Chapter 3

Logit

3.1 Choice probabilities

By far the easiest and most widely used discrete choice model is logit. Its popularity is due to the fact that the formula for the choice probabilities takes a closed form and is readily interpretable. Originally, the logit formula was derived by Luce (1959) from assumptions about the characteristics of choice probabilities, namely the "independence from irrelevant alternatives" property discussed in 3.3.2. Marschak (1960) showed that these axioms implied that the model is consistent with utility maximization. The relation of the logit formula to the distribution of unobserved utility (as opposed to the characteristics of choice probabilities) was developed by Marley, as cited by Luce and Suppes (1965), who showed that the extreme value distribution leads to the logit formula. McFadden (1974) completed the analysis by showing the opposite: that the logit formula for the choice probabilities necessarily implies that unobserved utility is distributed extreme value. In his Nobel lecture, McFadden (2001) provides a fascinating history of the development of this path-breaking model.

To derive the logit model, we use the general notation from Chapter 2 and add a specific distribution for unobserved utility. A decision-maker, labeled n, faces J alternatives. The utility that the decision-maker obtains from alternative j is decomposed into (1) a part labeled V_{nj} that is known by the researcher up to some parameters, and (2) an unknown part ε_{nj} that is treated by the researcher as random: $U_{nj} = V_{nj} + \varepsilon_{nj} \; \forall \, j$. The logit model is obtained by assuming that each ε_{nj} is distributed independently, identically extreme value. The distri-

bution is also called Gumbel and type I extreme value (and sometimes, mistakenly, Weibull.) The density for each unobserved component of utility is

$$f(\varepsilon_{nj}) = e^{-\varepsilon_{nj}} e^{-e^{-\varepsilon_{nj}}} \tag{3.1}$$

and the cumulative distribution is

$$F(\varepsilon_{nj}) = e^{-e^{-\varepsilon_{nj}}}. \tag{3.2}$$

The variance of this distribution is $\pi^2/6$. By assuming the variance is $\pi^2/6$, we are implicitly normalizing the scale of utility, as discussed in section 2.5. We return to this issue, and its impact on interpretation, in the next section. The mean of the extreme value distribution is not zero; however, the mean is immaterial since only differences in utility matter (see Chapter 2) and the difference between two random terms that have the same mean has itself a mean of zero.

The difference between two extreme value variables is distributed logistic. That is, if ε_{nj} and ε_{ni} are iid extreme value, then $\varepsilon^*_{nji} = \varepsilon_{nj} - \varepsilon_{ni}$ follows the logistic distribution

$$F(\varepsilon^*_{nji}) = \frac{e^{\varepsilon^*_{nji}}}{1 + e^{\varepsilon^*_{nji}}}. \tag{3.3}$$

This formula is sometimes used in describing binary logit models, i.e., models with two alternatives. Using the extreme value distribution for the errors (and hence the logistic distribution for the error differences) is nearly the same as assuming that the errors are independently normal. The extreme value gives slightly fatter tails than a normal, which means that it allows for slightly more aberrant behavior than the normal. Usually, however, the difference between extreme value and independent normal errors is indistinguishable empirically.

The key assumption is not so much the shape of the distribution as that the errors are independent of each other. This independence means that the unobserved portion of utility for one alternative is unrelated to the unobserved portion of utility for another alternative. It is a fairly restrictive assumption, and the development of other models such as those described in Chapters 4-6 has arisen largely for the purpose of avoiding this assumption and allowing for correlated errors.

It is important to realize that the independence assumption is not as restrictive as it might at first seem, and in fact can be interpreted

as a natural outcome of a well-specified model. Recall from Chapter 2 that ε_{nj} is defined as the difference between the utility that the decision-maker actually obtains, U_{nj}, and the representation of utility that the researcher has developed using observed variables, V_{nj}. As such, ε_{nj} and its distribution depend on the researcher's specification of representative utility; it is not defined by the choice situation *per se*. In this light, the assumption of independence obtains a different stature. Under independence, the error for one alternative provides no information to the researcher about the error for another alternative. Stated equivalently, the researcher has specified V_{nj} sufficiently that the remaining, unobserved portion of utility is essentially "white noise." In a deep sense, the ultimate goal of the researcher is to represent utility so well that the only remaining aspects constitute simply white noise. That is: the goal is to specify utility well enough that a logit model is appropriate. Seen in this way, the logit model is the ideal rather than a restriction.

If the researcher thinks that the unobserved portion of utility is correlated over alternatives given his specification of representative utility, then she has three options (1) use a different model that allows for correlated errors, such as those described in Chapters 4-6, (2) re-specify representative utility so that the source of the correlation is captured explicitly such that the remaining errors are independent, or (3) use the logit model under the current specification of representative utility considering the model to be an approximation. The viability of the last option depends, of course, on the goals of the research. Violations of the logit assumptions seem to have less impact when estimating average preferences than when forecasting substitution patterns. These issues are discussed in subsequence sections.

We now derive the logit choice probabilities, following McFadden (1974). The probability that decision-maker n chooses alternative i is

$$\begin{aligned} P_{ni} &= Prob(V_{ni} + \varepsilon_{ni} > V_{nj} + \varepsilon_{nj} \ \forall j \neq i) \\ &= Prob(\varepsilon_{nj} < \varepsilon_{ni} + V_{ni} - V_{nj} \ \forall j \neq i) \end{aligned} \tag{3.4}$$

If ε_{ni} is considered given, this expression is the cumulative distribution for each ε_{nj} evaluated at $\varepsilon_{ni} + V_{ni} - V_{nj}$, which, according to (3.2), is $exp(-exp-(\varepsilon_{ni} + V_{ni} - V_{nj})))$. Since the ε's are independent, this cumulative distribution over all $j \neq i$ is the product of the individual

cumulative distributions:

$$P_{ni} \mid \varepsilon_{ni} = \prod_{j \neq i} e^{-e^{-(\varepsilon_{ni} + V_{ni} - V_{nj})}}.$$

Of course, ε_{ni} is not given, and so the choice probability is the integral of $P_{ni} \mid \varepsilon_{ni}$ over all values of ε_{ni} weighted by its density (3.1):

$$P_{ni} = \int \left(\prod_{j \neq i} e^{-e^{-(\varepsilon_{ni} + V_{ni} - V_{nj})}} \right) e^{-\varepsilon_{ni}} e^{-e^{-\varepsilon_{ni}}} d\varepsilon_{ni}. \tag{3.5}$$

Some algebraic manipulation of this integral results in a succinct, closed-form expression:

$$P_{ni} = \frac{e^{V_{ni}}}{\sum_j e^{V_{nj}}} \tag{3.6}$$

which is the logit choice probability. The algebra that obtains (3.6) from (3.5) is given in the last section of this chapter, for readers who are interested.

