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* (IIA) property discussed in Section 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 converse: 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 independently, identically distributed extreme value. The distribution 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 relevance to 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

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

This formula is sometimes used in describing binary logit models, that is, 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 distribution 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 attains 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 her 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) respecify representative utility so that the source of the correlation is captured explicitly and thus 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 effect when estimating average preferences than when forecasting substitution patterns. These issues are discussed in subsequent sections.

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

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

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.

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 researcher 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 such that this small improvement doesn't help much. Similarly, if one alternative

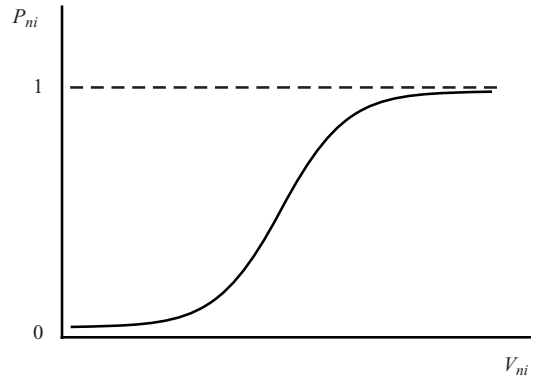

Figure 3.1. Graph of logit curve.

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 an electric heating system. Suppose that the utility the household obtains from each type of system depends only on the purchase price, the 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: $U_g = \beta_1 \mathrm{PP}_g + \beta_2 \mathrm{OC}_g + \varepsilon_g$ and $U_e = \beta_1 \mathrm{PP}_e + \beta_2 \mathrm{OC}_e + \varepsilon_e$, 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

$$P_g = \frac{e^{\beta_1 \mathrm{PP}_g + \beta_2 \mathrm{OC}_g}}{e^{\beta_1 \mathrm{PP}_g + \beta_2 \mathrm{OC}_g} + e^{\beta_1 \mathrm{PP}_e + \beta_2 \mathrm{OC}_e}} \tag{3.7}$$

and the probability of electric heating is the same but with $\exp(\beta_1 \mathrm{PP}_e + \beta_2 \mathrm{OC}_e)$ 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: $d\mathrm{U} = \beta_1\, d\mathrm{PP} + \beta_2\, d\mathrm{OC} = 0$. We then solve for the change in purchase price that keeps utility constant (i.e., satisfies this equation) for a change in operating costs: $\partial \mathrm{PP}/\partial \mathrm{OC} = -\beta_2/\beta_1$. 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

$$P_g = \frac{1}{1 + e^{(\beta_1 \mathrm{PP}_e + \beta_2 \mathrm{OC}_e) - (\beta_1 \mathrm{PP}_g + \beta_2 \mathrm{OC}_g)}}.$$

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: $U_o = \beta_1 \text{PP}_o + \beta_2 \text{OC}_o + \varepsilon_o$. With this extra option available, the probability that the household chooses a gas system is

$$P_g = \frac{e^{\beta_1 \text{PP}_g + \beta_2 \text{OC}_g}}{e^{\beta_1 \text{PP}_g + \beta_2 \text{OC}_g} + e^{\beta_1 \text{PP}_e + \beta_2 \text{OC}_e} + e^{\beta_1 \text{PP}_o + \beta_2 \text{OC}_o}},$$

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 $\sigma^2 \times (\pi^2/6)$. 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 $\pi^2/6$: $\text{Var}(\varepsilon_{nj}) = \text{Var}(\varepsilon^*_{nj}/\sigma) = (1/\sigma^2)$ $\text{Var}(\varepsilon^*_{nj}) = (1/\sigma^2) \times \sigma^2 \times (\pi^2/6) = \pi^2/6$. 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 the 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 effect of each observed variable *relative to* the variance of the unobserved factors. A larger variance in unobserved factors leads to smaller coefficients, even if the observed factors have the same effect 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 in 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

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

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 depend 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

$$U_{nj} = \alpha_n \mathrm{SR}_j + \beta_n \mathrm{PP}_j + \varepsilon_{nj}, \tag{3.8}$$

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

$$U_{nj} = \rho(M_n \mathrm{SR}_j) + \theta(\mathrm{PP}_j / I_n) + \varepsilon_{nj}.$$

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 $U_{nj} = \rho(M_n \mathrm{SR}_j) + \phi(M_n^2 \mathrm{SR}_j) + \theta(\mathrm{PP}_j / I_n) + \varepsilon_{nj}$, which results in a logit model with three variables entering the 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

$$U_{nj} = \rho(M_n \mathrm{SR}_j) + \mu_n \mathrm{SR}_j + \theta(\mathrm{PP}_j / I_n) + \eta_n \mathrm{PP}_j + \varepsilon_{nj}.$$

Since μ_n and η_n are not observed, the terms $\mu_n \mathrm{SR}_j$ and $\eta_n \mathrm{PP}_j$ become part of the unobserved component of utility,

$$U_{nj} = \rho(M_n \mathrm{SR}_j) + \theta(\mathrm{PP}_j / I_n) + \tilde{\varepsilon}_{nj},$$

where $\tilde{\varepsilon}_{nj} = \mu_n \mathrm{SR}_j + \eta_n \mathrm{PP}_j + \varepsilon_{nj}$. 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: $\mathrm{Cov}(\tilde{\varepsilon}_{nj}, \tilde{\varepsilon}_{nk}) = \mathrm{Var}(\mu_n)\mathrm{SR}_j \mathrm{SR}_k + \mathrm{Var}(\eta_n)\mathrm{PP}_j \mathrm{PP}_k \neq 0$ 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: $\mathrm{Var}(\tilde{\varepsilon}_{nj}) = \mathrm{Var}(\mu_n)\mathrm{SR}_j^2 + \mathrm{Var}(\eta_n)\mathrm{PP}_j^2 + \mathrm{Var}(\varepsilon_{nj})$, 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 latter than from 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 in the following discussion.

The Property of Independence from Irrelevant Alternatives

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

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

This ratio does not depend on any alternatives other than i and k. That is, the relative odds of choosing i over k are 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.

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: $P_c = P_{bb} = \frac{1}{2}$, 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, so 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 $P_c = P_{bb} = P_{rb} = \frac{1}{3}$, 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 $P_c = \frac{1}{2}$ and $P_{bb} = P_{rb} = \frac{1}{4}$. 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 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 of another alternative. For example, suppose a new transit mode is added that is similar to, 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, so 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 nonlinearly, 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 percentage. 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 percentage.

