2 Logit

2.1 Function Form of Choice Probabilities

By far the most widely used qualitative choice model is logit. Its popularity is due to the fact that the formula for logit choice probabilities is readily interpretable, particularly compared with other qualitative choice models, and the parameters of logit models are relatively inexpensive to estimate.

Following the discussion of section 1.3, the logit probabilities are derived under a particular assumption regarding the distribution of the unobserved portion of utility. The basic notation of section 1.3 will be repeated for convenience, followed by the specification of the logit probabilities.

Suppose a decisionmaker, denoted n, faces a set of J_n alternatives. The utility that the decisionmaker obtains from alternative i in J_n, denoted U_{in}, is decomposed into (1) a part that is known by the researcher, labeled as V_{in}, and (2) an unknown part that is assumed to be a random variable, labeled e_{in}. This is expressed as $U_{in} = V_{in} + e_{in}$. Recall that the known part of utility V_{in} is a function that depends on the observed characteristics of the alternative as faced by the decisionmaker (labeled z_{in}), the observed characteristics of the decisionmaker (s_n), and a vector of parameters (β) that are either known a priori by the researcher or estimated: $V_{in} = V(z_{in}, s_n, \beta)$. For notational simplicity this functional dependence is suppressed; however, it is important to remember that V_{in} depends on observed data and known or estimated parameters.

Assume that each e_{in}, for all i in J_n, is distributed independently, identically in accordance with the extreme value distribution.[1] Given this distribution for the unobserved components of utility, the probability that the decisionmaker will choose alternative i is

$$P_{in} = \frac{e^{V_{in}}}{\sum_{j \in J_n} e^{V_{jn}}}, \qquad \text{for all} \quad i \text{ in } J_n. \tag{2.1}$$

The proof of this fact, while straightforward, is tedious and not particularly illuminating; for readers who are interested it is given, along with the formula for the extreme value distribution, at the end of this chapter (see section 2.9).

Since the unobserved component of utility is assumed, through the extreme value distribution, to have zero mean, the observed part of utility, V_{in}, is often called representative, expected, or average utility. It should be clear in using these terms, however, that the expectation or average is over all possible values of factors unobserved by the researcher rather than by the decisionmaker.

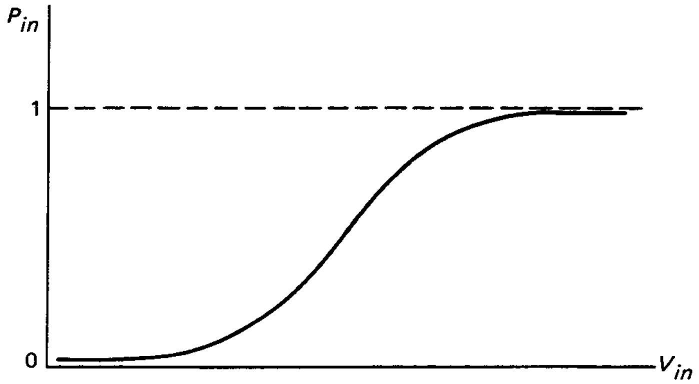

Figure 2.1
Graph of logit curve.

Three properties of the choice probabilities are important to note. First, each of the choice probabilities is necessarily between zero and one. If alternative *i* were as unpleasant as possible in the eyes of the decisionmaker, and hence its representative utility approached negative infinity, then P_{in} would approach zero. On the other hand, if alternative *i* were as wonderful as possible in the eyes of the decisionmaker, and hence its representative utility approached infinity, then P_{in} would approach one (given finite values for the representative utilities of the other alternatives).

Second, the choice probabilities necessarily sum to one:

$$\sum_{i \in J_n} P_{in} = \sum_{i \in J_n} \left(\frac{e^{V_{in}}}{\sum_{j \in J_n} e^{V_{jn}}} \right) = 1.$$

This follows from the fact that the choice set in a qualitative choice situation is exhaustive, so that the decisionmaker must choose one of the alternatives, and the alternatives are mutually exclusive, so that the decisionmaker cannot choose more than one alternative.

Third, the relation of the choice probability for an alternative to the representative utility of that alternative, holding the representative utilities of the other alternatives fixed, is sigmoid, or S-shaped (figure 2.1).

If the representative utility of one alternative is very low, compared with other alternatives, a small increase in the utility of this alternative will not much affect the probability of its being chosen; the other alternatives will still be generally preferred. Similarly, if one alternative is far superior to the others, so that its representative utility is very high, an additional increase in its utility will not much affect the probability of its being chosen; it will usually be chosen even without the extra utility. The point at which an increase in the representative utility of an alternative has the greatest effect on its probability of being chosen is when its representative utility is very similar to that of other alternatives. In this case, a small increase in the utility of one alternative could, in a way, "tip the balance," and thereby induce a large increase in the probability of the alternative being chosen.

Examples Consider first a binary choice situation: a household's choice between a gas or electric oven. Suppose that the utility the household obtains from each type of oven depends only on the oven's purchase price, operating cost, and the household's view of the relative accuracy or ease of cooking with the oven. The first two of these factors are observed by the researcher, but the researcher cannot observe the third factor. If the researcher considers the observed part of utility to be a linear function of the observed factors, then the utility of each type of oven can be written as $U_g = \beta_1 \mathrm{PP}_g + \beta_2 \mathrm{OC}_g + e_g$ and $U_r = \beta_1 \mathrm{PP}_r + \beta_2 \mathrm{OC}_r + e_r$, where the subscript g denotes gas and r denotes electric; PP and OC are the purchase price and operating cost, respectively, of the oven type denoted by the subscript; β_1 and β_2 are scalar parameters; and the subscript n denoting household is suppressed for convenience. The utility of a household is higher the less it has to pay for an oven, in either purchasing or operating it, since the household can purchaser other goods with the money saved. Therefore, β_1 and β_2 are negative.

The unobserved component of utility for each alternative, e_g and e_r, varies over household with respect to households' differing views of the accuracy and ease of cooking by each type of oven. If this component has an extreme value distribution, then the probability that the household will choose a gas oven is

$$P_g = \frac{e^{\beta_1 \mathrm{PP}_g + \beta_2 \mathrm{OC}_g}}{e^{\beta_1 \mathrm{PP}_g + \beta_2 \mathrm{OC}_g} + e^{\beta_1 \mathrm{PP}_r + \beta_2 \mathrm{OC}_r}}, \tag{2.2}$$

and the probability that it will choose an electric one is expressed analogously. Note that, since β_1 and β_2 are negative, the probability of choosing a gas oven decreases as the cost of purchasing or operating it increases (if the costs of an electric oven are constant). Furthermore, the probability of choosing a gas oven increases as the purchase price or operating cost of an electric oven increases.

A multinomial case is a simple extension of the binary case. Consider, for example, that the household could choose a microwave oven instead of either a gas or electric one (ignore, for the sake of this example, the possibility of owning a microwave in addition to a convection oven). Under analogous assumptions to those given, the utility of each type of oven is

$$U_g = \beta_1 \, \mathrm{PP}_g + \beta_2 \, \mathrm{OC}_g + e_g;$$

$$U_r = \beta_1 \, \mathrm{PP}_r + \beta_2 \, \mathrm{OC}_r + e_r;$$

$$U_m = \beta_1 \, \mathrm{PP}_m + \beta_2 \, \mathrm{OC}_m + e_m;$$

where the subscript m denotes microwave. Under the assumption that e_g, e_r, and e_m are each distributed independently extreme value, the probability that the household chooses a gas oven is

$$P_g = \frac{e^{\beta_1 \, \mathrm{PP}_g + \beta_2 \, \mathrm{OC}_g}}{e^{\beta_1 \, \mathrm{PP}_g + \beta_2 \, \mathrm{OC}_g} + e^{\beta_1 \, \mathrm{PP}_r + \beta_2 \, \mathrm{OC}_r} + e^{\beta_1 \, \mathrm{PP}_m + \beta_2 \, \mathrm{OC}_m}}. \tag{2.3}$$

The probabilities for electric and microwave ovens are analogous.