Representative utility is usually specified to be linear in parameters: $V_{nj} = \beta' x_{nj}$ where x_{nj} is a vector of observed variables relating to alternative j. With this specification, the logit probabilities become:

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

Under fairly general conditions, any function can be approximated arbitrarily closely by one that is linear in parameters. The assumption is therefore fairly benign. Importantly, McFadden (1974) demonstrated that the log-likelihood function with these choice probabilities is globally concave in parameters β, which helps in the numerical maximization procedures (as discussed in Chapter 8.) Numerous computer packages contain routines for estimation of logit models with linear-in-parameters representative utility.

The logit probabilities exhibit several desirable properties. First, P_{ni} is necessarily between zero and one, as required for a probability. When V_{ni} rises, reflecting an improvement in the observed attributes of the alternative, with V_{nj} $\forall\, j \neq i$ held constant, P_{ni} approaches one. And P_{ni} approaches zero when V_{ni} decreases, since the exponential in the numerator of (3.6) approaches zero as V_{ni} approaches $-\infty$. The logit probability for an alternative is never exactly zero. If the research

believes that an alternative has actually no chance of being chosen by a decision-maker, the researcher can exclude that alternative from the choice set. A probability of exactly 1 is obtained only if the choice set consists of a single alternative.

Second, the choice probabilities for all alternatives sum to one: $\sum_{i=1}^{J} P_{ni} = \sum_i exp(V_{ni}) / \sum_j exp(V_{nj}) = 1$. The decision-maker necessarily chooses one of the alternatives. The denominator in (3.6) is simply the sum of the numerator over all alternatives, which gives this summing-up property automatically. With logit, as well as with some more complex models such as the nested logit models of Chapter 4, interpretation of the choice probabilities is facilitated by recognition that the denominator serves to assure that the probabilities sum to one. In other models, such as mixed logit and probit, there is no denominator *per se* to interpret in this way.

The relation of the logit probability to representative utility is sigmoid, or S-shaped, as shown in Figure 3.1. This shape has implications for the impact of changes in explanatory variables. If the representative utility of an alternative is very low compared with other alternatives, a small increase in the utility of the alternative has little effect on the probability of its being chosen: the other alternatives are still sufficiently better that this small improvement doesn't help much. Similarly, if one alternative is far superior to the others in observed attributes, a further increase in its representative utility has little effect on the choice probability. The point at which the increase in representative utility has the greatest effect on the probability of its being chosen is when the probability is close to 0.5, meaning a 50-50 chance of the alternative being chosen. In this case, a small improvement "tips the balance" in people's choices, inducing a large change in probability. The sigmoid shape of logit probabilities is shared by most discrete choice models and has important implications for policy makers. For example, improving bus service in areas where the service is so poor that few travelers take the bus would be less effective, in terms of transit ridership, than making the same improvement in areas where bus service is already sufficiently good to induce a moderate share of travelers to choose it (but not so good that nearly everyone does.)

The logit probability formula is easily interpretable in the context of an example. Consider a binary choice situation first: a household's choice between a gas and electric heating system. Suppose that the

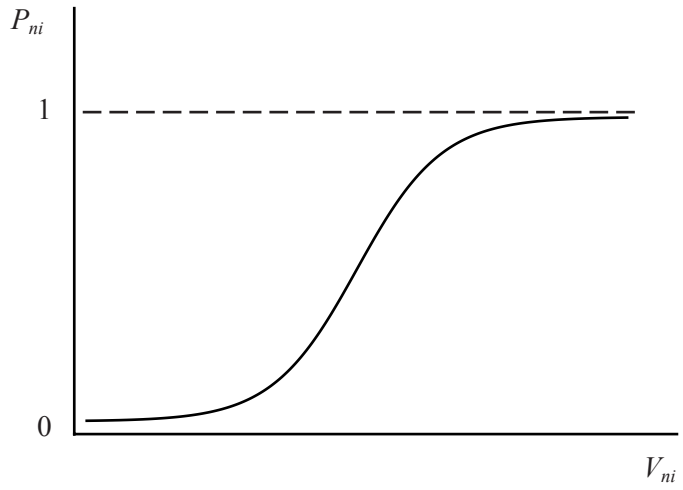

Figure 3.1: Graph of logit curve.

utility the household obtains from each type of system depends only on the purchase price, annual operating cost, and the household's view of the convenience and quality of heating with each type of system and the relative aesthetics of the systems within the house. The first two of these factors can be observed by the researcher, but the researcher cannot observe the others. If the researcher considers the observed part of utility to be a linear function of the observed factors, then the utility of each heating system can be written as: $U_g = \beta_1 PP_g + \beta_2 OC_g + \varepsilon_g$ and $U_e = \beta_1 PP_e + \beta_2 OC_e + \varepsilon_e$, where the subscripts g and e denote gas and electric, PP and OC are the purchase price and operating cost, β_1 and β_2 are scalar parameters, and the subscript n for the household is suppressed. Since higher costs mean less money to spend on other goods, we expect utility to drop as purchase price or operating cost rises (with all else held constant): $\beta_1 < 0$ and $\beta_2 < 0$.

The unobserved component of utility for each alternative, ε_g and ε_e, varies over households depending on how each household views the quality, convenience and aesthetics of each type of system. If these unobserved components are distributed iid extreme value, then the probability that the household will choose gas heating is

$$P_g = \frac{e^{\beta_1 PP_g + \beta_2 OC_g}}{e^{\beta_1 PP_g + \beta_2 OC_g} + e^{\beta_1 PP_e + \beta_2 OC_e}} \tag{3.7}$$

and the probability of electric heating is the same but with $exp(\beta_1 PP_e + \beta_2 OC_e)$ as the numerator. The probability of choosing a gas system decreases if its purchase price or operating cost rises while that of the electric system remains the same (assuming that β_1 and β_2 are negative as expected.)