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 percentage. The probability for large gas car would drop by ten percent, from .66 to .60, and that 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 overpredict the gas savings that result from the subsidy, since it overpredicts the substitution away from large gas cars (the "gas guzzlers") and underpredicts 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 IIA reflects reality (or an adequate approximation to reality), considerable advantages are gained by its employment. First, because of the IIA, 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, motorbiking, skateboarding, and so on. 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 *et al.* (1978). Two types of tests are suggested. First, the model can be reestimated 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 reestimated 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 *et al.* (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

is 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 researcher 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

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

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; for example, $V_{njt} = \beta' x_{njt}$, where x_{njt} is a vector of variables describing alternative j as faced by n in period t, then there is essentially no difference between the logit model with panel data and 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 0 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 the 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 there to be 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 respecify 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 Nonlinear Representative Utility

In some contexts, the researcher will find it useful to allow parameters to enter representative utility nonlinearly. Estimation is then more difficult, since the log-likelihood function may 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 may include parameters that are interpretable only when they enter utility nonlinearly. In these cases, the effort of writing one's own code can be warranted. Two examples illustrate this point.

Example 1: The Goods–Leisure Tradeoff

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 in the following, the parameters of the workers' utility function over goods and leisure enter nonlinearly 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

$$U = (1 - \beta) \ln G + \beta \ln L.$$

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 constraint 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 are 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

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

The cost of travel is divided by w^{β}, and the 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 nonlinearly. Since this parameter has meaning, the researcher might want to estimate it within this nonlinear 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. The 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 nonlinearity, 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. The population and employment in the combined zone are necessarily the sums of those 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 equality 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 effects 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 $\mathrm{CS}_n = (1/\alpha_n)\max_j(U_{nj})$, where α_n is the marginal utility of income: $dU_n/dY_n = \alpha_n$, with Y_n 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:

$$E(\mathrm{CS}_n) = \frac{1}{\alpha_n} E[\max_j(V_{nj} + \varepsilon_{nj})],$$

where the expectation is over all possible values of ε_{nj}. Williams (1977) and Small and Rosen (1981) show that, if each ε_{nj} is iid extreme value and utility is linear in income (so that α_n is constant with respect to income), then this expectation becomes

$$E(\mathrm{CS}_n) = \frac{1}{\alpha_n} \ln\left(\sum_{j=1}^{J} e^{V_{nj}}\right) + C, \tag{3.10}$$

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

Note that the argument 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 resemblance 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 Section 2.3, $E(\text{CS}_n)$ is the average consumer surplus in the subpopulation of people who have the same representative utilities as person n. The total consumer surplus in the population is calculated as the weighted sum of $E(\text{CS}_n)$ over a sample of decision makers, with the weights reflecting the numbers 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, $E(\text{CS}_n)$ 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:

$$\Delta E(\text{CS}_n) = \frac{1}{\alpha_n}\left[\ln\left(\sum_{j=1}^{J^1} e^{V_{nj}^1}\right) - \ln\left(\sum_{j=1}^{J^0} e^{V_{nj}^0}\right)\right],$$

where the superscripts 0 and 1 refer to before and after the change. The number of alternatives can change (e.g., a new alternative can be added) as well as the attributes of the alternatives. Since the unknown constant C enters expected consumer surplus both before and after the change, it drops out of the difference and can therefore be ignored when calculating changes in consumer surplus.

To calculate the change in consumer surplus, the researcher must know or have estimated the marginal utility of income, α_n. Usually a price or cost variable enters the representative utility, in which case the negative of its coefficient is α_n by definition. (A price or cost coefficient is negative; the negative of a negative coefficient gives a positive α_n.) For example, in the choice between car and bus, utility is $U_{nj} = \beta_1 t_{nj} + \beta_2 c_{nj}$, where t is time, c is cost, and both β_1 and β_2 are negative, indicating that utility decreases as the time or cost for a trip increases. The negative of the cost coefficient, $-\beta_2$, is the amount that utility rises due to a

one-dollar decrease in costs. A one-dollar reduction in costs is equivalent to a one-dollar increase in income, since the person gets to spend the dollar that he saves in travel costs just the same as if he got the extra dollar in income. The amount $-\beta_2$ is therefore the increase in utility from a one-dollar increase in income: the marginal utility of income. It is the same amount in this case for all n. If c_{nj} entered the representative utility interacting with characteristics of the person other than income, as in the product $c_{nj}H_n$, where H_n is household size, then the marginal utility of income would be $-\beta_2 H_n$, which varies over n.

Throughout this discussion, α_n has been assumed to be fixed for a given person independent of his income. The formula (3.10) for expected consumer surplus depends critically on the assumption that the marginal utility of income is independent from income. If the marginal utility of income changes with income, then a more complicated formula is needed, since α_n itself becomes a function of the change in attributes. McFadden (1999) and Karlstrom (2000) provide procedures for calculating changes in consumer surplus under these conditions.

The conditions for using expression (3.10) are actually less severe than stated. Since only changes in consumer surplus are relevant for policy analysis, formula (3.10) can be used if the marginal utility of income is constant over the range of implicit income changes that are considered by the policy. Thus, for policy changes that change consumer surplus by small amounts per person relative to income, the formula can be used even though the marginal utility of income in reality varies with income.

The assumption that α_n does not depend on income has implications for the specification of representative utility. As already discussed, α_n is usually taken as the absolute value of the coefficient of price or cost. Therefore, if the researcher plans to use her model to estimate changes in consumer surplus and wants to apply formula (3.10), this coefficient cannot be specified to depend on income. In the mode choice example, cost can be multiplied by household size, so that the cost coefficient, and hence the marginal utility of income, varies over households of different size. However, if the cost is divided by the household's income, then the coefficient of cost depends on income, violating the assumption needed for expression (3.10). This violation may not be important for small changes in consumer surplus, but certainly becomes important for large changes.

3.6 Derivatives and Elasticities

Since choice probabilities are a function of observed variables, it is often useful to know the extent to which these probabilities change in response to a change in some observed factor. For example, in a

household's choice of make and model of car to buy, a natural question is: to what extent will the probability of choosing a given car increase if the vehicle's fuel efficiency is improved? From competing manufacturers' points of view, a related question is: to what extent will the probability of households' choosing, say, a Toyota decrease if the fuel efficiency of a Honda improves?

