6 Mixed Logit

6.1 Choice Probabilities

Mixed logit is a highly flexible model that can approximate any random utility model (McFadden and Train, 2000). It obviates the three limitations of standard logit by allowing for random taste variation, unrestricted substitution patterns, and correlation in unobserved factors over time. Unlike probit, it is not restricted to normal distributions. Its derivation is straightforward, and simulation of its choice probabilities is computationally simple.

Like probit, the mixed logit model has been known for many years but has only become fully applicable since the advent of simulation. The first application of mixed logit was apparently the automobile demand models created jointly by Boyd and Mellman (1980) and Cardell and Dunbar (1980). In these studies, the explanatory variables did not vary over decision makers, and the observed dependent variable was market shares rather than individual customers' choices. As a result, the computationally intensive integration that is inherent in mixed logit (as explained later) needed to be performed only once for the market as a whole, rather than for each decision maker in a sample. Early applications on customer-level data, such as Train *et al*. (1987a) and Ben-Akiva *et al*. (1993), included only one or two dimensions of integration, which could be calculated by quadrature. Improvements in computer speed and in our understanding of simulation methods have allowed the full power of mixed logits to be utilized. Among the studies to evidence this power are those by Bhat (1998a) and Brownstone and Train (1999) on cross-sectional data, and Erdem (1996), Revelt and Train (1998), and Bhat (2000) on panel data. The description in the current chapter draws heavily from Train (1999).

Mixed logit models can be derived under a variety of different behavioral specifications, and each derivation provides a particular interpretation. The mixed logit model is *defined* on the basis of the functional form for its choice probabilities. Any behavioral specification whose

derived choice probabilities take this particular form is called a mixed logit model.

Mixed logit probabilities are the integrals of standard logit probabilities over a density of parameters. Stated more explicitly, a mixed logit model is any model whose choice probabilities can be expressed in the form

$$P_{ni} = \int L_{ni}(\beta) f(\beta)\, d\beta,$$

where $L_{ni}(\beta)$ is the logit probability evaluated at parameters β:

$$L_{ni}(\beta) = \frac{e^{V_{ni}(\beta)}}{\sum_{j=1}^{J} e^{V_{nj}(\beta)}}$$

and $f(\beta)$ is a density function. $V_{ni}(\beta)$ is the observed portion of the utility, which depends on the parameters β. If utility is linear in β, then $V_{ni}(\beta) = \beta' x_{ni}$. In this case, the mixed logit probability takes its usual form:

$$P_{ni} = \int \left(\frac{e^{\beta' x_{ni}}}{\sum_j e^{\beta' x_{nj}}} \right) f(\beta)\, d\beta. \tag{6.1}$$

The mixed logit probability is a weighted average of the logit formula evaluated at different values of β, with the weights given by the density $f(\beta)$. In the statistics literature, the weighted average of several functions is called a mixed function, and the density that provides the weights is called the mixing distribution. Mixed logit is a mixture of the logit function evaluated at different β's with $f(\beta)$ as the mixing distribution.

Standard logit is a special case where the mixing distribution $f(\beta)$ is degenerate at fixed parameters b: $f(\beta) = 1$ for $\beta = b$ and 0 for $\beta \neq b$. The choice probability (6.1) then becomes the simple logit formula

$$P_{ni} = \frac{e^{b' x_{ni}}}{\sum_j e^{b' x_{nj}}}.$$

The mixing distribution $f(\beta)$ can be discrete, with β taking a finite set of distinct values. Suppose β takes M possible values labeled $b_1, \ldots, b_M$, with probability s_m that $\beta = b_m$. In this case, the mixed logit becomes the *latent class model* that has long been popular in psychology and marketing; examples include Kamakura and Russell (1989) and Chintagunta *et al.* (1991). The choice probability is

$$P_{ni} = \sum_{m=1}^{M} s_m \left(\frac{e^{b_m' x_{ni}}}{\sum_j e^{b_m' x_{nj}}} \right).$$

This specification is useful if there are M segments in the population, each of which has its own choice behavior or preferences. The share of the population in segment m is s_m, which the researcher can estimate within the model along with the b's for each segment.

In most applications that have actually been called mixed logit (such as those cited in the introductory paragraphs in this chapter), $f(\beta)$ is specified to be continuous. For example, the density of β can be specified to be normal with mean b and covariance W. The choice probability under this density becomes

$$P_{ni} = \int \left(\frac{e^{\beta' x_{ni}}}{\sum_j e^{\beta' x_{nj}}} \right) \phi(\beta \mid b, W)\, d\beta,$$

where $\phi(\beta \mid b, W)$ is the normal density with mean b and covariance W. The researcher estimates b and W. The lognormal, uniform, triangular, gamma, or any other distribution can be used. As will be shown in Section 6.5, by specifying the explanatory variables and density appropriately, the researcher can represent any utility-maximizing behavior by a mixed logit model, as well as many forms of non-utility-maximizing behavior.