As in most discrete choice models, the ratio of coefficients in this example has economic meaning. In particular, the ratio β_2/β_1 represents the household's willingness to pay for operating cost reductions. If β_1 were estimated as -0.20 and β_2 as -1.14, these estimates would imply that households are willing to pay up to $-1.14/-0.20 = 5.70$ dollars more for a system whose annual operating costs are one dollar less. This relation is derived as follows. By definition, a household's willingness to pay for operating cost reductions is the increase in purchase price that keeps the household's utility constant given a reduction in operating costs. We take the total derivative of utility with respect to purchase price and operating cost and set this derivative to zero so that utility doesn't change: $\partial U = \beta_1 \partial PP + \beta_2 \partial OC = 0$. We then solve for the change in purchase price that keeps utility constant (i.e., satisfies this equation) for a change in operating costs: $\partial PP/\partial OC = -\beta_2/\beta_1$. The negative sign indicates that the two changes are in the opposite direction: to keep utility constant, purchase price rises when operating cost decreases.

In this binary choice situation, the choice probabilities can be expressed in another, even more succinct form. Dividing the numerator and denominator of (3.7) by the numerator, and recognizing that $exp(a)/exp(b) = exp(a-b)$, we have

$$P_g = \frac{1}{1 + e^{(\beta_1 PP_e + \beta_2 OC_e) - (\beta_1 PP_g + \beta_2 OC_g)}}.$$

In general, binary logit probabilities with representative utilities V_{n1} and V_{n2} can be written $P_{n1} = 1/(1 + exp(V_{n2} - V_{n1}))$ and $P_{n2} = 1/(1 + exp(V_{n1} - V_{n2}))$. If only demographics of the decision-maker, s_n, enter the model, and the coefficients of these demographic variables are normalized to zero for the first alternative (as described in Chapter 2), the probability of the first alternative is $P_{n1} = 1/(1 + e^{\alpha' s_n})$, which is the form that is used in most textbooks and computer manuals for binary logit.

Multinomial choice is a simple extension. Suppose there is a third type of heating system, namely oil-fueled. The utility of the oil system

is specified as the same form as for the electric and gas systems: $U_o = \beta_1 PP_o + \beta_2 OC_o + \varepsilon_o$. With this extra option being available, the probability that the household chooses a gas system is:

$$P_g = \frac{e^{\beta_1 PP_g + \beta_2 OC_g}}{e^{\beta_1 PP_g + \beta_2 OC_g} + e^{\beta_1 PP_e + \beta_2 OC_e} + e^{\beta_1 PP_o + \beta_2 OC_o}}$$

which is the same as (3.7) except that an extra term is included in the denominator to represent the oil heater. Since the denominator is larger while the numerator is the same, the probability of choosing a gas system is smaller when an oil system is an option than when not, as one would expect in the real world.

3.2 The scale parameter

In the previous section we derived the logit formula under the assumption that the unobserved factors are distributed extreme value with variance $\pi^2/6$. Setting the variance to $\pi^2/6$ is equivalent to normalizing the model for the scale of utility, as discussed in section 2.5. It is useful to make these concepts more explicit, to show the role that the variance of the unobserved factors plays in logit models.

In general, utility can be expressed as $U^*_{nj} = V_{nj} + \varepsilon^*_{nj}$ where the unobserved portion has variance $\sigma^2 * (\pi^2/6)$. That is, the variance is any number, re-expressed as a multiple of $\pi^2/6$. Since the scale of utility is irrelevant to behavior, utility can be divided by σ without changing behavior. Utility becomes $U_{nj} = V_{nj}/\sigma + \varepsilon_{nj}$ where $\varepsilon_{nj} = \varepsilon^*_{nj}/\sigma$. Now the unobserved portion has variance $\pi^2/6$: $Var(\varepsilon_{nj}) = Var(\varepsilon^*_{nj}/\sigma) = (1/\sigma^2)Var(\varepsilon^*_{nj}) = (1/\sigma^2) * \sigma^2 * (\pi^2/6) = \pi^2/6$. The choice probability is

$$P_{ni} = \frac{e^{V_{ni}/\sigma}}{\sum_j e^{V_{nj}/\sigma}}$$

which is the same formula as in equation 3.6 but with representative utility divided by σ. If V_{nj} is linear in parameters with coefficient β^*, the choice probabilities become:

$$P_{ni} = \frac{e^{(\beta^*/\sigma)' x_{ni}}}{\sum_j e^{(\beta^*/\sigma)' x_{nj}}}$$

Each of the coefficients is scaled by $1/\sigma$. The parameter σ is called the scale parameter, because it scales the coefficients to reflect the variance of the unobserved portion of utility.

Only the ratio β^*/σ can be estimated; β^* and σ are not separately identified. Usually, the model is expressed in its scaled form, with $\beta = \beta^*/\sigma$, which gives the standard logit expression

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

The parameters β are estimated, but for interpretation it is useful to recognize that these estimated parameters are actually estimates of the "original" coefficients β^* divided by the scale parameter σ. The coefficients that are estimated indicate the impact of each observed variable *relative to* the variance of the unobserved factors. A larger variance in unobserved factors leads to lower coefficients, even if the observed factors have the same impact on utility (i.e., higher σ means lower β even if β^* is the same.)

The scale parameter does not affect the ratio of any two coefficients, since it drops out of the ratio; for example, $\beta_1/\beta_2 = (\beta_1^*/\sigma)/(\beta_2^*/\sigma) = \beta_1^*/\beta_2^*$, where the subscripts refer to the first and second coefficients. Willingness to pay, values of time, and other measures of marginal rates of substitution are not affected by the scale parameter. Only the interpretation of the magnitudes of all coefficients is affected.