To address these questions, derivatives of the choice probabilities are calculated. The change in the probability that decision maker n chooses alternative i given a change in an observed factor, z_{ni}, entering the representative utility of that alternative (and holding the representative utility of other alternatives constant) is

$$\begin{aligned}\frac{\partial P_{ni}}{\partial z_{ni}} &= \frac{\partial\left(e^{V_{ni}}/\sum_j e^{V_{nj}}\right)}{\partial z_{ni}} \\ &= \frac{e^{V_{ni}}}{\sum e^{V_{nj}}}\frac{\partial V_{ni}}{\partial z_{ni}} - \frac{e^{V_{ni}}}{\left(\sum e^{V_{nj}}\right)^2}e^{V_{ni}}\frac{\partial V_{ni}}{\partial z_{ni}} \\ &= \frac{\partial V_{ni}}{\partial z_{ni}}\left(P_{ni} - P_{ni}^2\right) \\ &= \frac{\partial V_{ni}}{\partial z_{ni}}P_{ni}(1 - P_{ni}).\end{aligned}$$

If representative utility is linear in z_{ni} with coefficient β_z, the derivative becomes $\beta_z P_{ni}(1 - P_{ni})$. This derivative is largest when $P_{ni} = 1 - P_{ni}$, which occurs when $P_{ni} = .5$. It becomes smaller as P_{ni} approaches zero or one. The sigmoid probability curve in Figure 3.1 is consistent with these facts. Stated intuitively, the effect of a change in an observed variable is largest when the choice probabilities indicate a high degree of uncertainty regarding the choice. As the choice becomes more certain (i.e., the probabilities approach zero or one), the effect of a change in an observed variable lessens.

One can also determine the extent to which the probability of choosing a particular alternative changes when an observed variable relating to *another* alternative changes. Let z_{nj} denote an attribute of alternative j. How does the probability of choosing alternative i change as z_{nj} increases? We have

$$\begin{aligned}\frac{\partial P_{ni}}{\partial z_{nj}} &= \frac{\partial\left(e^{V_{ni}}/\sum_k e^{V_{nk}}\right)}{\partial z_{nj}} \\ &= -\frac{e^{V_{ni}}}{\left(\sum e^{V_{nk}}\right)^2}e^{V_{nj}}\frac{\partial V_{nj}}{\partial z_{nj}} \\ &= -\frac{\partial V_{nj}}{\partial z_{nj}}P_{ni}P_{nj}.\end{aligned}$$

When V_{nj} is linear in z_{nj} with coefficient β_z, then this cross-derivative becomes $-\beta_z P_{ni} P_{nj}$. If z_{nj} is a desirable attribute, so that β_z is positive, then raising z_{nj} decreases the probability of choosing each alternative other than j. Furthermore, the decrease in probability is proportional to the value of the probability before z_{nj} was changed.

A logically necessary aspect of derivatives of choice probabilities is that, when an observed variable changes, the changes in the choice probabilities sum to zero. This is a consequence of the fact that the probabilities must sum to one before and after the change; it is demonstrated for logit models as follows:

$$\begin{aligned}
\sum_{i=1}^{J} \frac{\partial P_{ni}}{\partial z_{nj}} &= \frac{\partial V_{nj}}{\partial z_{nj}} P_{nj}(1 - P_{nj}) + \sum_{i \neq j} \left(-\frac{\partial V_{nj}}{\partial z_{nj}} \right) P_{nj} P_{ni} \\
&= \frac{\partial V_{nj}}{\partial z_{nj}} P_{nj} \left[(1 - P_{nj}) - \sum_{i \neq j} P_{ni} \right] \\
&= \frac{\partial V_{nj}}{\partial z_{nj}} P_{nj} [(1 - P_{nj}) - (1 - P_{nj})] \\
&= 0.
\end{aligned}$$

In practical terms, if one alternative is improved so that the probability of its being chosen increases, the additional probability is necessarily drawn from other alternatives. To increase the probability of one alternative necessitates decreasing the probability of another alternative. While obvious, this fact is often forgotten by planners who want to improve demand for one alternative without reducing demand for other alternatives.

Economists often measure response by elasticities rather than derivatives, since elasticities are normalized for the variables' units. An elasticity is the percentage change in one variable that is associated with a one-percent change in another variable. The elasticity of P_{ni} with respect to z_{ni}, a variable entering the utility of alternative i, is

$$\begin{aligned}
E_{iz_{ni}} &= \frac{\partial P_{ni}}{\partial z_{ni}} \frac{z_{ni}}{P_{ni}} \\
&= \frac{\partial V_{ni}}{\partial z_{ni}} P_{ni}(1 - P_{ni}) \frac{z_{ni}}{P_{ni}} \\
&= \frac{\partial V_{ni}}{\partial z_{ni}} z_{ni}(1 - P_{ni}).
\end{aligned}$$

If representative utility is linear in z_{ni} with coefficient β_z, then $E_{iz_{ni}} = \beta_z z_{ni}(1 - P_{ni})$.

The cross-elasticity of P_{ni} with respect to a variable entering alternative j is

$$E_{iz_{nj}} = \frac{\partial P_{ni}}{\partial z_{nj}} \frac{z_{nj}}{P_{ni}}$$
$$= -\frac{\partial V_{nj}}{\partial z_{nj}} z_{nj} P_{nj},$$

which in the case of linear utility reduces to $E_{iz_{nj}} = -\beta_z z_{nj} P_{nj}$. As discussed in Section 3.3.2, this cross-elasticity is the same for all i: a change in an attribute of alternative j changes the probabilities for all other alternatives by the same percent. This property of the logit cross-elasticities is a manifestation, or restatement, of the IIA property of the logit choice probabilities.

3.7 Estimation

Manski and McFadden (1981) and Cosslett (1981) describe estimation methods under a variety of sampling procedures. We discuss in this section estimation under the most prominent of these sampling schemes. We first describe estimation when the sample is exogenous and all alternatives are used in estimation. We then discuss estimation on a subset of alternatives and with certain types of choice-based (i.e., nonexogenous) samples.