Tests for the need for a nondegenerate mixing distribution, as well as the adequacy of any given distribution, have been developed by McFadden and Train (2000) and Chesher and Santos-Silva (2002). Several studies have compared discrete and continuous mixing distributions within the context of mixed logit; see, for example, Wedel and Kamakura (2000) and Ainslie *et al.* (2001).

An issue of terminology arises with mixed logit models. There are two sets of parameters in a mixed logit model. First, we have the parameters β, which enter the logit formula. These parameters have density $f(\beta)$. The second set are parameters that describe this density. For example, if β is normally distributed with mean b and covariance W, then b and W are parameters that describe the density $f(\beta)$. Usually (though not always, as noted in the following text), the researcher is interested in estimating the parameters of f.

Denote the parameters that describe the density of β as θ. The more appropriate way to denote this density is $f(\beta \mid \theta)$. The mixed logit choice probabilities do not depend on the values of β. These probabilities are $P_{ni} = \int L_{ni}(\beta) f(\beta \mid \theta)\, d\beta$, which are functions of θ. The parameters β are integrated out. Thus, the β's are similar to the ε_{nj}'s, in that both are random terms that are integrated out to obtain the choice probability.

Under some derivations of the mixed logit model, the values of β have interpretable meaning as representing the tastes of individual decision makers. In these cases, the researcher might want to obtain information about the β's for each sampled decision maker, as well as the θ that describes the distribution of β's across decision makers. In Chapter 11, we describe how the researcher can obtain this information from estimates of θ and the observed choices of each decision maker. In the current chapter, we describe the estimation and interpretation of θ, using classical estimation procedures. In Chapter 12, we describe Bayesian procedures that provide information about θ and each decision maker's β simultaneously.

6.2 Random Coefficients

The mixed logit probability can be derived from utility-maximizing behavior in several ways that are formally equivalent but provide different interpretations. The most straightforward derivation, and most widely used in recent applications, is based on random coefficients. The decision maker faces a choice among J alternatives. The utility of person n from alternative j is specified as

$$U_{nj} = \beta_n' x_{nj} + \varepsilon_{nj},$$

where x_{nj} are observed variables that relate to the alternative and decision maker, β_n is a vector of coefficients of these variables for person n representing that person's tastes, and ε_{nj} is a random term that is iid extreme value. The coefficients vary over decision makers in the population with density $f(\beta)$. This density is a function of parameters θ that represent, for example, the mean and covariance of the β's in the population. This specification is the same as for standard logit except that β varies over decision makers rather than being fixed.

The decision maker knows the value of his own β_n and ε_{nj}'s for all j and chooses alternative i if and only if $U_{ni} > U_{nj}\ \forall j \neq i$. The researcher observes the x_{nj}'s but not β_n or the ε_{nj}'s. If the researcher observed β_n, then the choice probability would be standard logit, since the ε_{nj}'s are iid extreme value. That is, the probability *conditional* on β_n is

$$L_{ni}(\beta_n) = \frac{e^{\beta_n' x_{ni}}}{\sum_j e^{\beta_n' x_{nj}}}.$$

However, the researcher does not know β_n and therefore cannot condition on β. The unconditional choice probability is therefore the integral of

$L_{ni}(\beta_n)$ over all possible variables of β_n:

$$P_{ni} = \int \left(\frac{e^{\beta' x_{ni}}}{\sum_j e^{\beta' x_{nj}}} \right) f(\beta)\, d\beta,$$

which is the mixed logit probability (6.1).

The researcher specifies a distribution for the coefficients and estimates the parameters of that distribution. In most applications, such as Revelt and Train (1998), Mehndiratta (1996), and Ben-Akiva and Bolduc (1996), $f(\beta)$ has been specified to be normal or lognormal: $\beta \sim N(b, W)$ or $\ln \beta \sim N(b, W)$ with parameters b and W that are estimated. The lognormal distribution is useful when the coefficient is known to have the same sign for every decision maker, such as a price coefficient that is known to be negative for everyone. Revelt and Train (2000), Hensher and Greene (2001), and Train (2001) have used triangular and uniform distributions. With the uniform density, β is distributed uniformly between $b - s$ and $b + s$, where the mean b and spread s are estimated. The triangular distribution has positive density that starts at $b - s$, rises linearly to b, and then drops linearly to $b + s$, taking the form of a tent or triangle. The mean b and spread s are estimated, as with the uniform, but the density is peaked instead of flat. These densities have the advantage of being bounded on both sides, thereby avoiding the problem that can arise with normals and lognormals having unreasonably large coefficients for some share of decision makers. By constraining $s = b$, the researcher can assure that the coefficients have the same sign for all decision makers. Siikamaki (2001) and Siikamaki and Layton (2001) use the Rayleigh distribution (Johnson *et al.*, 1994), which is on one side of zero like the lognormal but, as these researchers found, can be easier for estimation than the lognormal. Revelt (1999) used truncated normals. As these examples indicate, the researcher is free to specify a distribution that satisfies his expectations about behavior in his own application.