So far we have assumed that the variance of the unobserved factors is the same for all decision-makers, since the same σ is used for all n. Suppose instead that the unobserved factors have greater variance for some decision-makers than others. In section 2.5, we discuss a situation where the variance of unobserved factors is different in Boston than Chicago. Denote the variance for all decision-makers in Boston as $(\sigma^B)^2(\pi^2/6)$ and that for decision-makers in Chicago as $(\sigma^C)^2(\pi^2/6)$. The ratio of variance in Chicago to that in Boston is $k = (\sigma^C/\sigma^B)^2$. The choice probabilities for people in Boston become

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

and for people in Chicago are

$$P_{ni} = \frac{e^{(\beta/\sqrt{k})' x_{ni}}}{\sum_j e^{(\beta/\sqrt{k})' x_{nj}}}$$

where $\beta = \beta^*/\sigma^B$. The ratio of variances k is estimated along with the coefficients β. The estimated β's are interpreted as being relative

to the variance of unobserved factors in Boston, and the estimated k provides information on the variance in Chicago relative to that in Boston. More complex relations can be obtained by allowing the variance for an observation to dependent on more factors. Also, data from different data sets can often be expected to have different variance for unobserved factors, giving a different scale parameter for each data set. Ben-Akiva and Morikawa (1990) and Swait and Louviere (1993) discuss these issues and provide more examples.

3.3 Power and limitations of logit

Three topics elucidate the power of logit models to represent choice behavior, as well as delineating the limits to that power. These topics are: taste variation, substitution patterns, and repeated choices over time. The applicability of logit models can be summarized as follows:

1. Logit can represent systematic taste variation (that is, taste variation that relates to observed characteristics of the decision-maker) but not random taste variation (differences in tastes that cannot be linked to observed characteristics).

2. The logit model implies proportional substitution across alternatives, given the researcher's specification of representative utility. To capture more flexible forms of substitution, other models are needed.

3. If unobserved factors are independent over time in repeated choice situations, then logit can capture the dynamics of repeated choice, including state-dependence. However, logit cannot handle situations where unobserved factors are correlated over time.

We elaborate each of these statements in the next three subsections.

3.3.1 Taste variation

The value or importance that decision-makers place on each attribute of the alternatives varies, in general, over decision-makers. For example, the size of a car is probably more important to households with many members than to smaller households. Low-income households are probably more concerned about the purchase price of a good, relative to its other characteristics, than higher-income households. In

choosing which neighborhood to live in, households with young children will be more concerned about the quality of schools than those without children. And so on. Decision-makers' tastes also vary for reasons that are not linked to observed demographic characteristics, just because different people are different. Two people who have the same income, education, etc., will make different choices, reflecting their individual preferences and concerns.

Logit models can capture taste variations, but only within limits. In particular, tastes that vary systematically with respect to observed variables can be incorporated in logit models, while tastes that vary with unobserved variables or purely randomly cannot be handled. The following example illustrates the distinction.

Consider households' choice among makes and models of cars to buy. Suppose for simplicity that the only two attributes of cars that the researcher observes are the purchase price, PP_j for make/model j, and inches of shoulder room, SR_j, which is a measure of the interior size of a car. The value that households place on these two attributes varies over households, and so utility is written as

$$U_{nj} = \alpha_n SR_j + \beta_n PP_j + \varepsilon_{nj}, \tag{3.8}$$

where α_n and β_n are parameters specific to household n.

The parameters vary over households reflecting differences in taste. Suppose for example that the value of shoulder room varies with the number of members in the households, M_n, but nothing else:

$$\alpha_n = \rho M_n,$$

so that as M_n increases, the value of shoulder room, α_n, also increases. Similarly, suppose the importance of purchase price is inversely related to income, I_n, so that low income households place more importance on purchase price:

$$\beta_n = \theta / I_n.$$

Substituting these relations into (3.8) produces

$$U_{nj} = \rho(M_n SR_j) + \theta(PP_j / I_n) + \varepsilon_{nj}.$$

Under the assumption that each ε_{nj} is iid extreme value, a standard logit model obtains with two variables entering representative utility, both of which are an interaction of a vehicle attribute with a household characteristic.

Other specifications for the variation in tastes can be substituted. For example, the value of shoulder room might be assumed to increase with household size, but at a decreasing rate, so that $\alpha_n = \rho M_n + \phi M_n^2$ where ρ is expected to be positive and ϕ negative. Then $U_{nj} = \rho(M_n SR_j) + \phi(M_n^2 SR_j) + \theta(PP_j/I_n) + \varepsilon_{nj}$, which results in a logit model with three variables entering representative utility.

The limitation of the logit model arises when we attempt to allow tastes to vary with respect to unobserved variables or purely randomly. Suppose for example that the value of shoulder room varied with household size plus some other factors (e.g., size of the people themselves, or frequency with which the household travels together) that are unobserved by the researcher and hence considered random:

$$\alpha_n = \rho M_n + \mu_n,$$

where μ_n is a random variable. Similarly, the importance of purchase price consists of its observed and unobserved components:

$$\beta_n = \theta/I_n + \eta_n.$$

Substituting into (3.8) produces

$$U_{nj} = \rho(M_n SR_j) + \mu_n SR_j + \theta(PP_j/I_n) + \eta_n PP_j + \varepsilon_{nj}.$$

Since μ_n and η_n are not observed, the terms $\mu_n SR_j$ and $\eta_n PP_j$ become part of the unobserved component of utility,

$$U_{nj} = \rho(M_n SR_j) + \theta(PP_j/I_n) + \tilde{\varepsilon}_{nj},$$

where $\tilde{\varepsilon}_{nj} = \mu_n SR_j + \eta_n PP_j + \varepsilon_{nj}$. The new error terms $\tilde{\varepsilon}_{nj}$ cannot possibly be distributed independently and identically as required for the logit formulation. Since μ_n and η_n enter each alternative, $\tilde{\varepsilon}_{nj}$ is necessarily correlated over alternatives: $Cov(\tilde{\varepsilon}_{nj}, \tilde{\varepsilon}_{nk}) = Var(\mu_n) SR_j SR_k + Var(\eta_n) PP_j PP_k \neq 0$ for any two cars j and k. Furthermore, since SR_j and PP_j vary over alternatives, the variance of $\tilde{\varepsilon}_{nj}$ varies over alternatives, violating the assumption of identically distributed errors: $Var(\tilde{\varepsilon}_{nj}) = Var(\mu_n) SR_j^2 + Var(\eta_n) PP_j^2 + Var(\varepsilon_{nj})$, which is different for different j.