Expression (2.3) has the same numerator in the binary case of expression (2.2), but the denominator is larger by the quantity $\exp(\beta_1 \, \mathrm{PP}_m + \beta_2 \, \mathrm{OC}_m)$. Therefore, as one would expect in the real world, the probability of choosing a gas oven is lower when the possibility of buying a microwave oven is available than when it is not.

2.2 The Independence from Irrelevant Alternatives Property

Three properties of logit probabilities are discussed in section 2.1, namely, that they (1) range from zero to one, (2) sum to one over alternatives, and (3) are a sigmoid or S-shaped function of representative utility. Each of these properties is quite reasonable, and in fact the first two are logically necessary. Logit probabilities also exhibit a property, however, that, at least in some contexts, is not desirable. This is called the independence from irrelevant alternatives property, or the IIA property for short.

The IIA property has been the focus of considerable discussion in the literature and not a small amount of confusion. In the next pages, the basics of the IIA property are presented first, followed by a discussion of relatively recent concepts that place the IIA property in clearer perspective.

IIA Basics

Consider the ratio of the choice probabilities for two alternatives, i and k:

$$\frac{P_{in}}{P_{kn}} = \frac{e^{V_{in}}/\sum_{j \in J_n} e^{V_{jn}}}{e^{V_{kn}}/\sum_{j \in J_n} e^{V_{jn}}}$$

$$= \frac{e^{V_{in}}}{e^{V_{kn}}} = e^{V_{in}-V_{kn}}.$$

Note that the ratio of these two probabilities does not depend on any alternatives other than i and k. That is, the ratios of probabilities is necessarily the same no matter what other alternatives are in J_n or what the characteristics of other alternatives are. Since the ratio is independent from alternatives other than i and k, it is said to be independent from "irrelevant" alternatives, that is, alternatives other than those for which the ratio is calculated.

While this property is an accurate reflection of reality in some choice situations, it is clearly inappropriate in other situations. Consider, for example, the classic red bus/blue bus problem. Suppose there is a traveler who has a choice of going by auto or taking a blue bus and that both alternatives have the same representative utility. Because the representative utilities are equal, the choice probabilities are equal ($P_{\mathrm{a}} = 1/2 = P_{\mathrm{bb}}$, where a denotes auto and bb denotes blue bus) and the ratio of probabilities is one ($P_{\mathrm{a}}/P_{\mathrm{bb}} = 1$).

Now suppose that a red bus were introduced and that the traveler considered the red bus to be exactly like the blue bus. Consequently, the ratio of probabilities for taking the two differently colored buses is one ($P_{\mathrm{bb}}/P_{\mathrm{rb}} = 1$, where rb denotes red bus). However, since in the logit model the ratio $P_{\mathrm{a}}/P_{\mathrm{bb}}$ is the same independent of the existence of other alternatives, this ratio remains constant at one. The only probabilities for which $P_{\mathrm{a}}/P_{\mathrm{bb}} = 1$ and $P_{\mathrm{bb}}/P_{\mathrm{rb}} = 1$ are $P_{\mathrm{a}} = P_{\mathrm{bb}} = P_{\mathrm{rb}} = 1/3$, which are the probabilities that the logit model predicts.

In real life, however, we would expect the probability of taking an auto to remain the same when a new bus is introduced that is essentially the same as the old bus. We would also expect the original probability of taking bus to be split, after the introduction of the new bus, between the two buses. That is, we would expect $P_{\mathrm{a}} = 1/2$ and $P_{\mathrm{bb}} = P_{\mathrm{rb}} = 1/4$. In this case, the logit model, because of its IIA property, overestimates the probability of taking either of the buses and underestimates the probability of taking an auto.

In cases like that of the red bus/blue bus, the IIA property of logit models is inappropriate. However, in situations in which the IIA property reflects reality, considerable advantages are gained by its employment. First, because of the IIA property, it is possible to estimate model parameters consistently on a subset of alternatives for each sampled decisionmaker. For example, in a situation with 100 alternatives, the researcher might (so as to reduce computer costs) 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 exclusion of alternatives not in the subset, exclusion of alternatives in estimation does not affect the consistency of the estimation. (Details of this type of estimation and its consistency are given in section 2.6.)

This fact has considerable practical importance. In analyzing choice situations for which the number of alternatives is large, estimating on a subset of alternatives can save substantial amounts of computer time and expense. At the extreme, the number of alternatives might be so large as to preclude estimation altogether (due to core capacity of computers) if it were not possible to utilize a subset of alternatives.

Another practical use of this ability to estimate on subsets of alternatives arises when a 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 identifying the factors that contribute to a worker's choice of taking an auto or a bus to work. The full set of alternative modes includes walking, bicycling, etc., in addition to auto and bus. However, the researcher, if he believed the IIA property to be appropriate in this case, could estimate a model with only the alternatives of bus and auto included for each sampled person, thereby saving considerable time and money. Sampled workers who did not choose either auto or bus would be excluded from the sample (since their chosen alternatives are not in the estimation subset) and the model would be estimated on the remaining sampled workers.

The IIA property also allows the researcher to predict demand for alternatives that do not currently exist, such as the demand for a new make of car, a new mode of travel, a new product, and so on. Consider, for example, a researcher examining households' choices of make and model of auto. If the researcher thinks that the IIA property is appropriate in this setting, he can estimate a model describing the choice of make and model of auto using currently available makes and models in the estimation, and then use the estimated model to calculate the probability that a household would choose a make and model that will be introduced shortly.

The appropriateness of this procedure is conceptually related to the consistency of estimation on a subset of alternatives. If the full set of alternatives is considered to be all the currently available makes and models plus the soon-to-be-introduced make and model, then estimation on currently available makes and models is equivalent to estimating on a subset of alternatives, which, as discussed, provides consistent estimates of the model parameters.

IIA Revisited

Despite its practical advantages, the IIA property is a restriction that is not realistic in many situations. Recent work has indicated, however, that the IIA property in logit models is not as restrictive as it might at first seem, or, in particular, as indicated by the red bus/blue bus problem.

McFadden (1975) has shown that any model that specifies choice probabilities, including models that do not exhibit IIA, can be expressed in the **form** of logit models. That is, it is possible to express any choice probability as

$$P_{in} = \frac{e^{W_{in}}}{\sum_{j \in J_n} e^{W_{jn}}},$$

where W_{jn}, for all j in J_n, is some function of observed data.

The proof is simple. Let $P^*_{in} = f(z_{in};\ z_{jn}, \text{ for all } j \neq i;\ s_n)$ be the "true" model, where z_{in} is observed data relating to alternative i as faced by decisionmaker n, and s_n is a vector of characteristics of the decisionmaker. Note that this specification is completely general; in particular, choice probabilities that do not exhibit IIA are allowed. Taking logs,

$$\log P^*_{in} = \log f(z_{in};\ z_{jn}, \text{for all } j \neq i;\ s_n).$$

Now, define $W_{in} = \log P^*_{in}$ and evaluate logit probabilities based on W_{in}:

$$P_{in} = \frac{e^{W_{in}}}{\sum_{j \in J_n} e^{W_{jn}}} = \frac{e^{\log P^*_{in}}}{\sum_{j \in J_n} e^{\log P^*_{jn}}} = \frac{P^*_{in}}{\sum_{j \in J_n} P^*_{jn}} = P^*_{in},$$

where the last equality is due to the fact that choice probabilities necessarily sum to one. This shows that logit probabilities, with the appropriate specification of W_{in}, equal the true probabilities. Stated another way, any choice model can, with an appropriate choice of W_{in}, be put into the logit form. This concept has given rise to the term "mother logit."

