitigate the difficulties with the AR simulator is to replace the 0–1 AR indicator with a smooth, strictly positive function. The simulation starts the same as with AR, by taking draws of the random terms and calculating the utility of each alternative for each draw: U_{nj}^r. Then, instead of determining whether alternative i has the highest utility (that is, instead of calculating the indicator function I^r), the simulated utilities $U_{nj}^r \ \forall j$ are entered into a function. Any function can be used for simulating P_{ni} as long as it rises when U_{ni}^r rises, declines when U_{nj}^r rises, is strictly positive, and has defined first and second derivatives with respect to $U_{nj}^r \ \forall j$. A function that is particularly convenient is the logit function, as suggested by McFadden (1989). Use of this function gives the *logit-smoothed AR simulator*.

The simulator is implemented in the following steps, which are the same as with the AR simulator except for step 3.

1. Draw a value of the J-dimensional vector of errors, ε_n, from a normal density with zero mean and covariance Ω. Label the draw ε_n^r with $r = 1$, and the elements of the draw as $\varepsilon_{n1}^r, \ldots, \varepsilon_{nJ}^r$.

2. Using these values of the errors, calculate the utility that each alternative obtains with these errors. That is, calculate $U_{nj}^r = V_{nj} + \varepsilon_{nj}^r \; \forall j$.
3. Put these utilities into the logit formula. That is, calculate

$$S_r = \frac{e^{U_{ni}^r/\lambda}}{\sum_j e^{U_{nj}^r/\lambda}},$$

where $\lambda > 0$ is a scale factor specified by the researcher and discussed in following text.
4. Repeat steps 1–3 many times. Label the number of repetitions (including the first) as R, so that r takes values of 1 through R.
5. The simulated probability is the number of accepts divided by the number of repetitions: $\breve{P}_{ni} = \frac{1}{R}\sum_{r=1}^{R} S^r$.

Since $S^r > 0$ for all finite values of U_{nj}^r, the simulated probability is strictly positive for any draws of the errors. It rises with U_{ni}^r and declines when U_{nj}^r, $j \neq i$, rises. It is smooth (twice differentiable), since the logit formula itself is smooth.

The logit-smoothed AR simulator can be applied to any choice model, simply by simulating the utilities under any distributional assumptions about the errors and then inserting the utilities into the logit formula. When applied to probit, Ben-Akiva and Bolduc (1996) have called it "logit-kernel probit."

The scale factor λ determines the degree of smoothing. As $\lambda \to 0$, S^r approaches the indicator function I^r. Figure 5.2 illustrates the situation for a two-alternative case. For a given draw of ε_n^r, the utility of the two alternatives is calculated. Consider the simulated probability for alternative 1. With AR, the 0–1 indicator function is zero if U_{n1}^r is below U_{n2}^r, and one if U_{n1}^r exceeds U_{n2}^r. With logit smoothing, the step function is replaced by a smooth sigmoid curve. The factor λ determines the proximity of the sigmoid to the 0–1 indicator. Lowering λ increases the scale of the utilities when they enter the logit function (since the utilities are divided by λ). Increasing the scale of utility increases the absolute difference between the two utilities. The logit formula gives probabilities that are closer to zero or one when the difference in utilities is larger. The logit-smoothed S^r therefore becomes closer to the step function as λ becomes closer to zero.

The researcher needs to set the value of λ. A lower value of λ makes the logit smoother a better approximation to the indicator function. However, this fact is a double-edged sword: if the logit smoother approximates the indicator function too well, the numerical difficulties of using

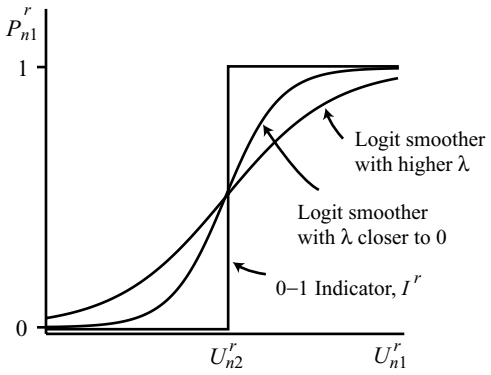

Figure 5.2. AR smoother.

the unsmoothed AR simulator will simply be reproduced in the logit-smoothed simulator. The researcher wants to set λ low enough to obtain a good approximation but not so low as to reintroduce numerical difficulties. There is little guidance on the appropriate level of λ. Perhaps the best approach is for the researcher to experiment with different λ's. The same draws of ε_n should be used with every λ, so as to assure that differences in results are due to the change in the λ rather than to differences in the draws.