Variations in tastes that are related to observed attributes of the decision maker are captured through specification of the explanatory variables and/or the mixing distribution. For example, cost might be divided by the decision maker's income to allow the value or relative importance of cost to decline as income rises. The random coefficient of this variable then represents the variation over people with the same income in the value that they place on cost. The mean valuation of cost declines with increasing income while the variance around the mean is fixed. Observed attributes of the decision maker can also enter $f(\beta)$, so that higher-order moments of taste variation can also depend on attributes

of the decision maker. For example, Bhat (1998a, 2000) specify $f(\beta)$ to be lognormal with mean and variance depending on decision maker characteristics.

6.3 Error Components

A mixed logit model can be used without a random-coefficients interpretation, as simply representing error components that create correlations among the utilities for different alternatives. Utility is specified as

$$U_{nj} = \alpha' x_{nj} + \mu_n' z_{nj} + \varepsilon_{nj},$$

where x_{nj} and z_{nj} are vectors of observed variables relating to alternative j, α is a vector of fixed coefficients, μ is a vector of random terms with zero mean, and ε_{nj} is iid extreme value. The terms in z_{nj} are error components that, along with ε_{nj}, define the stochastic portion of utility. That is, the unobserved (random) portion of utility is $\eta_{nj} = \mu_n' z_{nj} + \varepsilon_{nj}$, which can be correlated over alternatives depending on the specification of z_{nj}. For the standard logit model, z_{nj} is identically zero, so that there is no correlation in utility over alternatives. This lack of correlation gives rise to the IIA property and its restrictive substitution patterns. With nonzero error components, utility is correlated over alternatives: $\text{Cov}(\eta_{ni}, \eta_{nj}) = E(\mu_n' z_{ni} + \varepsilon_{ni})(\mu_n' z_{nj} + \varepsilon_{nj}) = z_{ni}' W z_{nj}$, where W is the covariance of μ_n. Utility is correlated over alternatives even when, as in most specifications, the error components are independent, such that W is diagonal.

Various correlation patterns, and hence substitution patterns, can be obtained by appropriate choice of variables to enter as error components. For example, an analog to nested logit is obtained by specifying a dummy variable for each nest that equals 1 for each alternative in the nest and zero for alternatives outside the nest. With K non-overlapping nests, the error components are $\mu_n' z_{nj} = \sum_{k=1}^{K} \mu_{nk} d_{jk}$, where $d_{jk} = 1$ if j is in nest k and zero otherwise. It is convenient in this situation to specify the error components to be independently normally distributed: μ_{nk} iid $N(0, \sigma_k)$. The random quantity μ_{nk} enters the utility of each alternative in nest k, inducing correlation among these alternatives. It does not enter any of the alternatives in other nests, thereby not inducing correlation between alternatives in the nest with those outside the nest. The variance σ_k captures the magnitude of the correlation. It plays an analogous role to the inclusive value coefficient of nested logit models.

To be more precise, the covariance between two alternatives in nest k is $\text{Cov}(\eta_{ni}, \eta_{nj}) = E(\mu_k + \varepsilon_{ni})(\mu_k + \varepsilon_{nj}) = \sigma_k$. The variance for each of the alternatives in nest k is $\text{Var}(\eta_{ni}) = E(\mu_k + \varepsilon_{ni})^2 = \sigma_k + \pi^2/6$, since

the variance of the extreme value term, ε_{ni}, is $\pi^2/6$ (see Section 3.1). The correlation between any two alternatives within nest k is therefore $\sigma_k/(\sigma_k + \pi^2/6)$. Constraining the variance of each nest's error component to be the same for all nests (i.e., constraining $\sigma_k = \sigma,\ k = 1, \ldots, K$) is analogous to constraining the log-sum coefficient to be the same for all nests in a nested logit. This constraint also assures that the mixed logit model is normalized for scale and level.

Allowing different variances for the random quantities for different nests is analogous to allowing the inclusive value coefficient to differ across nests in a nested logit. An analog to overlapping nests is captured with dummies that identify overlapping sets of alternatives, as in Bhat (1998a). An analog to heteroskedastic logit (discussed in Section 4.5) is obtained by entering an error component for each alternative. Ben-Akiva *et al.* (2001) provide guidance on how to specify these variables appropriately.