The logit model derived from the extreme value distribution, which exhibits the IIA property, is a special case of "mother logit." The term W_{in} in the mother logit model depends in general on all observed data including characteristics of alternatives other than i. However, V_{in} in equation (2.1), which can be called the standard logit model, depends only on characteristics of alternative i and of the decisionmaker; characteristics of alternatives other than i do not enter V_{in}. Therefore, when W_{in} depends only on characteristics of the decisionmaker and alternative i, mother logit becomes standard logit and exhibits IIA; otherwise, the mother logit model need not exhibit IIA.

What this discussion implies is that the logit specification can be used in situations for which IIA does not hold. All that is required is that additional variables be added to representative utility, in particular, variables that relate to alternatives other than the one for which the representative utility is designated.

An example of how this can be done, that is, of how adding terms to representative utility within the logit specification can enable the model to represent situations in which IIA does not hold, is provided by reexamining the red bus/blue bus problem. The representative utility of auto, red bus, and blue bus is assumed to be the same:

$$V_{\mathrm{a}} = V_{\mathrm{bb}} = V_{\mathrm{rb}}.$$

As discussed, the standard logit model gives equal probabilities for all three alternatives, while we know that the true probabilities are .5 for auto and .25 each for blue bus and red bus. However, if the term ln(1/2) is added to the representative utility of the two bus alternatives, then the logit model gives the true probabilities. The probability of auto is

$$\begin{aligned}P_{\mathrm{a}} &= \frac{e^{V_{\mathrm{a}}}}{e^{V_{\mathrm{bb}}+\ln(1/2)} + e^{V_{\mathrm{rb}}+\ln(1/2)} + e^{V_{\mathrm{a}}}}\\ &= \frac{e^{V_{\mathrm{a}}}}{(e^{V_{\mathrm{bb}}})(1/2) + (e^{V_{\mathrm{rb}}})(1/2) + e^{V_{\mathrm{a}}}}\\ &= \frac{e^{V_{\mathrm{a}}}}{2e^{V_{\mathrm{a}}}}\\ &= \frac{1}{2},\end{aligned}$$

where the next to last equality is due to the fact that $V_{\mathrm{bb}} = V_{\mathrm{rb}} = V_{\mathrm{a}}$. It can be similarly shown that $P_{bb} = P_{rb} = .25$. In summary, if appropriate terms are added to representative utility in the logit model, the red bus/blue bus "problem" is not a problem at all.

The difficulty, in general, is knowing what terms to add to representative utility to account for true probabilities not exhibiting IIA. In some cases, however, the researcher need not know the adjustment factor a priori, since it can be estimated. In the red bus/blue bus case, for example, the researcher need not know that ln(1/2) should be added to the representative utilities of the bus alternatives. Suppose the researcher estimates the model with all three alternatives in the choice set and includes a constant term in the specification of the representative utility of the bus alternatives; that is, suppose the researcher specified the representative utility of each alternative as

$$V_a^* = V_a;$$

$$V_{bb}^* = \alpha + V_{bb};$$

$$V_{rb}^* = \alpha + V_{rb}.$$

The estimation procedure would automatically estimate a value of α equal to ln(1/2). (This is due to the fact, explained in section 2.6, that the estimated value of a constant in the representative utility of each alternative is that at which the average estimated probability for each alternative exactly equals the share of sampled decisionmakers who actually chose that alternative. If the true shares for auto, blue bus, and red bus are .5, .25, and .25, respectively, and $V_a = V_{bb} = V_{rb}$, then the only value of α that would cause the estimated probabilities to equal these shares is ln(1/2).)

Using the logit model when the true probabilities do not exhibit IIA is not as problematic, therefore, as it at first appeared. There are three contexts, however, in which the problem still arises. First, in a situation like the red bus/blue bus case, if the researcher is estimating a model with all alternatives (three in the red bus/blue bus case) and does not include a constant in the representative utility of each alternative, then the estimation cannot incorporate the needed adjustment term. This implies that, whenever possible, the researcher should include constants in the representative utility of each alternative. Second, in a situation like the red bus/blue bus case, if the researcher estimates the model on a subset of alternatives (e.g., auto and blue bus) and then forecasts for a third alternative (e.g., red bus), then the estimated probability for the new alternative will not represent the true probability. This is because the representative utility of the new alternative will not incorporate the necessary adjustment. (If the researcher somehow knows the required adjustment factor, then he can apply it and calculate consistent probabilities for the new alternatives.) Third, if the situation is not like a red bus/blue bus case and an adjustment other than to the constant in representative utility is required to enable the logit specification to represent the true probabilities, then, unless the researcher can determine the necessary adjustments a priori, the estimated logit model will not represent the true probabilities.

2.3 Specification of Representative Utility

We turn now to several issues regarding the specification of representative utility. Since representative utility is usually assumed to be linear in parameters, this assumption is maintained through most of the section; nonlinear-in-parameters representative utility is discussed at the end of this section.

A linear-in-parameters representative utility function is written as

$$V_{in} = \beta w(z_{in}, s_n),$$

where w is a vector-valued function of the observed data and β is a vector of parameters. For notational simplicity the functional relation of the variables w to the observed data z_{in}, s_n can be suppressed by writing $V_{in} = \beta w_{in}$, where $w_{in} = w(z_{in}, s_n)$. The logit choice probabilities therefore become

$$P_{in} = \frac{e^{\beta w_{in}}}{\sum_{j \in J_n} e^{\beta w_{jn}}}, \qquad \text{for all} \quad i \text{ in } J_n.$$

Within this context, issues regarding the specification of representative utility are questions of what variables to enter as elements of w_{in}.

Alternative-Specific Constants

Recall that the utility that decisionmaker n obtains from alternative i in J_n is composed of a part observed by the researcher and a part not observed, $U_{in} = \beta w_{in} + e_{in}$. For a logit model, e_{in} is assumed to be distributed extreme value, which means it has zero mean. It will usually not be the case that the average of all unobserved factors that affect the decisionmaker's utility is zero. Suppose the average of e_{in} is α_i, a scalar parameter unknown to the researcher. Then the representative utility of alternative i can be expanded to include this constant:

$$U_{in} = \beta w_{in} + \alpha_i + e^*_{in},$$

where $e^*_{in} = e_{in} - \alpha_i$ and hence has zero mean. The parameter α_i is then estimated along with the other parameters of the model (i.e., with β) in the manner described in section 2.6. Conceptually, it is similar to the intercept term in a regression, except that in the decomposition of utility the left-hand-side variable, U_{in}, is not observed.

Including an alternative-specific constant for each alternative serves two functions in addition to providing a zero mean for unobserved utility. First, as demonstrated in section 2.6, the estimated values for the alternative-specific constant are those at which the average probability over the estimation sample for each alternative exactly equals the proportion of decisionmakers in the sample that actually chose that alternative. That is, a model estimated with alternative-specific constants will exactly reproduce the observed shares in the estimation sample. Second, for reasons that are discussed in section 2.2, the inclusion of alternative-specific constants can mitigate, and in some cases remove, inaccuracies due to logit's independence of irrelevant alternatives property.

While one speaks of entering an alternative-specific constant for each alternative, in actuality the constant for one alternative is necessarily normalized to zero and so constants are estimated for, at most, one fewer alternative than there are available. This is not a restriction of the model, only a normalization whose motivation is an aspect of the following topic.