This example illustrates the general point that when tastes vary systematically in the population in relation to observed variables, the variation can be incorporated into logit models. However, if taste

variation is at least partly random, logit is a misspecification. As an approximation, logit might be able to capture the average tastes fairly well even when tastes are random, since the logit formula seems to be fairly robust to misspecifications. The researcher might therefore choose to use logit even when she knows that tastes have a random component, for the sake of simplicity. However, there is no guarantee that a logit model will approximate the average tastes. And, even if it does, logit does not provide information on the distribution of tastes around the average. This distribution can be important in many situations, such as forecasting the penetration of a new product that appeals to a minority of people rather than to the average tastes. To incorporate random taste variation appropriately and fully, a probit or mixed logit model can be used instead.

3.3.2 Substitution patterns

When the attributes of one alternative improve (e.g., its price drops), the probability of its being chosen rises. Some of the people who would have chosen other alternatives under the original attributes now choose this alternative instead. Since probabilities sum to one over alternatives, an increase in the probability of one alternative necessarily means a decrease in probability for other alternatives. The pattern of substitution among alternatives has important implications in many situations. For example, when a cell phone manufacturer launches a new product with extra features, the firm is vitally interested in knowing the extent to which the new product will draw customers away from its other cell phones rather than from competitors' phones, since the firm makes more profit from the later than the former. Also, as we will see, the pattern of substitution affects the demand for a product and the change in demand when attributes change. Substitution patterns are therefore important even when the researcher is only interested in market share without being concerned about where the share comes from.

The logit model implies a certain pattern of substitution across alternatives. If substitution actually occurs in this way given the researcher's specification of representative utility, then the logit model is appropriate. However, to allow for more general patterns of substitution and to investigate which pattern is most accurate, more flexible models are needed. The issue can be seen in either of two ways, as a

restriction on the ratios of probabilities and/or as a restriction on the cross-elasticities of probabilities. We present each way of characterizing the issue below.

The independence from irrelevant alternatives (IIA) property

For any two alternatives i and k, the ratio of the logit probabilities is

$$\begin{aligned}\frac{P_{ni}}{P_{nk}} &= \frac{e^{V_{ni}}/\sum_j e^{V_{nj}}}{e^{V_{nk}}/\sum_j e^{V_{nj}}} \\ &= \frac{e^{V_{ni}}}{e^{V_{nk}}} = e^{V_{ni}-V_{nk}}.\end{aligned}$$

This ratio does not depend on any alternatives other than i and k. That is, the relative odds of choosing i over k is the same no matter what other alternatives are available or what the attributes of the other alternatives are. Since the ratio is independent from alternatives other than i and k, it is said to be independent from "irrelevant" alternatives. The logit model exhibits this "independence from irrelevant alternatives," or IIA, property.

In many settings, choice probabilities that exhibit IIA provide an accurate representation of reality. In fact, Luce (1959) considered IIA to be a property of appropriately specified choice probabilities. He derived the logit model directly from an assumption that choice probabilities exhibit IIA, rather than (as we have done) derive the logit formula from an assumption about the distribution of unobserved utility and then observe that IIA is a resulting property.

While the IIA property is realistic in some choice situations, it is clearly inappropriate in others, as first pointed out by Chipman (1960) and Debreu (1960). Consider the famous red bus/blue bus problem. A traveler has a choice of going to work by car or taking a blue bus. For simplicity assume that the representative utility of the two modes are the same, such that the choice probabilities are equal: $P_c = P_{bb} = 1/2$, where c is car and bb is blue bus. In this case, the ratio of probabilities is one: $P_c/P_{bb} = 1$.

Now suppose that a red bus is introduced and that the traveler considers the red bus to be exactly like the blue bus. The probability that the traveler will take the red bus is therefore the same as for the blue bus, such that the ratio of their probabilities is one: $P_{rb}/P_{bb} = 1$. However, in the logit model the ratio P_c/P_{bb} is the same whether or

not another alternative, in this case the red bus, exists. This ratio therefore remains at one. The only probabilities for which $P_c/P_{bb} = 1$ and $P_{rb}/P_{bb} = 1$ are $P_c = P_{bb} = P_{rb} = 1/3$, which are the probabilities that the logit model predicts.

In real life, however, we would expect the probability of taking a car to remain the same when a new bus is introduced that is exactly the same as the old bus. We would also expect the original probability of taking bus to be split between the two buses after the second one is introduced. That is, we would expect $P_c = 1/2$ and $P_{bb} = P_{rb} = 1/4$. In this case, the logit model, because of its IIA property, overestimates the probability of taking either of the buses and underestimates the probability of taking a car. The ratio of probabilities of car and blue bus, P_c/P_{bb}, actually changes with the introduction of the red bus, rather than remaining constant as required by the logit model.

This red bus/blue bus example is rather stark and unlikely to be encountered in the real world. However, the same kind of misprediction arises with logit models whenever the ratio of probabilities for two alternatives changes with the introduction or change in another alternative. For example, suppose a new transit mode is added that is similar, but not exactly like, the existing modes, such as an express bus along a line that already has standard bus service. This new mode might be expected to reduce the probability of regular bus by a greater proportion than it reduces the probability of car, such that ratio of probabilities for car and regular bus does not remain constant. The logit model would overpredict demand for the two bus modes in this situation. Other examples are given by, for example, Ortuzar (1983) and Brownstone and Train (1999).

Proportional substitution

The same issue can be expressed in terms of the cross-elasticities of logit probabilities. Let us consider changing an attribute of alternative j. We want to know the effect of this change on the probabilities for all the *other* alternatives. Section (3.6) derives the formula for the elasticity of P_{ni} with respect to a variable that enters the representative utility of alternative j:

$$E_{iz_{nj}} = -\beta_z z_{nj} P_{nj}$$

where z_{nj} is the attribute of alternative j as faced by person n and β_z is its coefficient (or, if the variable enters representative utility non-

linearly, then β_z is the derivative of V_{nj} with respect to z_{nj}).

This cross-elasticity is the same for all i: i does not enter the formula. An improvement in the attributes of an alternative reduces the probabilities for all the other alternatives by the same percent. If one alternative's probability drops by ten percent, then all the other alternatives' probabilities also drop by ten percent (except of course the alternative whose attribute changed; its probability rises due to the improvement.) A way of stating this phenomenon succinctly is that an improvement in one alternative draws proportionately from the other alternatives. Similarly for a decrease in the representative utility of an alternative: the probabilities for all other alternatives rise by the same percent.