Error-component and random-coefficient specifications are formally equivalent. Under the random-coefficient motivation, utility is specified as $U_{nj} = \beta_n' x_{nj} + \varepsilon_{nj}$ with random β_n. The coefficients β_n can be decomposed into their mean α and deviations μ_n, so that $U_{nj} = \alpha' x_{nj} + \mu_n' x_{nj} + \varepsilon_{nj}$, which has error components defined by $z_{nj} = x_{nj}$. Conversely, under an error-component motivation, utility is $U_{nj} = \alpha' x_{nj} + \mu_n' z_{nj} + \varepsilon_{nj}$, which is equivalent to a random-parameter model with fixed coefficients for variables x_{nj} and random coefficients with zero means for variables z_{nj}. If x_{nj} and z_{nj} overlap (in the sense that some of the same variables enter x_{nj} and z_{nj}), the coefficients of these variables can be considered to vary randomly with mean α and the same distribution as μ_n around their means.

Though random coefficients and error components are formally equivalent, the way a researcher thinks about the model affects the specification of the mixed logit. For example, when thinking in terms of random parameters, it is natural to allow each variable's coefficient to vary and perhaps even to allow correlations among the coefficients. This is the approach pursued by Revelt and Train (1998). However, when the primary goal is to represent substitution patterns appropriately through the use of error components, the emphasis is placed on specifying variables that can induce correlations over alternatives in a parsimonious fashion so as to provide sufficiently realistic substitution patterns. This is the approach taken by Brownstone and Train (1999). The goals differed in these studies, Revelt and Train being interested in the pattern of tastes, while Brownstone and Train were more concerned with prediction. The number of explanatory variables also differed, Revelt and Train examining 6 variables, so that estimating the joint distribution of their coefficients was a reasonable goal, while Brownstone and Train included

26 variables. Expecting to estimate the distribution of 26 coefficients is unreasonable, and yet thinking in terms of random parameters instead of error components can lead the researcher to such expectations. It is important to remember that the mixing distribution, whether motivated by random parameters or by error components, captures variance and correlations in unobserved factors. There is a natural limit on how much one can learn about things that are not seen.

6.4 Substitution Patterns

Mixed logit does not exhibit independence from irrelevant alternatives (IIA) or the restrictive substitution patterns of logit. The ratio of mixed logit probabilities, P_{ni}/P_{nj}, depends on all the data, including attributes of alternatives other than i or j. The denominators of the logit formula are inside the integrals and therefore do not cancel. The percentage change in the probability for one alternative given a change in the mth attribute of another alternative is

$$\begin{aligned} E_{ni x_{nj}^m} &= -\frac{1}{P_{ni}} \int \beta^m L_{ni}(\beta) L_{nj}(\beta) f(\beta)\, d\beta \\ &= -\int \beta^m L_{nj}(\beta) \left[\frac{L_{ni}(\beta)}{P_{ni}} \right] f(\beta)\, d\beta, \end{aligned}$$

where β^m is the mth element of β. This elasticity is different for each alternative i. A ten-percent reduction for one alternative need not imply (as with logit) a ten-percent reduction in each other alternative. Rather, the substitution pattern depends on the specification of the variables and mixing distribution, which can be determined empirically.

Note that the percentage change in probability depends on the correlation between $L_{ni}(\beta)$ and $L_{nj}(\beta)$ over different values of β, which is determined by the researcher's specification of variables and mixing distribution. For example, to represent a situation where an improvement in alternative j draws proportionally more from alternative i than from alternative k, the researcher can specify an element of x that is positively correlated between i and j but uncorrelated or negatively correlated between k and j, with a mixing distribution that allows the coefficient of this variable to vary.

6.5 Approximation to Any Random Utility Model

McFadden and Train (2000) show that any random utility model (RUM) can be approximated to any degree of accuracy by a mixed logit with appropriate choice of variables and mixing distribution. This proof is analogous to the RUM-consistent approximations provided by Dagsvik

(1994). An intuitive explanation can easily be provided. Suppose the true model is $U_{nj} = \alpha_n' z_{nj}$, where z_{nj} are variables related to alternative j, and α follows any distribution $f(\alpha)$. Any RUM can be expressed in this form. (The more traditional notation $U_{nj} = \beta_n' x_{nj} + \varepsilon_{nj}$ is obtained by letting $z_{nj}' = \langle x_{nj}', d_j \rangle$, $\alpha' = \langle \beta_n', \varepsilon_{nj} \rangle$, and $f(\alpha)$ be the joint density of β_n and ε_{nj} $\forall j$.) Conditional on α, the person's choice is fully determined, since U_{nj} is then known for each j. The conditional probability is therefore

$$q_{ni}(\alpha) = I(\alpha_n' z_{ni} > \alpha_n' z_{nj} \ \forall j \neq i),$$

where $I(\cdot)$ is the 1–0 indicator of whether the event in parentheses occurs. This conditional probability is deterministic in the sense that the probability is either zero or one: conditional on all the unknown random terms, the decision maker's choice is completely determined. The unconditional choice probability is the integral of $q_{ni}(\alpha)$ over α:

$$Q_{ni} = \int I(\alpha_n' z_{ni} > \alpha_n' z_{ni} \ \forall j \neq i) f(\alpha)\, d\alpha.$$