Differences in Representative Utility

A fundamental property of logit models is that only differences in representative utility affect the choice probabilities, not their absolute levels. Consider the probability of choosing alternative i. The standard expression for this probability is

$$P_{in} = \frac{e^{\beta w_{in}}}{\sum_{j \in J_n} e^{\beta w_{jn}}}.$$

The probability can equivalently be expressed in terms of the difference between each alternative's representative utility and the representative utility for any alternative in the choice set:

$$P_{in} = \frac{e^{\beta w_{in} - \beta w_{kn}}}{\sum_{j \in J_n} e^{\beta w_{jn} - \beta w_{kn}}},$$

where k is any alternative in J_n, including perhaps i. These two expressions are equal since

$$\frac{e^{\beta w_{in} - \beta w_{kn}}}{\sum_{j \in J_n} e^{\beta w_{jn} - \beta w_{kn}}} = \frac{e^{\beta w_{in}} \cdot e^{-\beta w_{kn}}}{\sum_{j \in J_n} e^{\beta w_{jn}} \cdot e^{-\beta w_{kn}}} = \frac{e^{\beta w_{in}}}{\sum_{j \in J_n} e^{\beta w_{jn}}}.$$

This fact has several implications. First, it allows the logit probabilities in binary choice situations to be expressed in a simplified form. Consider the choice between gas and electric water heaters. The probability of choosing a gas water heater is

$$P_{\mathrm{g}} = \frac{e^{\beta w_{\mathrm{g}}}}{e^{\beta w_{\mathrm{g}}} + e^{\beta w_{\mathrm{r}}}},$$

where the subscripts g and r denote gas and electricity, respectively, and the subscript n denoting decisionmaker is suppressed. This expression can be rewritten as

$$P_{\mathrm{g}} = \frac{1}{1 + e^{\beta w_{\mathrm{r}} - \beta w_{\mathrm{g}}}},$$

which is the form used in most of the binary logit literature.

Second, since only differences in representative utility matter, alternative-specific constants cannot meaningfully be entered in each alternative; as stated, at least one must be normalized to zero. Consider a binary choice situation in which the representative utility of the two alternatives are written as

$$V_{1n} = \beta w_{1n} + \alpha_1;$$

$$V_{2n} = \beta w_{2n} + \alpha_2.$$

The probabilities that result from these representative utilities are exactly the same as those that result from

$$V_{1n} = \beta w_{1n} + \alpha_1^*;$$

$$V_{2n} = \beta w_{2n},$$

in which $\alpha_1^* = \alpha_1 - \alpha_2$. In fact, any pair of alternative-specific constants whose difference is $\alpha_1 - \alpha_2$ is equivalent.

It is impossible to estimate a constant for each alternative in a choice set since the choice probabilities depend only on differences in the constants and an infinite number of combinations of constants have the same differences. By convention, the constant for one alternative is set equal to zero. The constant for each other alternative is then interpreted as the difference between the average impact of unobserved factors for the two alternatives.

Third, since only differences in representative utility are relevant, variables that do not vary over alternatives cannot affect the choice probabilities. For example, consider the choice of make and model of car. Let representative utility for each alternative i in J_n be $V_{in} = \beta_1 \mathrm{PP}_{in} + \beta_2 A_n$, where PP_{in} is the amount that person n must pay to purchase make/model i and A_n is the age of person n. In taking differences across alternatives, $V_{in} - V_{jn}$, the term $\beta_2 A_n$ drops out. The representative utility given is equivalent in terms of the decisionmaker's choices to $V_{in} = \beta_1 \mathrm{PP}_{in}$. Simply adding a constant to the utility of each alternative does not change the decisionmakers choices or, consequently, the choice probabilities.

If the researcher believes that a factor that does not vary over alternatives (e.g., any characteristic of the decisionmaker) affects the decisionmaker's choices, then it must be entered into representative utility in a meaningful fashion. In particular, it must interact with a variable that varies over alternatives.

In the example of the choice of make and model of car, the researcher might think that households with more members are more likely to purchase large cars because they value the extra room more than smaller families. This effect can be captured in the model by (1) defining a dummy variable that is one for large makes and models, then (2) interacting household size with this dummy variable, and finally (3) entering the interaction variable in representative utility:

$$V_{in} = \beta_1 \mathrm{PP}_{in} + \beta_2 M_n D_i,$$

where D_i is one if i is a large car and zero otherwise, and M_n is the number of members in household n. The coefficient β_2 represents a preference for large cars that increases with household size.

Another example is useful. Consider a household's choice of how many cars to own, with the alternatives being 0, 1, or 2. Suppose the only factor affecting this choice that the researcher observes is the number of members in the household, again labeled M_n for household n. Suppose further that the researcher feels that M_n affects the representative utility for each alternative differently, so that

$$V_{0n} = \beta_0 M_n;$$

$$V_{1n} = \beta_1 M_n;$$

$$V_{2n} = \beta_2 M_n;$$

where V_{0n}, V_{1n}, and V_{2n} are the representative utility of owning no, one, and two cars, respectively, and β_0, β_1, and β_2 are scalar parameters.

Recognizing that only differences in representative utility are relevant, two reformulations are necessary. First, one of the parameters β_0, β_1, or β_2 must be normalized to zero for reasons that are analogous to the normalization of one alternative-specific constant. An equivalent, normalized set of representative utilities is

$$V_{0n} = 0;$$

$$V_{1n} = \beta_1^* M_n;$$

$$V_{2n} = \beta_2^* M_n;$$

where $\beta_1^* = (\beta_1 - \beta_0)$ and $\beta_2^* = (\beta_2 - \beta_0)$. Second, even though M_n does not vary over alternatives, it enters with a different coefficient in each alternative. This is equivalent to M_n being interacted with dummy variables for each alternative. That is, an equivalent expression for representative utilities that explicitly recognizes this interaction is

$$V_{in} = \beta_1^* M_n D_i^1 + \beta_2^* M_n D_i^2, \qquad i = 0, 1, 2,$$

where D_i^1 equals one when $i = 1$ and zero otherwise, and D_i^2 equals one when $i = 2$ and zero otherwise. Note that since β_0 is zero by normalization, no variable is included that interacts M_n with a dummy for the alternatives of owning no cars.

The parameters β_1^* and β_2^* reflect the difference in the impact of M_n on representative utility for the alternative of owning one or two cars, respectively, compared with that of owning no cars. If β_1^* is positive, then increasing the number of members in the household increases the probability of owning one car relative to the probability of owning no car. If β_2^* is also positive, increasing M_n also increases the probability of owning two cars over owning none. Whether the probability of owning two cars increases relative to the probability of owning one car depends on whether β_2^* is larger than β_1^*. If β_2^* is greater than β_1^*, increasing the number of household members increases the probability of owning two cars over one and the probability of owning one car over none.

Taste Variation

The value, or importance, that decisionmakers place on each characteristic of the alternatives varies, in general, over decisionmakers. As discussed, for example, the size of a car was presumed to be more important to households with many members than smaller households. Other examples are readily identifiable. Low income households are probably more concerned about the purchase price of a good, relative to its other characteristics, than higher income households; younger decisionmakers might care more about the horsepower of a car than older people (or vice versa); in choosing a neighborhood to live in, households with young children will be more concerned about the accessibility and quality of schools than those without children; and so on. In addition, decisionmakers' tastes vary for reasons that are not observable or identifiable, just because people are different.

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 will demonstrate the distinction.