McFadden (1989) describes other smoothing functions. For all of them, the researcher must specify the degree of smoothing. An advantage of the logit smoother is its simplicity. Also, we will see in Chapter 6 that the logit smoother applied to a probit or any other model constitutes a type of mixed logit specification. That is, instead of seeing the logit smoother as providing an approximation that has no behavioral relation to the model (simply serving a numerical purpose), we can see it as arising from a particular type of error structure in the behavioral model itself. Under this interpretation, the logit formula applied to simulated utilities is not an approximation but actually represents the true model.

5.6.3. GHK Simulator

The most widely used probit simulator is called GHK, after Geweke (1989, 1991), Hajivassiliou (as reported in Hajivassiliou and McFadden, 1998), and Keane (1990, 1994), who developed the procedure. In a comparison of numerous probit simulators, Hajivassiliou *et al.* (1996) found GHK to be the most accurate in the settings that they examined. Geweke *et al.* (1994) found the GHK simulator works better than smoothed AR. Experience has confirmed its usefulness and relative accuracy (e.g., Borsch-Supan and Hajivassiliou, 1993).

The GHK simulator operates on utility differences. The simulation of probability P_{ni} starts by subtracting the utility of alternative i from each other alternative's utility. Importantly, the utility of a different alternative is subtracted depending on which probability is being simulated: for P_{ni}, U_{ni} is subtracted from the other utilities, while for P_{nj}, U_{nj} is subtracted. This fact is critical to the implementation of the procedure.

I will explain the GHK procedure first in terms of a three-alternative case, since that situation can be depicted graphically in two dimensions for utility differences. I will then describe the procedure in general for any number of alternatives. Bolduc (1993, 1999) provides an excellent alternative description of the procedure, along with methods to simulate the analytic derivatives of the probit probabilities. Keane (1994) provides a description of the use of GHK for transition probabilities.

Three Alternatives

We start with a specification of the behavioral model in utilities: $U_{nj} = V_{nj} + \varepsilon_{nj},\ j = 1, 2, 3$. The vector $\varepsilon'_n = \langle \varepsilon_{n1}, \varepsilon_{n2}, \varepsilon_{n3} \rangle \sim N(0, \Omega)$. We assume that the reseacher has normalized the model for scale and level, so that the parameters that enter Ω are identified. Also, Ω can be a parametric function of data, as with random taste variation, though we do not show this dependence in our notation.

Suppose we want to simulate the probability of the first alternative, P_{n1}. We reexpress the model in utility differences by subtracting the utility of alternative 1:

$$\begin{aligned} U_{nj} - U_{n1} &= (V_{nj} - V_{n1}) + (\varepsilon_{nj} - \varepsilon_{n1}), \\ \tilde{U}_{nj1} &= \tilde{V}_{nj1} + \tilde{\varepsilon}_{nj1}, \end{aligned}$$

for $j = 2, 3$. The vector $\tilde{\varepsilon}'_{n1} = \langle \tilde{\varepsilon}_{n21}, \tilde{\varepsilon}_{n31} \rangle$ is distributed $N(0, \tilde{\Omega}_1)$, where $\tilde{\Omega}_1$ is derived from Ω.

We take one more transformation to make the model more convenient for simulation. Namely, let L_1 be the Choleski factor of $\tilde{\Omega}_1$. Since $\tilde{\Omega}_1$ is 2×2 in our current illustration, L_1 is a lower-triangular matrix that takes the form

$$L_1 = \begin{pmatrix} c_{aa} & 0 \\ c_{ab} & c_{bb} \end{pmatrix}.$$

Using this Choleski factor, the original error differences, which are correlated, can be rewritten as linear functions of *uncorrelated* standard normal deviates:

$$\begin{aligned} \tilde{\varepsilon}_{n21} &= c_{aa}\eta_1, \\ \tilde{\varepsilon}_{n31} &= c_{ab}\eta_1 + c_{bb}\eta_2, \end{aligned}$$

where η_1 and η_2 are iid $N(0, 1)$. The error differences $\tilde{\varepsilon}_{n21}$ and $\tilde{\varepsilon}_{n31}$ are correlated because both of them depend on η_1. With this way of expressing the error differences, the utility differences can be written

$$\tilde{U}_{n21} = \tilde{V}_{n21} + c_{aa}\eta_1,$$
$$\tilde{U}_{n31} = \tilde{V}_{n31} + c_{ab}\eta_1 + c_{bb}\eta_2.$$