This pattern of substitution, which can be called "proportionate shifting," is a manifestation of the IIA property. The ratio of probabilities for alternatives i and k stays constant when an attribute of alternative j changes only if the two probabilities change by the same proportion. With superscript 0 denoting probabilities before the change and 1 after, the IIA property requires that

$$\frac{P^1_{ni}}{P^1_{nk}} = \frac{P^0_{ni}}{P^0_{nk}}$$

when an attribute of alternative j changes. This equality can only be maintained if each probability changes by the same proportion: $P^1_{ni} = \lambda P^0_{ni}$ and $P^1_{nk} = \lambda P^0_{nk}$ where both λ's are the same.

Proportionate substitution can be realistic for some situations, in which case the logit model is appropriate. In many settings, however, other patterns of substitution can be expected, and imposing proportionate substitution through the logit model can lead to unrealistic forecasts. Consider a situation that is important to the California Energy Commission (CEC), which has the responsibility of investigating policies to promote energy efficient vehicles in California and reducing the state's reliance on gasoline for cars. Suppose for the sake of illustration that there are three kinds of vehicles: large gas cars, small gas cars, and small electric cars. Suppose also that under current conditions the probabilities that a household will choose each of these vehicles are .66, .33, and .01, respectively. The CEC is interested in knowing the impact of subsidizing the electric cars. Suppose the subsidy is sufficient to raise the probability for the electric car from .01 to .10. By the logit model, the probability for each of the gas cars would

be predicted to drop by the same percent. The large gas car would drop by ten percent from .66 to .60 and the probability for the small gas car would drop by the same ten percent from .33 to .30. In terms of absolute numbers, the increased probability for the small electric car (.09) is predicted by the logit model to come twice as much from large gas cars (.06) as from small gas cars (0.03).

This pattern of substitution is clearly unrealistic. Since the electric car is small, subsidizing it can be expected to draw more from small gas cars than from large gas cars. In terms of cross-elasticities, we would expect the cross-elasticity for small gas cars with respect to an improvement in small electric cars to be higher than that for large gas cars. This difference is important in the CEC's policy analysis. The logit model will over-predict the gas savings that result from the subsidy, since it over-predicts the substitution away from large gas cars (the "gas guzzlers") and under-predicts the substitution away from small "gas-sipper" cars. From a policy perspective, this misprediction can be critical, causing a subsidy program to seem more beneficial than it actually is. This is the reason that the CEC uses models that are more general than logit to represent substitution across vehicles. The nested logit, probit and mixed logit models of Chapters 4-6 provide viable options for the researcher.

Advantages of IIA

As just discussed, the IIA property of logit can be unrealistic in many settings. However, when the IIA property reflects reality (or an adequate approximation to reality), considerable advantages are gained by its employment. First, because of the IIA property, it is possible to estimate model parameters consistently on a subset of alternatives for each sampled decision-maker. For example, in a situation with 100 alternatives, the researcher might, so as to reduce computer time, estimate on a subset of 10 alternatives for each sampled person, with the person's chosen alternative included as well as 9 alternatives randomly selected from the remaining 99. Since relative probabilities within a subset of alternatives are unaffected by the attributes or existence of alternatives not in the subset, exclusion of alternatives in estimation does not affect the consistency of the estimator. Details of this type of estimation are given in section (3.7.1). This fact has considerable practical importance. In analyzing choice situations for which the number

of alternatives is large, estimation on a subset of alternatives can save substantial amounts of computer time. At an extreme, the number of alternatives might be so large as to preclude estimation altogether if it were not possible to utilize a subset of alternatives.

Another practical use of the IIA property arises when the researcher is only interested in examining choices among a subset of alternatives and not among all alternatives. For example, consider a researcher who is interested in understanding the factors that affect workers' choice between car and bus modes for travel to work. The full set of alternative modes includes walking, bicycling, motor-biking, skate-boarding, etc. If the researcher believed that the IIA property holds adequately well in this case, she could estimate a model with only car and bus as the alternatives and exclude from the analysis sampled workers who used other modes. This strategy would save the researcher considerable time and expense developing data on the other modes without hampering her ability to examine the factors related to car and bus.

Tests of IIA

Whether IIA holds in a particular setting is an empirical question, amenable to statistical investigation. Tests of IIA were first developed by McFadden, Train and Tye (1978). Two types of tests are suggested. First, the model can be re-estimated on a subset of the alternatives. Under IIA, the ratio of probabilities for any two alternatives is the same whether or not other alternatives are available. As a result, if IIA holds in reality, then the parameter estimates obtained on the subset of alternatives will not be significantly different from those obtained on the full set of alternatives. A test of the hypothesis that the parameters on the subset are the same as the parameters on the full set constitutes a test of IIA. Hausman and McFadden (1984) provide an appropriate statistic for this type of test. Second, the model can be re-estimated with new, cross-alternative variables, that is, with variables from one alternative entering the utility of another alternative. If the ratio of probabilities for alternatives i and k actually depends on the attributes and existence of a third alternative j (in violation of IIA), then the attributes of alternative j will enter significantly the utility of alternatives i or k within a logit specification. A test of whether cross-alternative variables enter the model therefore constitutes a test of IIA. McFadden (1987) developed a procedure for performing this kind

of test with regressions: with the dependent variable being the residuals of the original logit model and the explanatory variables being appropriately specified cross-alternative variables. Train, Ben-Akiva and Atherton (1989) show how this procedure can be performed conveniently within the logit model itself.

The advent of models that do not exhibit IIA, and especially the development of software for estimating these models, makes testing IIA easier than before. For more flexible specifications, such as GEV and mixed logit, the simple logit model with IIA is a special case that arises under certain constraints on the parameters of the more flexible model. In these cases, IIA can be tested by testing these constraints. For example, a mixed logit model becomes a simple logit if the mixing distribution has zero variance. IIA can be tested by estimating a mixed logit and testing whether the variance of the mixing distribution is in fact zero.

A test of IIA as a constraint on a more general model necessarily operates under the maintained assumption that the more general model is itself an appropriate specification. The tests on subsets of alternatives (Hausman and McFadden, 1984) and cross-alternative variables (McFadden, 1987; Train et al., 1989), while more difficult to perform, operate under less restrictive maintained hypotheses. The counterpoint to this advantage, of course, is that, when IIA fails, these tests do not provide as much guidance on the correct specification to use instead of logit.