We can approximate this probability with a mixed logit. Scale utility by λ, so that $U_{nj}^* = (\alpha/\lambda) z_{nj}$. This scaling does not change the model, since behavior is unaffected by the scale of utility. Then add an iid extreme value term: ε_{nj}. The addition of the extreme value term does change the model, since it changes the utility of each alternative. We add it because doing so gives us a mixed logit. And, as we will show (this is the purpose of the proof), adding the extreme value term is innocuous. The mixed logit probability based on this utility is

$$P_{ni} = \int \left(\frac{e^{(\alpha/\lambda)' z_{ni}}}{\sum_j e^{(\alpha/\lambda)' z_{nj}}} \right) f(\alpha)\, d\alpha.$$

As λ approaches zero, the coefficients α/λ in the logit formula grow large, and P_{ni} approaches a 1–0 indicator for the alternative with the highest utility. That is, the mixed logit probability P_{ni} approaches the true probability Q_{ni} as λ approaches zero. By scaling the coefficients upward sufficiently, the mixed logit based on these scaled coefficients is arbitrarily close to the true model. Srinivasan and Mahmassani (2000) use this concept of raising the scale of coefficients to show that a mixed logit can approximate a probit model; the concept applies generally to approximate any RUM.

Recall that we added an iid extreme value term to the true utility of each alternative. These terms change the model, because the alternative with highest utility before the terms are added may not have highest utility

afterward (since a different amount is added to each utility). However, by raising the scale of utility sufficiently, we can be essentially sure that the addition of the extreme value terms has no effect. Consider a two-alternative example. Suppose, using the true model with its original scaling, that the utility of alternative 1 is 0.5 units higher than the utility of alternative 2, so that alternative 1 is chosen. Suppose we add an extreme value term to each alternative. There's a good chance, given the variance of these random terms, that the value obtained for alternative 2 will exceed that for alternative 1 by at least half a unit, so that alternative 2 now obtains the higher utility instead of 1. The addition of the extreme value terms thus changes the model, since it changes which alternative has the higher utility. Suppose, however, that we scale up the original utility by a factor of 10 (i.e., $\lambda = 0.10$). The utility for alternative 1 now exceeds the utility for alternative 2 by 5 units. It is highly unlikely that adding extreme value terms to these utilities will reverse this difference. That is, it is highly unlikely, in fact next to impossible, that the value of ε_{n2} that is added to the utility of alternative 2 is larger by 5 than the ε_{n1} that is added to the utility of alternative 1. If scaling up by 10 is not sufficient to assure that adding the extreme value term has no effect, then the original utilities can be scaled up by 100 or 1000. At some point, a scale will be found for which the addition of the extreme value terms has no effect. Stated succinctly, adding an extreme value term to true utility, which makes the model into a mixed logit, does not change utility in any meaningful way when the scale of the utility is sufficiently large. A mixed logit can approximate any RUM simply by scaling up utility sufficiently.

This demonstration is not intended to suggest that raising the scale of utility is actually how the researcher would proceed in specifying a mixed logit as an approximation to the true model. Rather, the demonstration simply indicates that if no other means for specifying a mixed logit to approximate the true model can be found, then this rescaling procedure can be used to attain the approximation. Usually, a mixed logit can be specified that adequately reflects the true model without needing to resort to an upward scaling of utility. For example, the true model will usually contain some iid term that is added to the utility of each alternative. Assuming an extreme value distribution for this term is perhaps close enough to reality to be empirically indistinguishable from other distributional assumptions for the iid term. In this case, the scale of utility is determined naturally by the variance of this iid term. The researcher's task is simply to find variables and a mixing distribution that capture the other parts of utility, namely, the parts that are correlated over alternatives or heteroskedastic.