3.7.1. Exogenous Sample

Consider first the situation in which the sample is exogenously drawn, that is, is either random or stratified random with the strata defined on factors that are exogenous to the choice being analyzed. If the sampling procedure is related to the choice being analyzed (for example, if mode choice is being examined and the sample is drawn by selecting people on buses and pooling them with people selected at toll booths), then more complex estimation procedures are generally required, as discussed in the next section. We also assume that the explanatory variables are exogenous to the choice situation. That is, the variables entering representative utility are independent of the unobserved component of utility.

A sample of N decision makers is obtained for the purpose of estimation. Since the logit probabilities take a closed form, the traditional maximum-likelihood procedures can be applied. The probability of person n choosing the alternative that he was actually observed to choose

can be expressed as

$$\prod_i (P_{ni})^{y_{ni}},$$

where $y_{ni} = 1$ if person n chose i and zero otherwise. Note that since $y_{ni} = 0$ for all nonchosen alternatives and P_{ni} raised to the power of zero is 1, this term is simply the probability of the chosen alternative.

Assuming that each decision maker's choice is independent of that of other decision makers, the probability of each person in the sample choosing the alternative that he was observed actually to choose is

$$L(\beta) = \prod_{n=1}^{N} \prod_i (P_{ni})^{y_{ni}},$$

where β is a vector containing the parameters of the model. The log-likelihood function is then

$$\text{LL}(\beta) = \sum_{n=1}^{N} \sum_i y_{ni} \ln P_{ni} \tag{3.11}$$

and the estimator is the value of β that maximizes this function. McFadden (1974) shows that $\text{LL}(\beta)$ is globally concave for linear-in-parameters utility, and many statistical packages are available for estimation of these models. When parameters enter the representative utility nonlinearly, the researcher may need to write her own estimation code using the procedures described in Chapter 8.

Maximum likelihood estimation in this situation can be reexpressed and reinterpreted in a way that assists in understanding the nature of the estimates. At the maximum of the likelihood function, its derivative with respect to each of the parameters is zero:

$$\frac{d\text{LL}(\beta)}{d\beta} = 0. \tag{3.12}$$

The maximum likelihood estimates are therefore the values of β that satisfy this first-order condition. For convenience, let the representative utility be linear in parameters: $V_{nj} = \beta' x_{nj}$. This specification is not required, but makes the notation and discussion more succinct. Using (3.11) and the formula for the logit probabilities, we show at the end of this subsection that the first-order condition (3.12) becomes

$$\sum_n \sum_i (y_{ni} - P_{ni}) x_{ni} = 0. \tag{3.13}$$

Rearranging and dividing both sides by N, we have

$$(3.14) \quad \frac{1}{N}\sum_n \sum_i y_{ni}x_{ni} = \frac{1}{N}\sum_n \sum_i P_{ni}x_{ni}.$$

This expression is readily interpretable. Let $\bar{x}$ denote the average of x over the alternatives chosen by the sampled individuals: $\bar{x} = (1/N)\sum_n \sum_i y_{ni}x_{ni}$. Let $\hat{x}$ be the average of x over the predicted choices of the sampled decision makers: $\hat{x} = (1/N)\sum_n \sum_i P_{ni}x_{ni}$. The observed average of x in the sample is $\bar{x}$, while $\hat{x}$ is the predicted average. By (3.14), these two averages are equal at the maximum likelihood estimates. That is, the maximum likelihood estimates of β are those that make the predicted average of each explanatory variable equal to the observed average in the sample. In this sense, the estimates induce the model to reproduce the observed averages in the sample.

This property of the maximum likelihood estimator for logit models takes on a special meaning for the alternative-specific constants. An alternative-specific constant is the coefficient of a dummy variable that identifies an alternative. A dummy for alternative j is a variable whose value in the representative utility of alternative i is $d_i^j = 1$ for $i = j$ and zero otherwise. By (3.14), the estimated constant is the one that gives

$$\frac{1}{N}\sum_n \sum_i y_{ni}d_i^j = \frac{1}{N}\sum_n \sum_i P_{ni}d_i^j,$$

$$S_j = \hat{S}_j,$$

where S_j is the share of people in the sample who chose alternative j, and $\hat{S}_j$ is the predicted share for alternative j. With alternative-specific constants, the predicted shares for the sample equal the observed shares. The estimated model is therefore correct on average within the sample. This feature is similar to the function of a constant in a linear regression model, where the constant assures that the average of the predicted value of the dependent variable equals its observed average in the sample.

The first-order condition (3.13) provides yet another important interpretation. The difference between a person's actual choice, y_{ni}, and the probability of that choice, P_{ni}, is a modeling error, or residual. The left-hand side of (3.13) is the sample covariance of the residuals with the explanatory variables. The maximum likelihood estimates are therefore the values of the β's that make this covariance zero, that is, make the residuals uncorrelated with the explanatory variables. This condition for logit estimates is the same as applies in linear regression models. For a regression model $y_n = \beta' x_n + \varepsilon_n$, the ordinary least squares estimates are the values of β that set $\sum_n (y_n - \beta' x_n)x_n = 0$. This fact is verified by solving for β: $\beta = (\sum_n x_n x_n')^{-1}(\sum_n x_n y_n)$, which is the formula

for the ordinary least squares estimator. Since $y_n - \beta' x_n$ is the residual in the regression model, the estimates make the residuals uncorrelated with the explanatory variables.

Under this interpretation, the estimates can be motivated as providing a sample analog to population characteristics. We have assumed that the explanatory variables are exogenous, meaning that they are uncorrelated in the population with the model errors. Since the variables and errors are uncorrelated in the population, it makes sense to choose estimates that make the variables and residuals uncorrelated in the sample. The estimates do exactly that: they provide a model that reproduces in the sample the zero covariances that occur in the population.

Estimators that solve equations of the form (3.13) are called method-of-moments estimators, since they use moment conditions (correlations in this case) between residuals and variables to define the estimator. We will return to these estimators when discussing simulation-assisted estimation in Chapter 10.