3.3.3 Panel data

In many settings, the researcher can observe numerous choices made by each decision-maker. For example, in labor studies, sampled people are observed to work or not work in each month over several years. Data on the current and past vehicle purchases of sampled households might be obtained by a researcher who is interested in the dynamics of car choice. In market research surveys, respondents are often asked a series of hypothetical choice questions, called "stated preference" experiments. For each experiment, a set of alternative products with different attributes are described and the respondent is asked to state which product he would choose. A series of such questions is asked, with the attributes of the products varying so as to determine how the respondent's choice changes when the attributes change. The re-

searcher therefore observes the sequence of choices by each respondent. Data that represent repeated choices like these are called panel data.

If the unobserved factors that affect decision-makers are independent over the repeated choices, then logit can be used to examine panel data in the same way as purely cross-sectional data. Any dynamics related to observed factors that enter the decision-process, such as state-dependence (by which the person's past choices influence their current choices) or lagged response to changes in attributes, can be accommodated. However, dynamics associated with unobserved factors cannot be handled, since the unobserved factors are assumed to be unrelated over choices.

The utility that decision-maker n obtains from alternative j in period or choice situation t is

$$U_{njt} = V_{njt} + \varepsilon_{njt} \; \forall \, j, t.$$

If ε_{njt} is distributed extreme value, independent over n, j, and, importantly, t, then, using the same proof as for (3.6), the choice probabilities are:

$$P_{nit} = \frac{e^{V_{nit}}}{\sum_j e^{V_{njt}}}. \tag{3.9}$$

Each choice situation by each decision-maker becomes a separate observation. If representative utility for each period is specified to depend only on variables for that period, e.g., $V_{njt} = \beta' x_{njt}$ where x_{njt} is a vector of variables describing alternative j as faced by n in period t, then there is essentially no difference in the logit model with panel data than with purely cross-sectional data.

Dynamic aspects of behavior can be captured by specifying representative utility in each period to depend on observed variables from other periods. For example, a lagged price response is represented by entering the price in period $t-1$ as an explanatory variable in the utility for period t. Prices in future periods can be entered, as by Adamowicz (1994), to capture consumers' anticipation of future price changes. Under the assumptions of the logit model, the dependent variable in previous periods can also be entered as an explanatory variable. Suppose for example that there is inertia, or habit formation, in people's choices such that they tend to stay with the alternative that they have previously chosen unless another alternative provides sufficiently higher utility to warrant a switch. This behavior is captured as $V_{njt} = \alpha y_{nj(t-1)} + \beta x_{njt}$ where $y_{njt} = 1$ if n chose j in period t and

zero otherwise. With $\alpha > 0$, the utility of alternative j in the current period is higher if alternative j was consumed in the previous period. The same specification can also capture a type of variety seeking. If α is negative, the consumer obtains higher utility from *not* choosing the same alternative that he chose in the last period. Numerous variations on these concepts are possible. Adamowicz (1994) enters the *number* of times the alternative has been chosen previously, rather than simply a dummy for the immediately previous choice. Erdem (1996) enters the *attributes* of previously chosen alternatives, with the utility of each alternative in the current period depending on the similarity of its attributes to these previously experienced attributes.

The inclusion of the lagged dependent variable does not induce inconsistency in estimation since, for a logit model, the errors are assumed to be independent over time. The lagged dependent variable $y_{nj(t-1)}$ is uncorrelated with the current error ε_{njt} due to this independence. The situation is analogous to linear regression models, where a lagged dependent variable can be added without inducing bias as long as the errors are independent over time.

Of course, the assumption of independent errors over time is severe. Usually, one would expect that there are some factors that are not observed by the researcher that affect each of the decision-makers' choices. In particular, if there are dynamics in the observed factors, then the researcher might expect there to be dynamics in the unobserved factors as well. In these situations, the researcher can either use a model such as probit or mixed logit that allows unobserved factors to be correlated over time, or re-specify representative utility to bring the sources of the unobserved dynamics into the model explicitly such that the remaining errors are independent over time.

3.4 Non-linear representative utility

In some contexts, the researcher will find it useful to allow parameters to enter representative utility nonlinearly. Estimation is more difficult since the log-likelihood function might not be globally concave, and computer routines are not as widely available as for logit models with linear-in-parameters utility. However, the aspects of behavior that the researcher is investigating might include parameters that are interpretable only when they enter utility non-linearly. In these cases, the effort of writing one's own code can be warranted. Two examples

illustrate this point.

Example 1: the goods/leisure trade-off

Consider a workers' choice of mode (car or bus) for trips to work. Suppose that workers also choose the number of hours to work based on the standard trade-off between goods and leisure. Train and McFadden (1978) developed a procedure for examining these interrelated choices. As we see below, the parameters of the workers' utility function over goods and leisure enter non-linearly in the utility for modes of travel.

Assume that workers' preferences regarding goods G and leisure L are represented by a Cobb-Douglas utility function of the form:

$$U = (1 - \beta)lnG + \beta lnL.$$

The parameter β reflects the worker's relative preference for goods and leisure, with higher β implying greater preference for leisure relative to goods. Each worker has a fixed amount of time (24 hours a day) and faces a fixed wage rate, w. In the standard goods-leisure model, the worker chooses the number of hours to work that maximizes U subject to the constraints that (1) the number of hours worked plus the number of leisure hours equals the number of hours available, and (2) the value of goods consumed equals the wage rate times the number of hours worked.

When mode choice is added to the model, the constraints on time and money change. Each mode takes a certain amount of time and costs a certain amount of money. Conditional on choosing car, the worker maximizes U subject to the constrain that (1) the number of hours worked plus the number of leisure hours equals the number of hours available *after the time spent driving to work in the car is subtracted* and (2) the value of goods consumed equals the wage rate times the number of hours worked *minus the cost of driving to work*. The utility associated with choosing to travel by car is the highest value of U that can be attained under these constraints. Similarly, the utility of taking the bus to work is the maximum value of U that can be obtained given the time and money that is left after the bus time and cost are subtracted. Train and McFadden derived the maximizing values of U conditional on each mode. For the U given above, these values are

$$U_j = -\alpha((c_j/w^{\beta}) + w^{1-\beta}t_j) \text{ for } j = \text{car and bus}$$

The cost of travel is divided by w^{β} and travel time is multiplied by $w^{1-\beta}$. The parameter β, which denotes workers' relative preference for goods and leisure, enters the mode choice utility non-linearly. Since this parameter has meaning, the researcher might want to estimate it within this non-linear utility rather than use a linear-in-parameters approximation.