Consider households' choices among makes and models of cars to buy. Suppose, for simplicity, that the only two characteristics of cars that the researcher observes is the purchase price (PP_i for make/model i) and inches of shoulder room (SR_i).[2] The value that different households place on these two characteristics varies over households, and so total utility can be written as

$$U_{in} = \alpha_n \mathrm{SR}_i + \beta_n \mathrm{PP}_i + e_{in}, \tag{2.4}$$

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 household (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 larger importance on purchase price:

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

Substituting these relations into (2.4) produces

$$U_{in} = \rho(M_n \mathrm{SR}_i) + \theta(\mathrm{PP}_i / I_n) + e_{in}.$$

Under the assumption that each e_{in} is an independently distributed extreme value, a standard logit model obtains with two variables entering representative utility, both of which are interactions of a vehicle characteristic 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 $U_{in} = \rho(M_n \mathrm{SR}_i) + \phi(M_n^2 \mathrm{SR}_i) + \theta(\mathrm{PP}_i / I_n) + e_{in}$, which results in a standard logit model with three variables entering representative utility.

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

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

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

$$\beta_n = \theta(1/I_n) + \eta_n.$$

Substituting into (2.4) produces

$$U_{in} = \rho(M_n \mathrm{SR}_i) + \mu_n \mathrm{SR}_i + \theta(\mathrm{PP}_i / I_n) + \eta_n \mathrm{PP}_i + e_{in}.$$

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

$$U_{in} = \rho(M_n \mathrm{SR}_i) + \theta(\mathrm{PP}_i / I_n) + \tilde{e}_{in},$$

where $\tilde{e}_{in} = \mu_n \mathrm{SR}_i + \eta_n \mathrm{PP}_i + e_{in}$. The new error term $\tilde{e}_{in}$ cannot possibly be distributed independently, identically random as required for the logit formulation. Since μ_n and η_n are constant over alternatives for each decisionmaker, $\tilde{e}_{in}$ is necessarily correlated over alternatives, violating the independence assumption (i.e., $\mathrm{cov}(\tilde{e}_{in}, \tilde{e}_{jn}) \neq 0$ for $j \neq i$). Furthermore, since SR_i and PP_i vary across alternatives, the variance of $\tilde{e}_{in}$ will vary over alternatives, violating the assumption of identically distributed errors (i.e., $\mathrm{Var}(\tilde{e}_{in}) \neq \mathrm{Var}(\tilde{e}_{jn})$ for $j \neq i$).

This example demonstrates the general point that when tastes vary systematically in the population in relation to observed variables, the variation can be incorporated in logit models. However, if taste variation is random, logit is inappropriate. A probit model, discussed in chapter 3, should be used instead.

Utility Theory as a Specification Tool

The researcher decides what variables to enter in representative utility on the basis of a priori information, both formal and informal. The researcher must decide not just which factors affect the choice probabilities, but how to enter them, that is, what types of interaction terms to specify and by what arithmetic operations, if any, to transform the variables (e.g., log, squared terms). It is often difficult to know the implications of various specifications of representative utility and to determine whether and how, for example, one specification is intrinsically different from another. In these situations, utility theory can often be a useful aid for interpreting and motivating specifications. The appropriate application of utility theory is different in each choice situation. However, an example will illustrate the point of how utility theory can aid in specifying variables to enter representative utility.

In logit models of workers' choice of mode (auto, bus, rail, etc.) for commuting, the wage of the worker often enters as an explanatory variable. In some cases (Train, 1980a, for example) the cost of travel is divided by the worker's wage to reflect the presumption that a worker with a high wage is less concerned about cost than a worker with a low wage. In other cases (McFadden, 1974, for example) travel time is multiplied by the worker's wage to reflect the presumption that a worker with a high wage is more concerned with lost time than a worker with a low wage.

Representative utility is assumed in both specifications to be of the form

$$V_{in} = \beta_n t_{in} + \theta_n c_{in},$$

where t_{in} is the time that it would take person n to travel to work by mode i, c_{in} is the cost of travel by mode i for person n, and β_n and θ_n are parameters specific to person n. With this formulation, the value of money **relative** to time is θ_n/β_n.

In the first specification, the parameters are assumed to vary as

$$\beta_n = \beta^A;$$

$$\theta_n = \theta^A/w_n;$$

where w_n is the wage of person n and β^A and θ^A are parameters constant over all people. The relative value of money compared with time therefore becomes $\theta^A/\beta^A w_n$. In the second specification

$$\beta_n = \beta^B w_n;$$

$$\theta_n = \theta^B;$$

so that the value of money relative to time is $\theta^B/\beta^B w_n$.

The relative value of time and money depends on wage in the same way in these two specifications. The question therefore arises: How are the two specifications different, or are they essentially the same? Does it matter which specification the researcher uses?

To address this question, the neoclassical theory of the tradeoff between goods and leisure is used to derive representative utility for workers' mode choice models. It is shown through this derivation that the two specifications have quite different implications regarding the worker's tradeoff between goods and leisure and that the shape of the worker's indifference mapping for goods and leisure determines the manner in which wage should enter representative utility.

Under the standard treatment of the goods/leisure tradeoff, a worker chooses how many hours to work and in doing so determines how many goods he can consume and how much leisure he has. For every extra hour worked, the worker has one less hour of leisure but can purchase more goods with the money earned in that hour. The worker values both goods and leisure and has a utility function that reflects his preferences regarding various combinations of goods and leisure. The worker chooses the amount to work that maximizes his utility subject to the constraints that (1) his leisure time is necessarily the total time available (24 hours per day) minus the amount worked and (2) the value of the goods he consumes is equal to the value of his wage earnings plus any unearned income.

The standard theory is expanded as follows to allow for the worker choosing a mode to work as well as the number of hours to work. Let the utility function be $U = U(G, L)$, where G is goods and L is leisure. Assuming the price index for goods is constant and normalized to one, the worker faces the constraints that

$$G = V + wW - c;$$

$$L = T - W - t;$$

where V is unearned income (i.e., not related to amount worked), W is the number of hours worked, w is the hourly wage rate, c is the cost of travel to work (which takes values c_i for each mode i), T is the total number of hours available, and t is the time required for travel to work (which takes value t_i for each mode).

We can determine the number of hours that the worker would choose to work **conditional** upon a particular mode being used to travel to work, and then examine the choice of mode. Given mode i, the worker chooses the number of hours to work that maximizes U subject to

$$G = V + w \cdot W - c_i; \tag{2.5}$$

$$L = T - W - t_i. \tag{2.6}$$

Substituting the maximizing value of W into U gives the utility that could be obtained given that mode i is chosen, labeled U_i^*. The worker then chooses the mode with the highest U_i^*.

Let us consider two polar cases for the $U(G, L)$. We shall find that the two specifications of representative utility used in mode choice models arise from these two cases.

CASE A: LET $U = \alpha_1 \log G + \alpha_2 L$ With this utility function, the worker will respond to additional unearned income by reducing the number of hours worked and not by consuming additional goods (to be shown as an intermediate result).

Substituting the constraints (2.5) and (2.6) into the utility function, we have

$$U_i = \alpha_1 \log(V + wW - c_i) + \alpha_2(T - W - t_i). \tag{2.7}$$

Maximizing U_i with respect to W,

$$\partial U_i/\partial W = \alpha_1 w/(V + wW - c_i) - \alpha_2 = 0,$$

so that the utility maximizing number of hours to work is

$$W = (\alpha_1/\alpha_2) + (c_i/w) - (V/w). \tag{2.8}$$

Substituting this into (2.5), we know that the utility maximizing amount of goods consumed is

$$G = V + w((\alpha_1/\alpha_2) + (c_i/w) - (V/w)) - c_i = w\alpha_1/\alpha_2.$$

Note that utility maximizing G does not vary with unearned income V, but that utility maximizing W decreases with V, implying that if a worker with the utility function given above were given additional unearned income, he would respond by reducing his work hours (i.e., increasing leisure) and not increasing consumption.