We asserted without proof that (3.13) is the first-order condition for the maximum likelihood estimator of the logit model. We give that proof now. The log-likelihood function (3.11) can be reexpressed as

$$\begin{aligned}
\mathrm{LL}(\beta) &= \sum_n \sum_i y_{ni} \ln P_{ni} \\
&= \sum_n \sum_i y_{ni} \ln \left(\frac{e^{\beta' x_{ni}}}{\sum_j e^{\beta' x_{nj}}} \right) \\
&= \sum_n \sum_i y_{ni} (\beta' x_{ni}) - \sum_n \sum_i y_{ni} \ln \Big(\sum_j e^{\beta' x_{nj}} \Big).
\end{aligned}$$

The derivative of the log-likelihood function then becomes

$$\begin{aligned}
\frac{d\mathrm{LL}(\beta)}{d\beta} &= \frac{\sum_n \sum_i y_{ni} (\beta' x_{ni})}{d\beta} - \frac{\sum_n \sum_i y_{ni} \ln \big(\sum_j e^{\beta' x_{nj}} \big)}{d\beta} \\
&= \sum_n \sum_i y_{ni} x_{ni} - \sum_n \sum_i y_{ni} \sum_j P_{nj} x_{nj} \\
&= \sum_n \sum_i y_{ni} x_{ni} - \sum_n \Big(\sum_j P_{nj} x_{nj} \Big) \sum_i y_{ni} \\
&= \sum_n \sum_i y_{ni} x_{ni} - \sum_n \Big(\sum_j P_{nj} x_{nj} \Big) \\
&= \sum_n \sum_i (y_{ni} - P_{ni}) x_{ni}.
\end{aligned}$$

Setting this derivative to zero gives the first-order condition (3.13).

Estimation on a Subset of Alternatives

In some situations, the number of alternatives facing the decision maker is so large that estimating model parameters is very expensive or even impossible. With a logit model, estimation can be performed on a subset of alternatives without inducing inconsistency. For example, a researcher examining a choice situation that involves 100 alternatives can estimate on a subset of 10 alternatives for each sampled decision maker, with the person's chosen alternative included as well as 9 alternatives randomly selected from the remaining 99. If all alternatives have the same chance of being selected into the subset, then estimation proceeds on the subset of alternatives as if it were the full set. If alternatives have unequal probability of being selected, more complicated estimation procedures may be required. The procedure is described as follows.

Suppose that the researcher has used some specific method for randomly selecting alternatives into the subset that is used in estimation for each sampled decision maker. Denote the full set of alternatives as F and a subset of alternatives as K. Let $q(K \mid i)$ be the probability under the researcher's selection method that subset K is selected given that the decision maker chose alternative i. Assuming that the subset necessarily includes the chosen alternative, we have $q(K \mid i) = 0$ for any K that does not include i. The probability that person n chooses alternative i from the full set is P_{ni}. Our goal is to derive a formula for the probability that the person chooses alternative i *conditional* on the researcher selecting subset K for him. This conditional probability is denoted $P_n(i \mid K)$.

This conditional probability is derived as follows. The joint probability that the researcher selects subset K and the decision maker chooses alternative i is $\text{Prob}(K, i) = q(K \mid i)P_{ni}$. The joint probability can also be expressed with the opposite conditioning as $\text{Prob}(K, i) = P_n(i \mid K)Q(K)$ where $Q(K) = \sum_{j \in F} P_{nj} q(K \mid j)$ is the probability of the researcher selecting subset K marginal over all the alternatives that the person could choose. Equating these two expressions and solving for $P_n(i \mid K)$, we have

$$
\begin{aligned}
P_n(i \mid K) &= \frac{P_{ni} q(K \mid i)}{\sum_{j \in F} P_{nj} q(K \mid j)} \\
&= \frac{e^{V_{ni}} q(K \mid i)}{\sum_{j \in F} e^{V_{nj}} q(K \mid j)} \\
&= \frac{e^{V_{ni}} q(K \mid i)}{\sum_{j \in K} e^{V_{nj}} q(K \mid j)}, \qquad (3.15)
\end{aligned}
$$

where the second line has canceled out the denominators of P_{ni} and

$P_{nj}\ \forall j$, and the third equality uses the fact that $q(K \mid j) = 0$ for any j not in K.

Suppose that the researcher has designed the selection procedure so that $q(K \mid j)$ is the same for all $j \in K$. This property occurs if, for example, the researcher assigns an equal probability of selection to all nonchosen alternatives, so that the probability of selecting j into the subset when i is chosen by the decision maker is the same as for selecting i into the subset when j is chosen. McFadden (1978) calls this the "uniform conditioning property," since the subset of alternatives has a uniform (equal) probability of being selected conditional on any of its members being chosen by the decision maker. When this property is satisfied, $q(K \mid j)$ cancels out of the preceding expression, and the probability becomes

$$P_n(i \mid K) = \frac{e^{V_{ni}}}{\sum_{j \in K} e^{V_{nj}}},$$

which is simply the logit formula for a person who faces the alternatives in subset K.

The conditional log-likelihood function under the uniform conditioning property is

$$\text{CLL}(\beta) = \sum_n \sum_{i \in K_n} y_{ni} \ln \frac{e^{V_{ni}}}{\sum_{j \in K_n} e^{V_{nj}}},$$

where K_n is the subset selected for person n. This function is the same as the log-likelihood function given in (3.11) except that the subset of alternatives K_n replaces, for each sampled person, the complete set. Maximization of CLL provides a consistent estimator of β. However, since information is excluded from CLL that LL incorporates (i.e., information on alternatives not in each subset), the estimator based on CLL is not efficient.

Suppose that the researcher designs a selection process that does not exhibit the uniform conditioning property. In this case, the probability $q(K \mid i)$ can be incorporated into the model as a separate variable. The expression in (3.15) can be rewritten as

$$P_n(i \mid K) = \frac{e^{V_{ni} + \ln q(K \mid i)}}{\sum_{j \in K} e^{V_{nj} + \ln q(K \mid j)}}.$$

A variable z_{nj} calculated as $\ln q(K_n \mid j)$ is added to the representative utility of each alternative. The coefficient of this variable is constrained to 1 in estimation.

The question arises: why would a researcher ever want to design a selection procedure that does not satisfy the uniform conditioning

property, since satisfying the property makes estimation so straightforward? An illustration of the potential benefit of nonuniform conditioning is provided by Train *et al.* (1987a) in their study of telecommunications demand. The choice situation in their application included an enormous number of alternatives representing portfolios of calls by time of day, distance, and duration. The vast majority of alternatives were hardly ever chosen by anyone in the population. If alternatives had been selected with equal probability for each alternative, it was quite likely that the resulting subsets would consist nearly entirely of alternatives that were hardly ever chosen, coupled with the person's chosen alternative. Comparing a person's chosen alternative with a group of highly undesirable alternatives provides little information about the reasons for a person's choice. To avoid this problem, alternatives were selected in proportion to the shares for the alternatives in the population (or, to be precise, estimates of the population shares). This procedure increased the chance that relatively desirable alternatives would be in each subset of alternatives that was used in estimation.