The probability of alternative 1 is $P_{n1} = \text{Prob}(\tilde{U}_{n21} < 0 \text{ and } \tilde{U}_{n31} < 0) = \text{Prob}(\tilde{V}_{n21} + \tilde{\varepsilon}_{n21} < 0 \text{ and } \tilde{V}_{n31} + \tilde{\varepsilon}_{n31} < 0)$. This probability is hard to evaluate numerically in terms of the $\tilde{\varepsilon}$'s, because they are correlated. However, using the transformation based on the Choleski factor, the probability can be written in a way that involves independent random terms. The probability becomes a function of the one-dimensional standard cumulative normal distribution:

$$\begin{aligned}
P_{n1} &= \text{Prob}(\tilde{V}_{n21} + c_{aa}\eta_1 < 0 \text{ and } \tilde{V}_{n31} + c_{ab}\eta_1 + c_{bb}\eta_2 < 0) \\
&= \text{Prob}(\tilde{V}_{n21} + c_{aa}\eta_1 < 0) \\
&\quad \times \text{Prob}(\tilde{V}_{n31} + c_{ab}\eta_1 + c_{bb}\eta_2 < 0 \mid \tilde{V}_{n21} + c_{aa}\eta_1 < 0) \\
&= \text{Prob}(\eta_1 < -\tilde{V}_{n21}/c_{aa}) \\
&\quad \times \text{Prob}(\eta_2 < -(\tilde{V}_{n31} + c_{ab}\eta_1)/c_{bb} \mid \eta_1 < -\tilde{V}_{n21}/c_{aa}) \\
&= \Phi\left(\frac{-\tilde{V}_{n21}}{c_{aa}}\right) \times \int_{\eta_1=-\infty}^{-\tilde{V}_{n21}/c_{aa}} \Phi\left(\frac{-\tilde{V}_{n31} + c_{ab}\eta_1}{c_{bb}}\right)\phi(\eta_1)\, d\eta_1,
\end{aligned}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution evaluated at the point in the parentheses, and $\phi(\cdot)$ is the standard normal density. The first factor, $\Phi(-\tilde{V}_{n21}/c_{aa})$, is easy to calculate: it is simply the cumulative standard normal evaluated at $-\tilde{V}_{n21}/c_{aa}$. Computer packages contain fast routines for calculating the cumulative standard normal. The second factor is an integral. As we know, computers cannot integrate, and we use simulation to approximate integrals. This is the heart of the GHK procedure: using simulation to approximate the integral in P_{n1}.

Let us examine this integral more closely. It is the integral over a truncated normal, namely, over η_1 up to $-\tilde{V}_{n21}/c_{aa}$. The simulation proceeds as follows. Draw a value of η_1 from a standard normal density truncated above at $-\tilde{V}_{n21}/c_{aa}$. For this draw, calculate the factor $\Phi(-(\tilde{V}_{n31} + c_{ab}\eta_1)/c_{bb})$. Repeat this process for many draws, and average the results. This average is a simulated approximation to $\int_{\eta_1=-\infty}^{-\tilde{V}_{n21}/c_{aa}} \Phi(-(\tilde{V}_{n31} + c_{ab}\eta_1)/c_{bb})\phi(\eta_1)\, d\eta_1$. The simulated probability is then obtained by multiplying this average by the value of $\Phi(-\tilde{V}_{n21}/c_{aa})$, which is calculated exactly. Simple enough!

The question arises: how do we take a draw from a truncated normal? We describe how to take draws from truncated univariate distributions in Section 9.2.4. The reader may want to jump ahead and quickly view

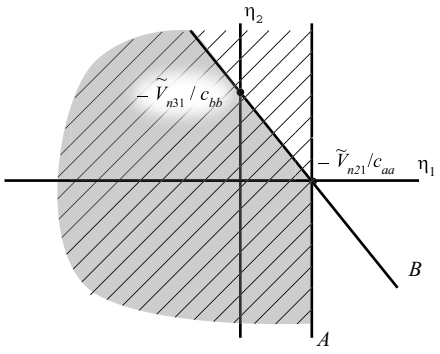

Figure 5.3. Probability of alternative 1.

that section. For truncated normals, the process is to take a draw from a standard uniform, labeled μ. Then calculate $\eta = \Phi^{-1}(\mu\Phi(-\tilde{V}_{n21}/c_{aa}))$. The resulting η is a draw from a normal density truncated from above at $-\tilde{V}_{n21}/c_{aa}$.