6.6 Simulation

Mixed logit is well suited to simulation methods for estimation. Utility is $U_{nj} = \beta_n' x_{nj} + \varepsilon_{nj}$, where the the coefficients β_n are distributed with density $f(\beta \mid \theta)$, where θ refers collectively to the parameters of this distribution (such as the mean and covariance of β). The researcher specifies the functional form $f(\cdot)$ and wants to estimate the parameters θ. The choice probabilities are

$$P_{ni} = \int L_{ni}(\beta) f(\beta \mid \theta)\, d\beta,$$

where

$$L_{ni}(\beta) = \frac{e^{\beta' x_{ni}}}{\sum_{j=1}^{J} e^{\beta' x_{nj}}}.$$

The probabilities are approximated through simulation for any given value of θ: (1) Draw a value of β from $f(\beta \mid \theta)$, and label it β^r with the superscript $r = 1$ referring to the first draw. (2) Calculate the logit formula $L_{ni}(\beta^r)$ with this draw. (3) Repeat steps 1 and 2 many times, and average the results. This average is the simulated probability:

$$\check{P}_{ni} = \frac{1}{R} \sum_{r=1}^{R} L_{ni}(\beta^r),$$

where R is the number of draws. $\check{P}_{ni}$ is an unbiased estimator of P_{ni} by construction. Its variance decreases as R increases. It is strictly positive, so that $\ln \check{P}_{ni}$ is defined, which is useful for approximating the log-likelihood function below. $\check{P}_{ni}$ is smooth (twice differentiable) in the parameters θ and the variables x, which facilitates the numerical search for the maximum likelihood function and the calculation of elasticities. And $\check{P}_{ni}$ sums to one over alternatives, which is useful in forecasting.

The simulated probabilities are inserted into the log-likelihood function to give a simulated log likelihood:

$$\text{SLL} = \sum_{n=1}^{N} \sum_{j=1}^{J} d_{nj} \ln \check{P}_{nj},$$

where $d_{nj} = 1$ if n chose j and zero otherwise. The maximum simulated likelihood estimator (MSLE) is the value of θ that maximizes SLL. The properties of this estimator are discussed in Chapter 10. Usually, different draws are taken for each observation. This procedure maintains independence over decision makers of the simulated probabilities that

enter SLL. Lee (1992) describes the properties of MSLE when the same draws are used for all observations.

The simulated mixed logit probability can be related to accept–reject (AR) methods of simulation. AR simulation is described in Section 5.6 for probit models, but it is applicable more generally. For any random utility model, the AR simulator is constructed as follows: (1) A draw of the random terms is taken. (2) The utility of each alternative is calculated from this draw, and the alternative with the highest utility is identified. (3) Steps 1 and 2 are repeated many times. (4) The simulated probability for an alternative is calculated as the proportion of draws for which that alternative has the highest utility. The AR simulator is unbiased by construction. However, it is not strictly positive for any finite number of draws. It is also not smooth, but rather a step function: constant within ranges of parameters for which the identity of the alternative with the highest utility does not change for any draws, and with jumps where changes in the parameters change the identity of the alternative with the highest utility. Numerical methods for maximization based on the AR simulator are hampered by these characteristics. To address these numerical problems, the AR simulator can be smoothed by replacing the 0–1 indicator with the logit formula. As discussed in Section 5.6.2, the logit-smoothed AR simulator can approximate the AR simulator arbitrarily closely by scaling utility appropriately.

The mixed logit simulator can be seen as a logit-smoothed AR simulator of any RUM: draws of the random terms are taken, utility is calculated for these draws, the calculated utilities are inserted into the logit formula, and the results are averaged. The theorem that a mixed logit can approximate any random utility model (Section 6.5) can be viewed from this perspective. We know from Section 5.6.2 that the logit-smoothed AR simulator can be arbitrarily close to the AR simulator for any model, with sufficient scaling of utility. Since the mixed logit simulator is equivalent to a logit-smoothed AR simulator, the simulated mixed logit model can be arbitrarily close to the AR simulator of any model.

6.7 Panel Data

The specification is easily generalized to allow for repeated choices by each sampled decision maker. The simplest specification treats the coefficients that enter utility as varying over people but being constant over choice situations for each person. Utility from alternative j in choice situation t by person n is $U_{njt} = \beta_n x_{njt} + \varepsilon_{njt}$, with ε_{njt} being iid extreme value over time, people, and alternatives. Consider a sequence of

alternatives, one for each time period, $\mathbf{i} = \{i_1, \ldots, i_T\}$. Conditional on β the probability that the person makes this sequence of choices is the product of logit formulas:

$$\mathbf{L}_{n\mathbf{i}}(\beta) = \prod_{t=1}^{T} \left[\frac{e^{\beta_n' x_{ni_t t}}}{\sum_j e^{\beta_n' x_{njt}}} \right] \tag{6.2}$$

since the ε_{njt}'s are independent over time. The unconditional probability is the integral of this product over all values of β:

$$P_{n\mathbf{i}} = \int \mathbf{L}_{n\mathbf{i}}(\beta) f(\beta)\, d\beta. \tag{6.3}$$

The only difference between a mixed logit with repeated choices and one with only one choice per decision maker is that the integrand involves a *product* of logit formulas, one for each time period, rather than just one logit formula. The probability is simulated similarly to the probability with one choice period. A draw of β is taken from its distribution. The logit formula is calculated for each period, and the product of these logits is taken. This process is repeated for many draws, and the results are averaged.