3.7.2. Choice-Based Samples

In some situations, a sample drawn on the basis of exogenous factors would include few people who have chosen particular alternatives. For example, in the choice of water heaters, a random sample of households in most areas would include only a small number who had chosen solar water-heating systems. If the researcher is particularly interested in factors that affect the penetration of solar devices, a random sample would need to be very large to assure a reasonable number of households with solar heat.

In situations such as these, the researcher might instead select the sample, or part of the sample, on the basis of the choice being analyzed. For example, the researcher examining water heaters might supplement a random sample of households with households that are known (perhaps through sales records at stores if the researcher has access to these records) to have recently installed solar water heaters.

Samples selected on the basis of decision makers' choices can be purely choice-based or a hybrid of choice-based and exogenous. In a purely choice-based sample, the population is divided into those that choose each alternative, and decision makers are drawn randomly within each group, though at different rates. For example, a researcher who is examining the choice of home location and is interested in identifying the factors that contribute to people choosing one particular community might draw randomly from within that community at the rate of one out

of L households, and draw randomly from all other communities at a rate of one out of M, where M is larger than L. This procedure assures that the researcher has an adequate number of people in the sample from the area of interest. A hybrid sample is like the one drawn by the researcher interested in solar water heating, in which an exogenous sample is supplemented with a sample drawn on the basis of the households' choices.

Estimation of model parameters with samples drawn at least partially on the basis of the decision maker's choice is fairly complex in general, and varies with the exact form of the sampling procedure. For interested readers, Ben-Akiva and Lerman (1985, pp. 234–244) provide a useful discussion. One result is particularly significant, since it allows researchers to estimate logit models on choice-based samples without becoming involved in complex estimation procedures. This result, due to Manski and Lerman (1977), can be stated as follows. If the researcher is using a *purely* choice-based sample and includes an alternative-specific constant in the representative utility for each alternative, then estimating a logit model as if the sample were exogenous produces consistent estimates for all the model parameters except the alternative-specific constants. Furthermore, these constants are biased by a known factor and can therefore be adjusted so that the adjusted constants are consistent. In particular, the expectation of the estimated constant for alternative j, labeled $\hat{\alpha}_j$, is related to the true constant α_j^* by

$$E(\hat{\alpha}_j) = \alpha_j^* - \ln(A_j/S_j),$$

where A_j is the share of decision makers in the population who chose alternative j, and S_j is the share in the choice-based sample who chose alternative j. Consequently, if A_j is known (that is, if population shares are known for each alternative), then a consistent estimate of the alternative-specific constant is the constant $\hat{\alpha}_j$ that is estimated on the choice-based sample *plus* the log of the ratio of the population share to the sample share.

3.8 Goodness of Fit and Hypothesis Testing

We discuss goodness of fit and hypothesis testing in the context of logit models, where the log-likelihood function is calculated exactly. The concepts apply to other models, with appropriate adjustment for simulation variance, when the log-likelihood function is simulated rather than calculated exactly.

3.8.1. Goodness of Fit

A statistic called the *likelihood ratio index* is often used with discrete choice models to measure how well the models fit the data. Stated more precisely, the statistic measures how well the model, with its estimated parameters, performs compared with a model in which all the parameters are zero (which is usually equivalent to having no model at all). This comparison is made on the basis of the log-likelihood function, evaluated at both the estimated parameters and at zero for all parameters.

The likelihood ratio index is defined as

$$\rho = 1 - \frac{\text{LL}(\hat{\beta})}{\text{LL}(0)},$$

where $\text{LL}(\hat{\beta})$ is the value of the log-likelihood function at the estimated parameters and LL(0) is its value when all the parameters are set equal to zero. If the estimated parameters do no better, in terms of the likelihood function, than zero parameters (that is, if the estimated model is no better than no model), then $\text{LL}(\hat{\beta}) = \text{LL}(0)$ and so $\rho = 0$. This is the lowest value that ρ can take (since if $\text{LL}(\hat{\beta})$ were less than LL(0), then $\hat{\beta}$ would not be the maximum likelihood estimate).

At the other extreme, suppose the estimated model was so good that each sampled decision maker's choice could be predicted perfectly. In this case, the likelihood function at the estimated parameters would be one, since the probability of observing the choices that were actually made is one. And, since the log of one is zero, the log-likelihood function would be zero at the estimated parameters. With $\text{LL}(\hat{\beta}) = 0$, $\rho = 1$. This is the highest value that ρ can take. In summary, the likelihood ratio index ranges from zero, when the estimated parameters are no better than zero parameters, to one, when the estimated parameters perfectly predict the choices of the sampled decision makers.

It is important to note that the likelihood ratio index is not at all similar in its interpretation to the R^2 used in regression, despite both statistics having the same range. R^2 indicates the percentage of the variation in the dependent variable that is "explained" by the estimated model. The likelihood ratio has no intuitively interpretable meaning for values between the extremes of zero and one. It is the percentage increase in the log-likelihood function above the value taken at zero parameters (since $\rho = 1 - \text{LL}(\hat{\beta})/\text{LL}(0) = (\text{LL}(0) - \text{LL}(\hat{\beta}))/\text{LL}(0)$). However, the meaning of such a percentage increase is not clear. In comparing two models estimated on the same data and with the same set

of alternatives (such that LL(0) is the same for both models), it is usually valid to say that the model with the higher ρ fits the data better. But this is saying no more than that increasing the value of the log-likelihood function is preferable. Two models estimated on samples that are not identical or with a different set of alternatives for any sampled decision maker cannot be compared via their likelihood ratio index values.

Another goodness-of-fit statistic that is sometimes used, but should actually be avoided, is the "percent correctly predicted." This statistic is calculated by identifying for each sampled decision maker the alternative with the highest probability, based on the estimated model, and determining whether or not this was the alternative that the decision maker actually chose. The percentage of sampled decision makers for which the highest-probability alternative and the chosen alternative are the same is called the percent correctly predicted.