Substituting the utility maximizing W into the utility function (i.e., substituting (2.8) into (2.7)) gives

$$\begin{aligned} U_i^* &= \alpha_1 \log(V + w((\alpha_1/\alpha_2) + (c_i/w) - (V/w)) - c_i) \\ &\quad + \alpha_2(T - (\alpha_1/\alpha_2) - (c_i/w) + (V/w) - t_i) \\ &= \alpha_1 \log(w\alpha_1/\alpha_2) + \alpha_2 T - \alpha_1 + (\alpha_2 V/w) - \alpha_2((c_i/w) + t_i). \end{aligned}$$

In the choice of mode, all terms that do not vary over i drop out (since only difference in utility matter), and so

$$U_i^* = -\alpha_2((c_i/w) + t_i).$$

In this case, the correct specification of representative utility is for cost to be divided by wage and time not to be interacted.

CASE B: LET $U = \alpha_1 G + \alpha_2 \log L$ Using analogous steps to those for case A, we can show that (1) this U implies that the worker would consume all additional unearned income in goods and would not reduce the number of hours worked at all and (2) the maximum utility that the worker can receive conditional upon mode i is

$$U_i^* = \alpha_1 V + \alpha_1 Tw - \alpha_2 - \alpha_1(t_i w + c_i) + \alpha_2 \log((\alpha_2/\alpha_1)w).$$

The representative utility in the choice of mode includes only those terms that vary over modes and therefore takes the form

$$U_i^* = -\alpha_1(t_i w + c_i);$$

that is, time is multiplied by wage and cost is not interacted.

These two cases show that if the researcher believes that a worker would respond to additional unearned income by working fewer hours (i.e., that $U = \alpha_1 \log G + \alpha_2 L$ reflects workers' preferences), then he should enter cost divided by wage. On the other hand, if he feels that workers would purchase additional goods and not reduce work hours (i.e. if $U = \alpha_1 G + \alpha_2 \log L$), then he should enter time multiplied by wage.

Nonlinear-in-Parameters Representative Utility

Thus far in this section, representative utility has been assumed to be linear in parameters. This assumption is maintained in the great majority of applications. Since, under fairly general conditions, any parametric function can be approximated arbitrarily closely by a function that is linear in parameters, the assumption does not necessarily introduce significant errors.

In some situations, however, it is useful to specify representative utility as not being linear in parameters. Estimation is more difficult and computer routines are not as widely available for logit models with nonlinear-in-parameters representative utility. However, the additional accuracy or information obtained might warrant the additional effort and expense.

Such cases arise when the form of representative utility can be determined theoretically and contains parameters that enter nonlinearly but nevertheless are interesting or important to estimate. Two examples will illustrate this point.

Example 1 In the previous example concerning the goods/leisure tradeoff, the conclusion was reached that (1) if the researcher believed that workers would respond to additional unearned income by reducing the amount they worked but not consuming more goods, then representative utility in a mode choice model should be $-\alpha_1((c_i/w) + t_i)$, where c_i and t_i are the time and cost of mode i, respectively, and w is the wage of the worker; however, (2) if the researcher felt that workers would consume any additional income in goods and not work less, then the appropriate representative utility is $-\alpha_1(c_i + wt_i)$.

It is probably the case, however, that neither of these two extreme situations accurately describe workers' behavior. If given additional unearned income, workers would probably consume somewhat more goods and reduce somewhat the number of hours they worked. The same type of analysis as given above for the extreme cases can be used to construct a more realistic in-between case. In particular, suppose, using the same notation, that workers' preferences regarding goods and leisure are represented by a utility function of the form

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

With this utility function, workers respond to additional unearned income by consuming more goods and working less. It can be shown (see Train and McFadden, 1978, for details) that the representative utility entering a mode choice model, given this utility function for goods and leisure, is

$$U_i = -\alpha((c_i/w^{\beta}) + w^{1-\beta}t_i).$$

That is, the cost of travel is divided by w^{β} and travel time is multiplied by $w^{1-\beta}$. This is a generalization of the polar cases described, since (1) as β approaches one, U_i becomes $-\alpha((c_i/w) + t_i)$, and (2) as β approaches zero, U_i becomes $-\alpha(c_i + wt_i)$. Estimating the general form of U_i, even though β enters nonlinearly, is valuable since it is more realistic on theoretical grounds and the estimated value of β provides information on workers' preference mapping for goods and leisure.

Example 2 Logit models have been used to describe urban travelers' choice of destination conditional upon their deciding to take a trip within the metropolitan area in which they live. Generally, the metropolitan area is partitioned into zones, and models are specified for the probability that a person taking a trip within the city will choose to go to a particular zone. Representative utility for each zone depends in these models on the time and cost of travel to the zone plus a variety of variables, such as residential population and retail employment in the zone, that reflect reasons that travelers would want to visit the zone. These latter variables are called "attraction" variables; label them by the vector a_i for zone i. Since it is these attraction variables that give rise to parameters entering nonlinearly, assume for simplicity that representative utility depends only on these variables, so that

$$V_{in} = f(a_i, \beta),$$

where β is a vector of parameters.

The difficulty in specifying representative utility (that is, in determining f) comes in recognizing that, since zonal definitions are largely arbitrary, an accurate model would not be sensitive to different zonal definitions. In particular, if two zones are combined, it would be desirable for the model to predict a probability of choosing the combined zone that is equal to the sum of the probabilities that it predicted of choosing each of the two original zones. This consideration places restrictions on the form that V_{in} can take. Consider zones i and k, which, when combined, are labeled zone c. Since population and employment in the combined zone is the sum of that in the two original zones, we have $a_c = a_i + a_k$. Also, in order for the model to give the same probabilities for choosing these zones before and after merger, V_{in} must be specified such that

$$P_{in} + P_{kn} = P_{cn};$$

$$(e^{V_{in}} + e^{V_{kn}}) \Big/ \Big(e^{V_{in}} + e^{V_{kn}} + \sum_{\substack{j \in J \\ j \neq i,k}} e^{V_{jn}}\Big) = (e^{V_{cn}}) \Big/ \Big(e^{V_{cn}} + \sum_{\substack{j \in J \\ j \neq i,k}} e^{V_{jn}}\Big).$$

This equality holds only if

$$\exp(V_{in}) + \exp(V_{kn}) = \exp(V_{cn}). \tag{2.9}$$

If we let $V_{in} = \ln(\beta a_i)$ for all i, then relation (2.9) holds given that $a_i + a_k = a_c$.

Therefore, to specify a destination choice model that is not sensitive to the definition of zones, representative utility must be specified with parameters inside a log operation. Special computer routines have been written to estimate such parameters.

2.4 Derivatives and Elasticities of Choice Probabilities

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 a competing manufacturers point of view, a related question is to what extent will the probability of choosing a given car decrease if the fuel efficiency of a competing make and model improves.