Past and future exogenous variables can be added to the utility in a given period to represent lagged response and anticipatory behavior, as described in Section 5.5 in relation to probit with panel data. However, unlike probit, lagged dependent variables can be added in a mixed logit model without changing the estimation procedure. Conditional on β_n, the only remaining random terms in the mixed logit are the ε_{nj}'s, which are independent over time. A lagged dependent variable entering U_{njt} is uncorrelated with these remaining error terms for period t, since these terms are independent over time. The conditional probabilities (conditional on β) are therefore the same as in equation (6.2), but with the x's including lagged dependent variables. The unconditional probability is then the integral of this conditional probability over all values of β, which is just equation (6.3). In this regard, mixed logit is more convenient than probit for representing state dependence, since lagged dependent variables can be added to mixed logit without adjusting the probability formula or simulation method. Erdem (1996) and Johannesson and Lundin (2000) exploit this advantage to examine habit formation and variety seeking within a mixed logit that also captures random taste variation.

If choices and data are not observed from the start of the process (i.e., from the first choice situation that the person faces), the issue of initial conditions must be confronted, just as with probit. The researcher must

somehow represent the probability of the first observed choice, which depends on the previous, unobserved choices. Heckman and Singer (1986) provide ways to handle this issue. However, when the researcher observes the choice process from the beginning, the initial conditions issue does not arise. In this case, the use of lagged dependent variables to capture inertia or other types of state dependence is straightforward with mixed logit. Stated-preference data (that is, answers to a series of choice situations posed to respondents in a survey) provide a prominent example of the researcher observing the entire sequence of choices.

In the specification so far and in nearly all applications, the coefficients β_n are assumed to be constant over choice situations for a given decision maker. This assumption is appropriate if the decision maker's tastes are stable over the time period that spans the repeated choices. However, the coefficients associated with each person can be specified to vary over time in a variety of ways. For example, each person's tastes might be serially correlated over choice situations, so that utility is

$$\begin{aligned} U_{njt} &= \beta_{nt} x_{njt} + \varepsilon_{njt}, \\ \beta_{nt} &= b + \tilde{\beta}_{nt}, \\ \tilde{\beta}_{nt} &= \rho \tilde{\beta}_{nt-1} + \mu_{nt}, \end{aligned}$$

where b is fixed and μ_{nt} is iid over n and t. Simulation of the probability for the sequence of choices proceeds as follows:

1. Draw μ^r_{n1} for the initial period, and calculate the logit formula for this period using $\beta^r_{n1} = b + \mu^r_{n0}$.
2. Draw μ^r_{n2} for the second period, calculate $\beta_{n2} = b + \rho\mu^r_{n1} + \mu^r_{n2}$, and then calculate the logit formula based on this β^r_{n2}.
3. Continue for all T time periods.
4. Take the product of the T logits.
5. Repeat steps 1–4 for numerous sequences of draws.
6. Average the results.

The burden placed on simulation is greater than with coefficients being constant over time for each person, requiring T times as many draws.

6.8 Case Study

As illustration, consider a mixed logit of anglers' choices of fishing sites (Train, 1999). The specification takes a random-coefficients form. Utility is $U_{njt} = \beta_n x_{njt} + \varepsilon_{njt}$, with coefficients β_n varying over anglers but not over trips for each angler. The probability of the sequence of sites chosen by each angler is given by equation (6.3).

The sample consists of 962 river trips taken in Montana by 258 anglers during the period of July 1992 through August 1993. A total of 59 possible river sites were defined, based on geographical and other relevant factors. Each site contains one or more of the stream segments used in the Montana River Information System. The following variables enter as elements of x for each site:

1. Fish stock, measured in units of 100 fish per 1000 feet of river.
2. Aesthetics rating, measured on a scale of 0 to 3, with 3 being the highest.
3. Trip cost: cost of traveling from the angler's home to the site, including the variable cost of driving (gas, maintenance, tires, oil) and the value of time spent driving (with time valued at one-third the angler's wage.)
4. Indicator that the *Angler's Guide to Montana* lists the site as a major fishing site.
5. Number of campgrounds per U.S. Geological Survey (USGS) block in the site.
6. Number of state recreation access areas per USGS block in the site.
7. Number of restricted species at the site.
8. Log of the size of the site, in USGS blocks.