We can now put this all together to give the explicit steps that are used for the GHK simulator in our three-alternative case. The probability of alternative 1 is

$$P_{n1} = \Phi\left(\frac{-\tilde{V}_{n21}}{c_{aa}}\right) \times \int_{\eta_1=-\infty}^{-\tilde{V}_{n21}/c_{aa}} \Phi\left(\frac{-\tilde{V}_{n31} + c_{ab}\eta_1}{c_{bb}}\right)\phi(\eta_1)\,d\eta_1.$$

This probability is simulated as follows:

1. Calculate $k = \Phi(-\tilde{V}_{n21}/c_{aa})$.
2. Draw a value of η_1, labeled η_1^r, from a truncated standard normal truncated at $-\tilde{V}_{n21}/c_{aa}$. This is accomplished as follows:
   (a) Draw a standard uniform μ^r.
   (b) Calculate $\eta_1^r = \Phi^{-1}(\mu^r\Phi(-\tilde{V}_{n21}/c_{aa}))$.
3. Calculate $g^r = \Phi(-(\tilde{V}_{n31} + c_{ab}\eta_1^r)/c_{bb})$.
4. The simulated probability for this draw is $\check{P}_{n1}^r = k \times g^r$.
5. Repeat steps 1–4 R times, and average the results. This average is the simulated probability: $\check{P}_{n1} = (1/R)\sum \check{P}_{n1}^r$.

A graphical depiction is perhaps useful. Figure 5.3 shows the probability for alternative 1 in the space of independent errors η_1 and η_2. The x-axis is the value of η_1, and the y-axis is the value of η_2. The line labeled A is where η_1 is equal to $-\tilde{V}_{n21}/c_{aa}$. The condition that η_1 is below $-\tilde{V}_{n21}/c_{aa}$ is met in the striped area to the left of line A. The line labeled B is where $\eta_2 = -(\tilde{V}_{n31} + c_{ba}\eta_1)/c_{bb}$. Note that the y-intercept is where $\eta_1 = 0$, so that $\eta_2 = -\tilde{V}_{n31}/c_{bb}$ at this point. The slope of the line is $-c_{ba}/c_{bb}$. The condition that $\eta_2 < -(\tilde{V}_{n31} + c_{ba}\eta_1)/c_{bb}$ is satisfied below line B. The shaded area is where η_1 is to the left of line A and

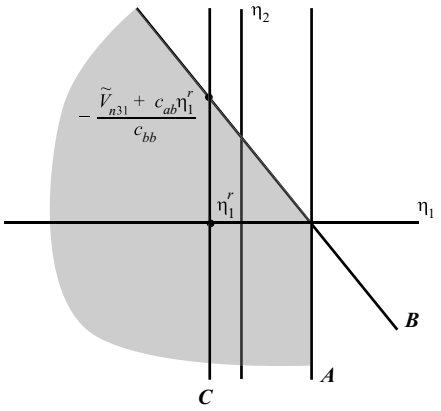

Figure 5.4. Probability that η_2 is in the correct range, given η_1^r.

η_2 is below line B. The mass of density in the shaded area is therefore the probability that alternative 1 is chosen.

The probability (i.e., the shaded mass) is the mass of the striped area times the proportion of this striped mass that is below line B. The striped area has mass $\Phi(-\tilde{V}_{n21}/c_{aa})$. This is easy to calculate. For any given value of η_1, the portion of the striped mass that is below line B is also easy to calculate. For example, in Figure 5.4, when η_1 takes the value η_1^r, then the probability that η_2 is below line B is the share of line C's mass that is below line B. This share is simply $\Phi(-(\tilde{V}_{n31} + c_{ab}\eta_1^r)/c_{bb})$. The portion of the striped mass that is below line B is therefore the average of $\Phi(-(\tilde{V}_{n31} + c_{ab}\eta_1^r)/c_{bb})$ over all values of η_1 that are to the left of line A. This average is simulated by taking draws of η_1 to the left of line A, calculating $\Phi(-(\tilde{V}_{n31} + c_{ab}\eta_1^r)/c_{bb})$ for each draw, and averaging the results. The probability is then this average times the mass of the striped area, $\Phi(-\tilde{V}_{n21}/c_{aa})$.

General Model

We can now describe the GHK simulator in general terms quickly, since the basic logic has already been discussed. This succinct expression serves to reinforce the concept that the GHK simulator is actually easier than it might at first appear.