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

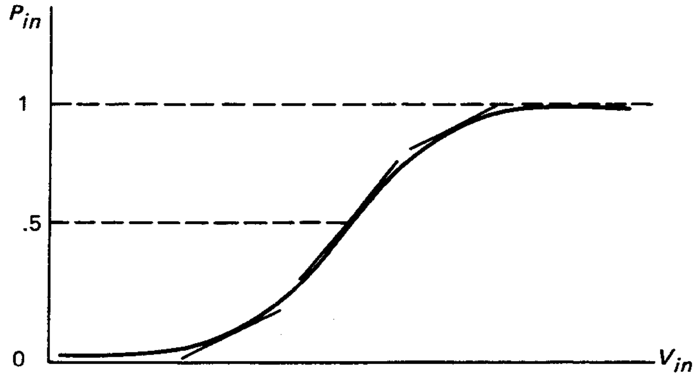

Figure 2.2
Slope of the logit curve.

$$
\begin{aligned}
\frac{\partial P_{in}}{\partial y_{in}} &= \frac{\partial[(e^{V_{in}})(\sum_{j\in J_n} e^{V_{jn}})^{-1}]}{\partial y_{in}}\\
&= \left(\frac{\partial V_{in}}{\partial y_{in}}\right)(e^{V_{in}})(\sum e^{V_{jn}})^{-1} - (e^{V_{in}})(\sum e^{V_{jn}})^{-2}(e^{V_{in}})\left(\frac{\partial V_{in}}{\partial y_{in}}\right)\\
&= \frac{\partial V_{in}}{\partial y_{in}}(P_{in} - P_{in}^2)\\
&= \left(\frac{\partial V_{in}}{\partial y_{in}}\right)P_{in}(1 - P_{in}).
\end{aligned}
$$

Usually V_{in} is linear in the observed variables, with parameters as coefficients. If the coefficient of y_{in} is the scalar β_y, then $\partial V_{in}/\partial y_{in} = \beta_y$ and $\partial P_{in}/\partial y_{in} = \beta_y P_{in}(1 - P_{in})$. This derivative is particularly easy to evaluate. Note that, since β_y is constant, the derivative is largest when $P_{in} = 1 - P_{in}$, which occurs when $P_{in} = 1/2$, and becomes smaller as P_{in} approaches zero or one. This fact is a natural result of the sigmoid shape of the logit function. Consider figure 2.2. The derivative of the choice probability at any level of y_{in} is the slope of the probability curve at that point. This slope is obviously highest at $P_{in} = 1/2$ and becomes lower as P_{in} moves in either direction away from 1/2.

Stated intuitively, the effect of a change in an observed variable is highest 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 given 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 y_{jn} denote an attribute of alternative j (e.g., the fuel efficiency of vehicle j). How does the probability of choosing alternative i change as the y_{jn} increases?

$$\frac{\partial P_{in}}{\partial y_{jn}} = \frac{\partial (e^{V_{in}})(\sum_{j \in J_n} e^{V_{jn}})^{-1}}{\partial y_{jn}}$$

$$= -(e^{V_{in}})(\sum e^{V_{jn}})^{-2}(e^{V_{jn}})\frac{\partial V_{jn}}{\partial y_{jn}}$$

$$= -(\partial V_{jn}/\partial y_{jn})P_{in}P_{jn}.$$

In the case of V_{jn} being linear in observed variables, with a scalar coefficient β_y for y_{jn}, then

$$\partial P_{in}/\partial y_{jn} = -\beta_y P_{in} P_{jn}.$$

If y_{jn} is a desirable attribute such that β_y is positive, then increasing y_{jn} decreases the probability of choosing each alternative other than j. Furthermore, the decrease in probability is proportional to the value of the probability before y_{jn} was changed.

This latter fact is a property of logit models that can be undesirable in some situations. For example, consider a traveler's choice among auto, bus, and rail. If the probability of taking an auto is .60 and bus and rail each have a .20 probability, then an improvement in the attributes of bus travel (e.g., a reduction in its price) would reduce the probability of taking an auto three times as much as the probability of going by rail. If in reality most of the additional bus probability is drawn from the rail mode, then the standard logit model is inappropriate. The underlying problem in this situation is the independence of irrelevant alternatives (IIA) property of logit models, which is discussed in section 2.2. The solution is to take one of the corrective measures indicated in that discussion or to utilize a model, such as probit or GEV, as described in chapters 3 and 4, respectively, that does not exhibit the IIA property.

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 necessary consequence of the fact that the probabilities must sum to one before and after the change; it is demonstrated as follows:

$$\begin{aligned}\sum_{i\in J_n}\frac{\partial P_{in}}{\partial y_{jn}} &= \left(\frac{\partial V_{jn}}{\partial y_{jn}}\right)P_{jn}(1-P_{jn}) + \sum_{\substack{i\in J_n\\ i\neq j}}\left(-\frac{\partial V_{jn}}{\partial y_{jn}}\right)P_{jn}P_{in}\\ &= \left(\frac{\partial V_{jn}}{\partial y_{jn}}\right)P_{jn}\left[(1-P_{jn}) - \sum_{\substack{i\in J_n\\ i\neq j}}P_{in}\right]\\ &= \left(\frac{\partial V_{jn}}{\partial y_{jn}}\right)P_{jn}[(1-P_{jn})-(1-P_{jn})]\\ &= 0.\end{aligned}$$

In practical terms, if one alternative is improved so that its probability of being chosen increases, the additional probability is necessarily "drawn" from other alternatives. That is, 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.[3]

Economists often measure response by elasticities rather than derivatives, since elasticities are normalized for the variables' units. An elasticity is the percent change in one variable that is associated with a percent change in another variable. The elasticity of choice probabilities with respect to observed factors affecting the probabilities are now given. The elasticity of P_{in} with respect to y_{in}, a variable entering the utility of alternative i, is

$$\begin{aligned}E_{iy_i} &= (\partial P_{in}/\partial y_{in})(y_{in}/P_{in})\\ &= (\partial V_{in}/\partial y_{in})P_{in}(1-P_{in})(y_{in}/P_{in})\\ &= (\partial V_{in}/\partial y_{in})y_{in}(1-P_{in}).\end{aligned}$$

If representative utility is linear in y_{in}, with coefficient β_y, then

$$E_{iy_i} = \beta_y y_{in}(1-P_{in}).$$

The elasticity of P_{in} with respect to a variable entering alternative $j \neq i$, called a cross-elasticity, is calculated as

$$E_{iy_j} = -(\partial V_{jn}/\partial y_{jn})y_{jn}P_{jn},$$

which in the case of linear utility reduces to

$$E_{iy_j} = -\beta_y y_{jn}P_{jn}.$$

2.5 Average Probabilities, Derivatives, and Elasticities

Different individuals facing the same set of alternatives will, in general, have different representative utility for each alternative, because the characteristics of each alternative vary over people (e.g., the time required to travel to work by auto varies by place of home and place of work) and because individuals' characteristics (such as income, age, etc.) vary in the population. Decisionmakers with different representative utility for each alternative will have different choice probabilities. And, given that derivatives and elasticities depend on the choice probabilities, different individuals will be predicted to respond differently to changes in factors entering representative utility.

Usually a researcher is interested in the average probability or average response within a population, rather than the probability or response of any one individual. Methods for predicting population behavior with qualitative choice models are discussed in detail in chapter 6. It is useful at this point, however, to introduce the most straightforward and widely used method and to warn against an erroneous method that is nevertheless common.

Suppose the researcher has a random or stratified random[4] sample of individuals drawn from a population. Aggregate, or population, variables are predicted by taking the weighted average of the variables calculated for each individual. For example, to calculate the average probability in the population for a particular alternative, choice probabilities are calculated for each individual on the basis of the individual's characteristics and the characteristics of the alternatives as faced by the individual. The average probability for alternative i is then estimated as

$$\bar{P}_i = \sum_n w_n P_{in},$$

where w_n is sampling weight associated with individual n, and the summation is over all sampled individuals. If the sample is purely random, then w_n is the same for all sampled individuals and equals $1/N$, where N is the sample size. For stratified random samples, w_n varies over strata.

The number of individuals in the population predicted to choose alternative i is estimated as the average probability for alternative i times the population size:

$$N_i = M\bar{P}_i,$$

where M is the number of decisionmakers in the population and N_i is the estimated number that will choose alternative i. Average derivatives and elasticities are calculated similarly as the weighted average of individual derivatives and elasticities.