This statistic incorporates a notion that is opposed to the meaning of probabilities and the purpose of specifying choice probabilities. The statistic is based on the idea that the decision maker is predicted by the researcher to choose the alternative for which the model gives the highest probability. However, as discussed in the derivation of choice probabilities in Chapter 2, the researcher does not have enough information to predict the decision maker's choice. The researcher has only enough information to state the probability that the decision maker will choose each alternative. In stating choice probabilities, the researcher is saying that if the choice situation were repeated numerous times (or faced by numerous people with the same attributes), each alternative would be chosen a certain proportion of the time. This is quite different from saying that the alternative with the highest probability will be chosen each time.

An example may be useful. Suppose an estimated model predicts choice probabilities of .75 and .25 in a two-alternative situation. Those probabilities mean that if 100 people faced the representative utilities that gave these probabilities (or one person faced these representative utilities 100 times), the researcher's best prediction of how many people would choose each alternative are 75 and 25. However, the "percent correctly predicted" statistic is based on the notion that the best prediction for each person is the alternative with the highest probability. This notion would predict that one alternative would be chosen by all 100 people while the other alternative would never be chosen. The procedure misses the point of probabilities, gives obviously inaccurate market shares, and seems to imply that the researcher has perfect information.

3.8.2. Hypothesis Testing

As with regressions, standard t-statistics are used to test hypotheses about individual parameters in discrete choice models, such as whether the parameter is zero. For more complex hypotheses, a likelihood ratio test can nearly always be used, as follows. Consider a null hypothesis H that can be expressed as constraints on the values of the parameters. Two of the most common hypotheses are (1) several parameters are zero, and (2) two or more parameters are equal. The constrained maximum likelihood estimate of the parameters (labeled $\hat{\beta}^H$) is that value of β that gives the highest value of LL without violating the constraints of the null hypothesis H. Define the ratio of likelihoods, $R = L(\hat{\beta}^H)/L(\hat{\beta})$, where $\hat{\beta}^H$ is the (constrained) maximum value of the likelihood function (not logged) under the null hypothesis H, and $\hat{\beta}$ is the unconstrained maximum of the likelihood function. As in likelihood ratio tests for models other than those of discrete choice, the test statistic defined as $-2 \log R$ is distributed chi-squared with degrees of freedom equal to the number of restrictions implied by the null hypothesis. Therefore, the test statistic is $-2(\text{LL}(\hat{\beta}^H) - \text{LL}(\hat{\beta}))$. Since the log likelihood is always negative, this is simply two times the (magnitude of the) difference between the constrained and unconstrained maximums of the log-likelihood function. If this value exceeds the critical value of chi-squared with the appropriate degrees of freedom, then the null hypothesis is rejected.

Null Hypothesis I: The Coefficients of Several Explanatory Variables Are Zero

To test this hypothesis, estimate the model twice: once with these explanatory variables included, and a second time without them (since excluding the variables forces their coefficients to be zero). Observe the maximum value of the log-likelihood function for each estimation; two times the difference in these maximum values is the value of the test statistic. Compare the test statistic with the critical value of chi-squared with degrees of freedom equal to the number of explanatory variables excluded from the second estimation.

Null Hypothesis II: The Coefficients of the First Two Variables Are the Same

To test this hypothesis, estimate the model twice: once with each of the explanatory variables entered separately, including the first two;

then with the first two variables replaced by one variable that is the sum of the two variables (since adding the variables forces their coefficients to be equal). Observe the maximum value of the log-likelihood function for each of the estimations. Multiply the difference in these maximum values by two, and compare this figure with the critical value of chi-squared with one degree of freedom.

3.9 Case Study: Forecasting for a New Transit System

One of the earliest applications of logit models, and a prominent test of their capabilities, arose in the mid-1970s in the San Francisco Bay area. A new rail system, called the Bay Area Rapid Transit (BART), had been built. Daniel McFadden obtained a grant from the National Science Foundation to apply logit models to commuters' mode choices in the Bay area and to use the models to predict BART ridership. I was lucky enough to serve as his research assistant on this project. A sample of commuters was taken before BART was open for service. Mode choice models were estimated on this sample. These estimates provided important information on the factors that enter commuters' decisions, including their value of time savings. The models were then used to forecast the choices that the sampled commuters would make once BART became available. After BART had opened, the commuters were recontacted and their mode choices were observed. The predicted share taking BART was compared with the observed share. The models predicted quite well, far more accurately than the procedures used by the BART consultants, who had not used discrete choice models.

The project team collected data on 771 commuters before BART was opened. Four modes were considered to be available for the trip to work: (1) driving a car by oneself, (2) taking the bus and walking to the bus stop, (3) taking the bus and driving to the bus stop, and (4) carpooling. The time and cost of travel on each mode were determined for each commuter, based on the location of the person's home and work. Travel time was differentiated as walk time (for the bus–walk mode), wait time (for both bus modes), and on-vehicle time (for all the modes). Characteristics of the commuter were also collected, including income, household size, number of cars and drivers in the household, and whether the commuter was the head of the household. A logit model with linear-in-parameters utility was estimated on these data.

The estimated model is shown in Table 3.1, which is reproduced from Train (1978). The cost of travel was divided by the commuter's wage to reflect the expectation that workers with lower wages are more

Table 3.1. *Logit model of work trip mode choice*

Explanatory Variable[a]		Coefficient	*t*-Statistic
Cost divided by post-tax wage, minutes (1–4)		−0.0284	4.31
Auto on-vehicle time, minutes (1, 3, 4)		−0.0644	5.65
Transit on-vehicle time, minutes (2, 3)		−0.0259	2.94
Walk time, minutes (2, 3)		−0.0689	5.28
Transfer wait time, minutes (2, 3)		−0.0538	2.30
Number of transfers (2, 3)		−0.1050	0.78
Headway of first bus, minutes (2, 3)		−0.0318	3.18
Family income with ceiling $7500 (1)		0.00000454	0.05
Family income – $7500 with floor 0, ceiling $3000 (1)		−0.0000572	0.43
Family income – $10,500 with floor 0, ceiling $5000 (1)		−0.0000543	0.91
Number of drivers in household (1)		1.02	4.81
Number of drivers in household (3)		0.990	3.29
Number of drivers in household (4)		0.872	4.25
Dummy if worker is head of household (1)		0.627	3.37
Employment density at work location (1)		−0.0016	2.27
Home location in or near central business district (1)		−0.502	4.18
Autos per driver with ceiling one (1)		5.00	9.65
Autos per driver with ceiling one (3)		2.33	2.74
Autos per driver with ceiling one (4)		2.38	5.28
Auto alone dummy (1)		−5.26	5.93
Bus with auto access dummy (3)		−5.49	5.33
Carpool dummy (4)		−3.84	6.36
Likelihood ratio index	0.4426		
Log likelihood at convergence	−595.8		
Number of observations	771		
Value of time saved as a percentage of wage:			
Auto on-vehicle time	227		3.20
Transit on-vehicle time	91		2.43
Walk time	243		3.10
Transfer wait time	190		2.01