Utility is expressed as

$$U_{nj} = V_{nj} + \varepsilon_{nj}, \qquad j = 1, \ldots, J,$$
$$\varepsilon_n' = \langle \varepsilon_{n1}, \ldots, \varepsilon_{nJ} \rangle, \qquad \varepsilon_n : J \times 1,$$
$$\varepsilon_n \sim N(0, \Omega).$$

Transform to utility differences against alternative i:

$$\begin{aligned}
&\tilde{U}_{nji} = \tilde{V}_{nji} + \tilde{\varepsilon}_{nji}, \qquad j \neq i,\\
&\tilde{\varepsilon}'_{ni} = \langle \tilde{\varepsilon}_{n1}, \ldots, \tilde{\varepsilon}_{nJ} \rangle, \qquad \text{where } \ldots \text{is over all except } i,\\
&\tilde{\varepsilon}_{ni} : (J-1) \times 1,\\
&\tilde{\varepsilon}_{ni} \sim N(0, \tilde{\Omega}_i),
\end{aligned}$$

where $\tilde{\Omega}_i$ is derived from Ω.

Reexpress the errors as a Choleski transformation of iid standard normal deviates.

$$\begin{aligned}
&L_i \qquad \text{s.t.} \quad L_i L'_i = \Omega_i,\\
&L_i = \begin{pmatrix}
c_{11} & 0 & \cdots & \cdots & \cdots & 0\\
c_{21} & c_{22} & 0 & \cdots & \cdots & 0\\
c_{31} & c_{32} & c_{33} & 0 & \cdots & 0\\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots
\end{pmatrix}.
\end{aligned}$$

Then, stacking utilities $\tilde{U}'_{ni} = (\tilde{U}_{n1i}, \ldots, \tilde{U}_{nJi})$, we get the vector form of the model,

$$\tilde{U}_{ni} = \tilde{V}_{ni} + L_i \eta_n,$$

where $\eta'_n = \langle \eta_{1n}, \ldots, \eta_{J-1,n} \rangle$ is a vector of iid standard normal deviates: $\eta_{nj} \sim N(0, 1)\ \forall j$. Written explicitly, the model is

$$\begin{aligned}
\tilde{U}_{n1i} &= \tilde{V}_{n1i} + c_{11}\eta_1,\\
\tilde{U}_{n2i} &= \tilde{V}_{n2i} + c_{21}\eta_1 + c_{22}\eta_2,\\
\tilde{U}_{n3i} &= \tilde{V}_{n3i} + c_{31}\eta_1 + c_{32}\eta_2 + c_{33}\eta_3,
\end{aligned}$$

and so on. The choice probabilities are

$$\begin{aligned}
P_{ni} &= \text{Prob}(\tilde{U}_{nji} < 0 \ \ \forall j \neq i)\\
&= \text{Prob}\left(\eta_1 < \frac{-\tilde{V}_{n1i}}{c_{11}}\right)\\
&\quad \times \text{Prob}\left(\eta_2 < \frac{-(\tilde{V}_{n2i} + c_{21}\eta_1)}{c_{22}} \middle| \eta_1 < \frac{-\tilde{V}_{n1i}}{c_{11}}\right)\\
&\quad \times \text{Prob}\left(\eta_3 < \frac{-(\tilde{V}_{n3i} + c_{31}\eta_1 + c_{32}\eta_2)}{c_{33}} \middle|\right.\\
&\qquad \left. \eta_1 < \frac{-\tilde{V}_{n1i}}{c_{11}} \ \text{ and } \ \eta_2 < \frac{-(\tilde{V}_{n2i} + c_{21}\eta_1)}{c_{22}}\right).\\
&\quad \times \cdots.
\end{aligned}$$

The GHK simulator is calculated as follows:

1. Calculate
$$\text{Prob}\left(\eta_1 < \frac{-\tilde{V}_{n1i}}{c_{11}}\right) = \Phi\left(\frac{-\tilde{V}_{n1i}}{c_{11}}\right).$$
2. Draw a value of η_1, labeled η_1^r, from a truncated standard normal truncated at $-\tilde{V}_{1in}/c_{11}$. This draw is obtained as follows:
   (a) Draw a standard uniform μ_1^r.
   (b) Calculate $\eta_1^r = \Phi^{-1}(\mu_1^r \Phi(-\tilde{V}_{n1i}/c_{11}))$.
3. Calculate
$$\text{Prob}\left(\eta_2 < \frac{-(\tilde{V}_{n2i} + c_{21}\eta_1)}{c_{22}}\middle|\eta_1 = \eta_1^r\right) = \Phi\left(\frac{-\left(\tilde{V}_{n2i} + c_{21}\eta_1^r\right)}{c_{22}}\right).$$
4. Draw a value of η_2, labeled η_2^r, from a truncated standard normal truncated at $-(\tilde{V}_{n2i} + c_{21}\eta_1^r)/c_{22}$. This draw is obtained as follows:
   (a) Draw a standard uniform μ_2^r.
   (b) Calculate $\eta_2^r = \Phi^{-1}(\mu_2^r \Phi(-(\tilde{V}_{n2i} + c_{21}\eta_1^r)/c_{22}))$.
5. Calculate
$$\text{Prob}\left(\eta_3 < \frac{-(\tilde{V}_{n3i} + c_{31}\eta_1 + c_{32}\eta_2)}{c_{33}}\middle|\eta_1 = \eta_1^r\ ,\ \eta_2 = \eta_2^r\right) = \Phi\left(\frac{-\left(\tilde{V}_{n3i} + c_{31}\eta_1^r + c_{32}\eta_2^r\right)}{c_{33}}\right).$$
6. And so on for all alternatives but i.
7. The simulated probability for this rth draw of $\eta_1, \eta_2, \ldots$ is calculated as
$$\begin{aligned}\breve{P}_{ni}^r = {} & \Phi\left(\frac{-\tilde{V}_{n1i}}{c_{11}}\right) \\ & \times \Phi\left(\frac{-\left(\tilde{V}_{n2i} + c_{21}\eta_1^r\right)}{c_{22}}\right) \\ & \times \Phi\left(\frac{-\left(\tilde{V}_{n3i} + c_{31}\eta_1^r + c_{32}\eta_2^r\right)}{c_{33}}\right) \\ & \times \cdots.\end{aligned}$$

8. Repeat steps 1–7 many times, for $r = 1, \ldots, R$.
9. The simulated probability is

$$\check{P}_{in} = \frac{1}{R} \sum_{r} \check{P}_{in}^{r}.$$

GHK Simulator with Maximum Likelihood Estimation

There are several issues that need to be addressed when using the GHK simulator in maximum likelihood estimation. First, in the log-likelihood function, we use the probability of the decision maker's chosen alternative. Since different decision makers choose different alternatives, P_{ni} must be calculated for different i's. The GHK simulator takes utility differences against the alternative for which the probability is being calculated, and so different utility differences must be taken for decision makers who chose different alternatives. Second, for a person who chose alternative i, the GHK simulator uses the covariance matrix $\tilde{\Omega}_i$, while for a person who chose alternative j, the matrix $\tilde{\Omega}_j$ is used. Both of these matrices are derived from the same covariance matrix Ω of the original errors. We must assure that the parameters in $\tilde{\Omega}_i$ are consistent with those in $\tilde{\Omega}_j$, in the sense that they both are derived from a common Ω. Third, we need to assure that the parameters that are estimated by maximum likelihood imply covariance matrices $\Omega_j \ \forall j$ that are positive definite, as a covariance matrix must be. Fourth, as always, we must make sure that the model is normalized for scale and level of utility, so that the parameters are identified.

Researchers use various procedures to address these issues. I will describe the procedure that I use.

To assure that the model is identified, I start with the covariance matrix of scaled utility differences with the differences taken against the first alternative. This is the matrix $\tilde{\Omega}_1$, which is $(J-1) \times (J-1)$. To assure that the covariance matrix is positive definite, I parameterize the model in terms of the Choleski factor of $\tilde{\Omega}_1$. That is, I start with a lower-triangular matrix that is $(J-1) \times (J-1)$ and whose top-left element is 1:

$$L_1 = \begin{pmatrix} 1 & 0 & \cdots & \cdots & \cdots & 0 \\ c_{21} & c_{22} & 0 & \cdots & \cdots & 0 \\ c_{31} & c_{32} & c_{33} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \end{pmatrix}.$$

The elements $c_{k\ell}$ of this Choleski factor are the parameters that are estimated in the model. Any matrix that is the product of a lower-triangular full-rank matrix multiplied by itself is positive definite. So by using the elements of L_1 as the parameters, I am assured that $\tilde{\Omega}_1$ is positive definite for any estimated values of these parameters.

The matrix Ω for the J nondifferenced errors is created from L_1. I create a $J \times J$ Choleski factor for Ω by adding a row of zeros at the top of L_1 and a column of zeros at the left. The resulting matrix is

$$L = \begin{pmatrix} 0 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & 1 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & c_{21} & c_{22} & 0 & \cdots & \cdots & 0 \\ 0 & c_{31} & c_{32} & c_{33} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots \end{pmatrix}.$$

Then Ω is calculated as LL'. With this Ω, I can derive $\tilde{\Omega}_j$ for any j. Note that Ω constructed in this way is fully general (i.e., allows any substitution pattern), since it utilizes all the parameters in the normalized $\tilde{\Omega}_1$.