Example 2: geographic aggregation

Models have been developed and widely used for travelers' choice of destination for various types of trips, such as shopping trips, within a metropolitan area. Usually, the metropolitan area is partitioned into "zones," and the models give the probability that a person will choose to travel to a particular zone. Representative utility for each zone depends on the time and cost of travel to the zone plus a variety of variables, such as residential population and retail employment, that reflect reasons that people might want to visit the zone. These latter variables are called "attraction" variables; label them by the vector a_j for zone j. Since it is these attraction variables that give rise to parameters entering non-linearity, assume for simplicity that representative utility depends only on these variables.

The difficulty in specifying representative utility comes in recognizing that the researcher's decision of how large an area to include in each zone is fairly arbitrary. It would be useful to have a model that is not sensitive to the level of aggregation in the zonal definitions. If two zones are combined, it would be useful for the model to give a probability of traveling to the combined zone that is the same as the sum of the probabilities of traveling to the two original zones. This consideration places restrictions on the form of representative utility.

Consider zones j and k which, when combined, are labeled zone c. Population and employment in the combined zone are necessarily the sum of that in the two original zones: $a_j + a_k = a_c$. In order for the models to give the same probability for choosing these zones before and after their merger, the model must satisfy:

$$P_{nj} + P_{nk} = P_{nc}$$

which for logit models takes the form

$$\frac{e^{V_{nj}} + e^{V_{nk}}}{e^{V_{nj}} + e^{V_{nk}} + \sum_{\ell \neq j,k} e^{V_{n\ell}}} = \frac{e^{V_{nc}}}{e^{V_{nc}} + \sum_{\ell \neq j,k} e^{V_{n\ell}}}$$

This equality holds only when $exp(V_{nj}) + exp(V_{nk}) = exp(V_{nc})$. If representative utility is specified as $V_{n\ell} = ln(\beta' a_\ell)$ for all zones ℓ, then the inequality holds: $exp(ln(\beta' a_j)) + exp(ln(\beta' a_k)) = \beta' a_j + \beta' a_k = \beta' a_c = exp(ln(\beta' a_c))$. Therefore, to specify a destination choice model that is not sensitive to the level of zonal aggregation, representative utility needs to be specified with parameters inside a log operation.

3.5 Consumer Surplus

For policy analysis, the researcher is often interested in measuring the change in consumer surplus that is associated with a particular policy. For example, if a new alternative is being considered, such as building a light rail system in a city, then it is important to measure the benefits of the project to see if they warrant the costs. Similarly, a change in the attributes of an alternative can have an impact on consumer surplus that is important to assess. Degradation of the water quality of rivers harms the anglers who can no longer fish as effectively at the damaged sites. Measuring this harm in monetary terms is a central element of legal action against the polluter. Often the distributional impacts of a policy are important to assess, such as how the burden of a tax is borne by different population groups.

Under the logit assumptions, the consumer surplus associated with a set of alternatives takes a closed form that is easy to calculate. By definition, a person's consumer surplus is the utility, in dollar terms, that the person receives in the choice situation. The decision-maker chooses the alternative that provides the greatest utility. Consumer surplus is therefore $CS_n = (1/\alpha_n) max_j(U_{nj} \ \forall j)$, where α_n is the marginal utility of income: $dU_n/dY_n = \alpha_n$ with Y_n being the income of person n. The division by α_n translates utility into dollars since $1/\alpha_n = dY_n/dU_n$. The researcher does not observe U_{nj} and therefore cannot use this expression to calculate the decision-maker's consumer surplus. Instead, the researcher observes V_{nj} and knows the distribution of the remaining portion of utility. With this information, the researcher is able to calculate the expected consumer surplus:

$$E(CS_n) = \frac{1}{\alpha_n} E[max(V_{nj} + \varepsilon_{nj} \ \forall j)],$$

where the expectation is over all possible values of the ε_{nj}'s. Williams (1977) and Small and Rosen (1981) show that, if each ε_{nj} is iid extreme

value and utility is linear in income (such that α_n is constant with respect to income), then this expectation becomes:

$$E(CS_n) = \frac{1}{\alpha_n} ln(\sum_{j=1}^{J} e^{V_{nj}}) + C. \tag{3.10}$$

where C is an unknown constant that represents the fact that the absolute level of utility cannot be measured. As we see below, this constant is irrelevant from a policy perspective and can be ignored.

Note that the term in parentheses in this expression is the denominator of the logit choice probability (3.6). Aside from the division and addition of constants, expected consumer surplus in a logit model is simply the log of the denominator of the choice probability. It is often called "the log-sum term." This relation between the two formulas has no economic meaning, in the sense that there is nothing about a denominator in a choice probability that makes it necessarily related to consumer surplus. It is simply the outcome of the mathematical form of the extreme value distribution. However, the relation makes calculation of expected consumer surplus very easy, which is another of the many conveniences of logit.

Under the standard interpretation for the distribution of errors, as described in the last paragraph of 2.3, $E(CS_n)$ is the average consumer surplus in the subpopulation of people who have the same representative utilities as person n. Total consumer surplus in the population is calculated as the weighted sum of $E(CS_n)$ over a sample of decision-makers, with the weights reflecting the number of people in the population who face the same representative utilities as the sampled person.

The change in consumer surplus that results from a change in the alternatives and/or the choice set is calculated from (3.10). In particular, $E(CS_n)$ is calculated twice: first under the conditions before the change and again under the conditions after the change. The difference between the two results is the change in consumer surplus:

$$\Delta E(CS_n) = \frac{1}{\alpha_n} [ln(\sum_{j=1}^{J^1} e^{V^1_{nj}}) - ln(\sum_{j=1}^{J^0} e^{V^0_{nj}})],$$

where the superscripts 0 and 1 refer to before and after the change. The number of alternatives can change (e.g., a new alternative added) as well as the attributes of the alternatives. Since the unknown constant C enters expected consumer surplus both before and after the