The coefficients of variables 4–7 can logically take either sign; for example, some anglers might like having campgrounds and others prefer the privacy that comes from not having nearby campgrounds. Each of these coefficients is given an independent normal distribution with mean and standard deviation that are estimated. The coefficients for trip cost, fish stock, and aesthetics rating of the site are expected to have the same sign for all anglers, with only their magnitudes differing over anglers. These coefficients are given independent lognormal distributions. The mean and standard deviation of the log of the coefficient are estimated, and the mean and standard deviation of the coefficient itself are calculated from these estimates. Since the lognormal distribution is defined over the positive range and trip cost is expected to have a negative coefficient for all anglers, the negative of trip cost enters the model. The coefficient for the log of size is assumed to be fixed. This variable allows for the fact that the probability of visiting a larger site is higher than that for a smaller site, all else equal. Having the coefficient of this variable vary over people, while possible, would not be particularly meaningful. A version of the model with correlated coefficients is given by Train (1998). The site choice model is part of an overall model, given by Desvousges *et al.* (1996), of the joint choice of trip frequency and site choice.

Table 6.1. *Mixed logit model of river fishing site choice*

Variable	Parameter	Value	Std. Error
Fish stock	Mean of ln(coefficient)	−2.876	0.6066
	Std. dev. of ln(coefficient)	1.016	0.2469
Aesthetics	Mean of ln(coefficient)	−0.794	0.2287
	Std. dev. of ln(coefficient)	0.849	0.1382
Total cost (neg.)	Mean of ln(coefficient)	−2.402	0.0631
	Std. dev. of ln(coefficient)	0.801	0.0781
Guide lists as major	Mean coefficient	1.018	0.2887
	Std. dev. of coefficient	2.195	0.3518
Campgrounds	Mean coefficient	0.116	0.3233
	Std. dev. of coefficient	1.655	0.4350
Access areas	Mean coefficient	−0.950	0.3610
	Std. dev. of coefficient	1.888	0.3511
Restricted species	Mean coefficient	−0.499	0.1310
	Std. dev. of coefficient	0.899	0.1640
Log(size)	Mean coefficient	0.984	0.1077
Likelihood ratio index		0.5018	
SLL at convergence		−1932.33	

Simulation was performed using one thousand random draws for each sampled angler. The results are given in Table 6.1. The standard deviation of each random coefficient is highly significant, indicating that these coefficients do indeed vary in the population.

Consider first the normally distributed coefficients. The estimated means and standard deviations of these coefficients provide information on the share of the population that places a positive value on the site attribute and the share that places a negative value. The distribution of the coefficient of the indicator that the *Angler's Guide to Montana* lists the site as a major site obtains an estimated mean of 1.018 and estimated standard deviation of 2.195, such that 68 percent of the distribution is above zero and 32 percent below. This implies that being listed as a major site in the *Angler's Guide to Montana* is a positive inducement for about two-thirds of anglers and a negative factor for the other third, who apparently prefer more solitude. Campgrounds are preferred by about half (53 percent) of anglers and avoided by the other half. And about one-third of anglers (31 percent) are estimated to prefer having numerous access areas, while the other two-thirds prefer there being fewer access areas.

Consider now the lognormal coefficients. Coefficient β^k follows a lognormal if the log of β^k is normally distributed. We parameterize the lognormal distribution is terms of the underlying normal. That is, we

estimate parameters m and s that represent the mean and variance of the log of the coefficient: $\ln \beta^k \sim N(m, s)$. The mean and variance of β^k are then derived from the estimates of m and s. The median is $\exp(m)$, the mean is $\exp(m + s/2)$, and the variance is $\exp(2m + s)\,[\exp(s) - 1]$. The point estimates imply that the coefficients of fish stock, aesthetics, and trip cost have the following median, mean, and standard deviations:

Variable	Median	Mean	Std. Dev.
Fish stock	0.0563	0.0944	0.1270
Aesthetics	0.4519	0.6482	0.6665
Trip cost	0.0906	0.1249	0.1185

The ratio of an angler's fish stock coefficients to the trip cost coefficient is a measure of the amount that the angler is willing to pay to have additional fish in the river. Since the ratio of two independent lognormally distributed terms is also lognormally distributed, we can calculate moments for the distribution of willingness to pay. The log of the ratio of the fish stock coefficient to the trip cost coefficient has estimated mean -0.474 and standard deviation of 1.29. The ratio itself therefore has median 0.62, mean 1.44, and standard deviation 2.96. That is, the average willingness to pay to have the fish stock raised by 100 fish per 1000 feet of river is estimated to be \$1.44, and there is very wide variation in anglers' willingness to pay for additional fish stock. Similarly, \$9.87 is the estimated average willingness to pay for a site that has an aesthetics rating that is higher by 1, and again the variation is fairly large.

As this application illustrates, the mixed logit provides more information than a standard logit, in that the mixed logit estimates the extent to which anglers differ in their preferences for site attributes. The standard deviations of the coefficients enter significantly, indicating that a mixed logit provides a significantly better representation of the choice situation than standard logit, which assumes that coefficients are the same for all anglers. The mixed logit also allows for the fact that several trips are observed for each sampled angler and that each angler's preferences apply to each of the angler's trips.