Utility is expressed in vector form stacked by alternatives: $U_n = V_n + \varepsilon_n$, $\varepsilon_n \sim N(0, \Omega)$. Consider a person who has chosen alternative i. For the log-likelihood function, we want to calculate P_{ni}. Recall the matrix M_i that we introduced in Section 5.1. Utility differences are taken using this matrix: $\tilde{U}_{ni} = M_i U_n$, $\tilde{V}_{ni} = M_i V_n$, and $\tilde{\varepsilon}_{ni} = M_i \varepsilon_n$. The covariance of the error differences $\tilde{\varepsilon}_{ni}$ is calculated as $\tilde{\Omega}_i = M_i \Omega M_i'$. The Choleski factor of $\tilde{\Omega}_i$ is taken and labeled L_i. (Note that L_1 obtained here will necessarily be the same as the L_1 that we used at the beginning to parameterize the model.) The person's utility is expressed as: $\tilde{U}_{ni} = \tilde{V}_{ni} + L_i \eta_n$, where η_n is a $(J-1)$-vector of iid standard normal deviates. The GHK simulator is applied to this expression.

This procedure satisfies all of our requirements. The model is necessarily normalized for scale and level, since we parameterize it in terms of the Choleski factor L_1 of the covariance of *scaled* error *differences*, $\tilde{\Omega}_1$. Each $\tilde{\Omega}_i$ is consistent with each other $\tilde{\Omega}_j$ for $j \neq i$, because they are both derived from the same Ω (which is constructed from L_1). Each $\tilde{\Omega}_i$ is positive definite for any values of the parameters, because the parameters are the elements of L_1. As stated earlier, any matrix that is the product of a lower-triangular matrix multiplied by itself is positive definite, and so $\tilde{\Omega}_1 = LL'$ is positive definite. And each of the other $\tilde{\Omega}_j$'s, for $j = 2, \ldots, J$, is also positive definite, since they are constructed to be consistent with Ω_1, which is positive definite.

GHK as Importance Sampling

As I described in the three-alternative case, the GHK simulator provides a simulated approximation of the integral

$$\int_{\eta_1=-\infty}^{-\tilde{V}_{n21}/c_{aa}} \Phi\left(\frac{-\tilde{V}_{n31}+c_{ab}\eta_1}{c_{bb}}\right)\phi(\eta_1)\,d\eta_1.$$

The GHK simulator can be interpreted in another way that is often useful.

Importance sampling is a way of transforming an integral to be more convenient to simulate. The procedure is described in Section 9.2.7, and readers may find it useful to jump ahead to view that section. Importance sampling can be summarized as follows. Consider any integral $\bar{t} = \int t(\varepsilon)g(\varepsilon)\,d\varepsilon$ over a density g. Suppose that another density exists from which it is easy to draw. Label this other density $f(\varepsilon)$. The density g is called the target density, and f is the generating density. The integral can be rewritten as $\bar{t} = \int [t(\varepsilon)g(\varepsilon)/f(\varepsilon)]f(\varepsilon)\,d\varepsilon$. This integral is simulated by taking draws from f, calculating $t(\varepsilon)g(\varepsilon)/f(\varepsilon)$ for each draw, and averaging the results. This procedure is called importance sampling because each draw from f is weighted by g/f when taking the average of t; the weight g/f is the "importance" of the draw from f. This procedure is advantageous if (1) it is easier to draw from f than g, and/or (2) the simulator based on $t(\varepsilon)g(\varepsilon)/f(\varepsilon)$ has better properties (e.g., smoothness) than the simulator based on $t(\varepsilon)$.

The GHK simulator can be seen as making this type of transformation, and hence as being a type of importance sampling. Let η be a vector of $J-1$ iid standard normal deviates. The choice probability can be expressed as

$$P_{ni} = \int I(\eta \in B)g(\eta)\,d\eta, \tag{5.7}$$

where $B = \{\eta \text{ s.t. } \tilde{U}_{nji} < 0 \ \forall\ j \neq i\}$ is the set of η's that result in i being chosen; $g(\eta) = \phi(\eta_1)\cdots\phi(\eta_{J-1})$ is the density, where ϕ denotes the standard normal density; and the utilities are

$$\begin{aligned}
\tilde{U}_{n1i} &= \tilde{V}_{n1i} + c_{11}\eta_1,\\
\tilde{U}_{n2i} &= \tilde{V}_{n2i} + c_{21}\eta_1 + c_{22}\eta_2,\\
\tilde{U}_{n3i} &= \tilde{V}_{n3i} + c_{31}\eta_1 + c_{32}\eta_2 + c_{33}\eta_3,
\end{aligned}$$

and so on.