[a] Variable enters modes in parentheses and is zero in other modes. Modes: 1. Auto alone. 2. Bus with walk access. 3. Bus with auto access. 4. Carpool.

concerned about cost than higher-paid workers. On-vehicle time enters separately for car and bus travel to indicate that commuters might find time spent on the bus to be more, or less, bothersome than time spent driving in a car. Bus travel often involves transfers, and these transfers can be onerous for travelers. The model therefore includes the number of transfers and the expected wait time at the transfers. The headway (i.e., the time between scheduled buses) for the first bus line that the

commuter would take is included as a measure of the maximum amount of time that the person would need to wait for this bus.

The estimated coefficients of cost and the various time components provide information on the value of time. By definition, the value of time is the extra cost that a person would be willing to incur to save time. The utility takes the form $U_{nj} = \alpha c_{nj}/w_n + \beta t_{nj} + \ldots$, where c is cost and t is time. The total derivative with respect to changes in time and cost is $dU_{nj} = (\alpha/w_n)\, dc_{nj} + \beta\, dt_{nj}$, which we set equal to zero and solve for dc/dt to find the change in cost that keeps utility unchanged for a change in time: $dc/dt = -(\beta/\alpha)w_n$. The value of time is therefore a proportion β/α of the person's wage. The estimated values of time are reported at the bottom of Table 3.1. The time saved from riding on the bus is valued at 91 percent of wage $((-.0259/-.0284) \times 100)$, while the time saved from driving in a car is worth more than twice as much: 227 percent of wage. This difference suggests that commuters consider driving to be considerably more onerous than riding the bus, when evaluated on a per-minute basis. Commuters apparently choose cars not because they like driving *per se* but because driving is usually quicker. Walking is considered more bothersome than waiting for a bus (243 percent of wage versus 190 percent), and waiting for a bus is more bothersome than riding the bus.

Income enters the representative utility of the auto-alone alternative. It enters in a piecewise linear fashion to allow for the possibility that additional income has a different impact depending on the overall level of income. None of the income variables enters significantly. Apparently dividing travel cost by wage picks up whatever effect income might have on the mode choice of a commuter. That is, higher wages induce the commuter to be less concerned about travel costs but do not induce a predilection for driving beyond the impact through cost. The number of people and the number of vehicles per driver in the household have a significant effect on mode choice, as expected. Alternative-specific constants are included, with the constant for the bus–walk alternative normalized to zero.

The model in Table 3.1 was used to predict the mode choices of the commuters after BART was open for service. The choice set was considered to be the four modes listed previously plus two BART modes, differentiated by whether the person takes the bus or drives to the BART station. Table 3.2 presents the forecasted and actual shares for each mode. BART demand was forecast to be 6.3 percent, compared with an actual share of 6.2 percent. This close correspondence is remarkable.

The figures in Table 3.2 tend to mask several complications that arose in the forecasting. For example, walking to the BART station was

Table 3.2. *Predictions for after BART opened*

	Actual Share	Predicted Share
Auto alone	59.90	55.84
Bus with walk access	10.78	12.51
Bus with auto access	1.426	2.411
BART with bus access	0.951	1.053
BART with auto access	5.230	5.286
Carpool	21.71	22.89

originally included as a separate mode. The model forecasted this option very poorly, overpredicting the number of people who would walk to BART by a factor of twelve. The problem was investigated and found to be primarily due to differences between the experience of walking to BART stations and that of walking to the bus, given the neighborhoods in which the BART stations are located. These issues are discussed at greater length by McFadden *et al.* (1977).

3.10 Derivation of Logit Probabilities

It was stated without proof in Section 3.1 that if the unobserved component of utility is distributed iid extreme value for each alternative, then the choice probabilities take the form of equation (3.6). We now derive this result. From (3.5) we have

$$P_{ni} = \int_{s=-\infty}^{\infty} \left(\prod_{j \neq i} e^{-e^{-(s+V_{ni}-V_{nj})}} \right) e^{-s} e^{-e^{-s}} ds,$$

where s is ε_{ni}. Our task is to evaluate this integral. Noting that $V_{ni} - V_{ni} = 0$ and then collecting terms in the exponent of e, we have

$$\begin{aligned} P_{ni} &= \int_{s=-\infty}^{\infty} \left(\prod_{j} e^{-e^{-(s+V_{ni}-V_{nj})}} \right) e^{-s} ds \\ &= \int_{s=-\infty}^{\infty} \exp\left(-\sum_{j} e^{-(s+V_{ni}-V_{nj})} \right) e^{-s} ds \\ &= \int_{s=-\infty}^{\infty} \exp\left(-e^{-s} \sum_{j} e^{-(V_{ni}-V_{nj})} \right) e^{-s} ds. \end{aligned}$$

Define $t = \exp(-s)$ such that $-\exp(-s)\,ds = dt$. Note that as s approaches infinity, t approaches zero, and as s approaches negative

infinity, t becomes infinitely large. Using this new term,

$$\begin{aligned} P_{ni} &= \int_{\infty}^{0} \exp\left(-t \sum_j e^{-(V_{ni}-V_{nj})}\right)(-dt) \\ &= \int_{0}^{\infty} \exp\left(-t \sum_j e^{-(V_{ni}-V_{nj})}\right) dt \\ &= \left. \frac{\exp\left(-t \sum_j e^{-(V_{ni}-V_{nj})}\right)}{-\sum_j e^{-(V_{ni}-V_{nj})}} \right|_{0}^{\infty} \\ &= \frac{1}{\sum_j e^{-(V_{ni}-V_{nj})}} = \frac{e^{V_{ni}}}{\sum_j e^{V_{nj}}}, \end{aligned}$$

as required.