The direct way to simulate this probability is to take draws of η, calculate $I(\eta \in B)$ for each draw, and average the results. This is the AR simulator. This simulator has the unfortunate properties that it can be zero and is not smooth.

For GHK we draw η from a different density, not from $g(\eta)$. Recall that for GHK, we draw η_1 from a standard normal density truncated at $-\tilde{V}_{n1i}/c_{11}$. The density of this truncated normal is $\phi(\eta_1)/\Phi(-\tilde{V}_{n1i}/c_{11})$, that is, the standard normal density normalized by the total probability below the truncation point. Draws of η_2, η_3, and so on are also taken from truncated densities, but with different truncation points. Each of these truncated densities takes the form $\phi(\eta_j)/\Phi(\cdot)$ for some truncation point in the denominator. The density from which we draw for the GHK simulator is therefore

$$f(\eta) = \begin{cases} \frac{\phi(\eta_1)}{\Phi(-\tilde{V}_{n1i}/c_{11})} \times \frac{\phi(\eta_2)}{\Phi(-(\tilde{V}_{n2i}+c_{21}\eta_1)/c_{22})} \times \cdots & \text{for } \eta \in B, \\ 0 & \text{for } \eta \notin B. \end{cases} \tag{5.8}$$

Note that we only take draws that are consistent with the person choosing alternative i (since we draw from the correctly truncated distributions). So $f(\eta) = 0$ for $\eta \notin B$.

Recall that for a draw of η within the GHK simulator, we calculate:

$$\begin{aligned} \check{P}_{in}(\eta) = {} & \Phi\left(\frac{-\tilde{V}_{n1i}}{c_{11}}\right) \\ & \times \Phi\left(\frac{-(\tilde{V}_{n2i} + c_{21}\eta_1)}{c_{22}}\right) \\ & \times \cdots. \end{aligned} \tag{5.9}$$

Note that this expression is the denominator of $f(\eta)$ for $\eta \in B$, given in equation (5.8). Using this fact, we can rewrite the density $f(\eta)$ as

$$f(\eta) = \begin{cases} g(\eta)/\check{P}_{ni}(\eta) & \text{for } \eta \in B, \\ 0 & \text{for } \eta \notin B. \end{cases}$$

With this expression for $f(\eta)$, we can prove that the GHK simulator, $\check{P}_{in}(\eta)$, is unbiased for $P_{ni}(\eta)$:

$$\begin{aligned} E(\check{P}_{in}(\eta)) &= \int \check{P}_{in}(\eta) f(\eta)\, d\eta \\ &= \int_{\eta \in B} \check{P}_{in}(\eta) \frac{g(\eta)}{\check{P}_{in}(\eta)}\, d\eta \qquad \text{by (5.6.3)} \\ &= \int_{\eta \in B} g(\eta)\, d\eta \\ &= \int I(\eta \in B) g(\eta)\, d\eta \\ &= P_{in}. \end{aligned}$$

The interpretation of GHK as an importance sampler is also obtained from this expression:

$$\begin{aligned} P_{in} &= \int I(\eta \in B) g(\eta)\, d\eta \\ &= \int I(\eta \in B) g(\eta) \frac{f(\eta)}{f(\eta)} d\,\eta \\ &= \int I(\eta \in B) \frac{g(\eta)}{g(\eta)/\check{P}_{in}(\eta)} f(\eta)\, d\eta \qquad \text{by (5.6.3)} \\ &= \int I(\eta \in B) \check{P}_{in}(\eta) f(\eta)\, d\eta \\ &= \int \check{P}_{in}(\eta) f(\eta)\, d\eta, \end{aligned}$$

where the last equality is because $f(\eta) > 0$ only when $\eta \in B$. The GHK procedure takes draws from $f(\eta)$, calculates $\check{P}_{in}(\eta)$ for each draw, and averages the results. Essentially, GHK replaces the 0–1 $I(\eta \in B)$ with smooth $\check{P}_{in}(\eta)$ and makes the corresponding change in the density from $g(\eta)$ to $f(\eta)$.