An alternative method of estimating average probabilities and responses is common but not consistent. Instead of calculating the probabilities and responses for a sample of decisionmakers and then taking averages, this alternative approach is to calculate probabilities and responses for an **average** decisionmaker and consider these to be in some way representative of average population behavior. For example, consider a mode choice model in which each traveler chooses between auto and transit on the basis of the cost and time associated with each. A consistent way to estimate the average probability of auto is to determine the times and costs faced by each person in a sample, calculate the probability of choosing auto for each person, and take the weighted average of these probabilities. The simpler, but inconsistent method, is to calculate the average cost and time associated with each mode and determine the probability of choosing auto given these average costs and times.

The inconsistency of this approach results from the fact that the choice probabilities, derivatives, and elasticities are nonlinear functions of the observed data and, as is well known, the average value of a nonlinear function over a range of data is not equal to the value of the function evaluated at the average of the data. The point can be made visually. Consider figure 2.3, which gives the probabilities of choosing a particular alternative for two individuals with representative utility for this alternative of a and b (assuming the representative utility of other alternatives is the same for the two individuals). The average probability is the average of the probabilities for the two individuals, namely, $(P_a + P_b)/2$. The probability evaluated at the average representative utility is given by the point on the logit curve above $(a + b)/2$. As shown for this case, the average probability is above the probability at the average representative utility. In general, the probability evaluated at the average utility underestimates the average probability when the individuals' choice probabilities are low and overestimates when they are high.

Estimating average responses by calculating derivatives and elasticities at the average representative utility is usually even more problematic than for average probabilities. Consider figure 2.4, depicting two individuals with representative utility a and b. The derivative of the choice probability

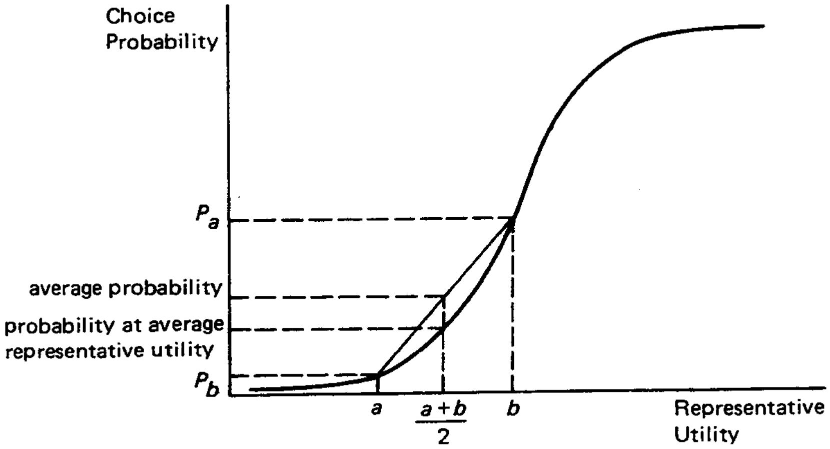

Figure 2.3
Difference between average probability and probability calculated at average representative utility.

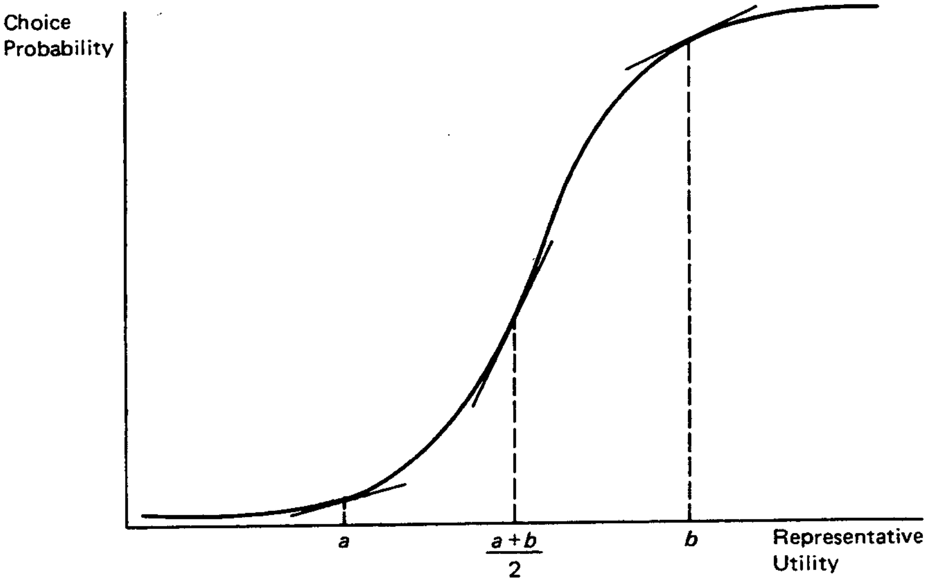

Figure 2.4
Difference between average response and response calculated at average representative utility.

for a change in representative utility is small for both of these individuals (the slope of the logit curve above a and b). Consequently, the average derivative is also small. However, the derivative at the average representative utility is very high (the slope above $(a + b)/2$). Estimating average responses in this way can be seriously misleading. In fact, Talvitie (1976) has found, in a mode choice situation, that elasticities at the average representative utility can be as much as two to three times greater or less than the average of individual elasticities.

2.6 Estimation

In order to calculate the choice probabilities for a particular decisionmaker the researcher must (unless a priori information is utilized) estimate the value of the parameter vector β. If the researcher does not intend to predict decisionmakers' choices, he might nevertheless be interested in knowing the value of β for other reasons. For example, suppose an auto manufacturer considers the utility consumers obtain from a make and model of vehicle to be $-\beta_1$ PP + β_2 FE plus a term for unobserved factors, where PP is the purchase price of the vehicles and FE is its fuel efficiency. This manufacturer could use information on the relative value of β_1 and β_2 to decide whether to incorporate into the vehicle that he produces a device that would increase its fuel efficiency but also increase its price.

Standard Estimation on Exogenous Sample

Assume that the researcher observes the choices of a sample of decisionmakers, along with the characteristics of each decisionmaker and each alternative faced by the decisionmaker. 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 described later in this section.

The parameter vector β is estimated by maximum likelihood methods, which can be described as follows for exogenously chosen samples. (For readers who are unfamiliar with maximum likelihood estimation, a straightforward discussion is given on pages 69–71 of the widely used econometrics text by Pindyck and Rubinfeld, 1981.) Consider one sampled decisionmaker, labeled n. The probability of person n choosing the alternative that he was actually observed to choose is

$$\prod_{i\in J_n} P_{in}^{\delta_{in}},$$

where δ_{in} is one if person n chose alternative i, and zero otherwise. Note that since $\delta_{in} = 0$ for all nonchosen alternatives and $P_{in}^0 = 1$, this term is simply the probability of the chosen alternative.

Consider now the entire sample. Since each decisionmaker's choice is independent of that of other decisionmakers, the probability of each person in the sample choosing the alternative that he was observed actually to choose is

$$L = \prod_{n\in N} \prod_{i\in J_n} P_{in}^{\delta_{in}}, \tag{2.10}$$

where N is the set of decisionmakers in the sample. This expression is simply the probability of each person's chosen alternative multiplied across all people in the sample.

Each P_{in} in expression (2.10) is a function of β and the observed data. Holding the observed data fixed, L can therefore be considered a function of β and written $L(\beta)$. In particular, it is the likelihood function for β, giving for each value of β the probability that the sampled decisionmakers would choose the alternatives that they actually did choose. The value of β that gives the highest such probability, that is, that maximizes the likelihood function, is called the maximum likelihood estimate of β. Under fairly general conditions, this estimator is consistent and efficient, as is usually the case with maximum likelihood estimators (see McFadden, 1973).

Rather than deal with the likelihood function itself, it is usually easier to maximize the log of the likelihood function. (Since the log operation is a monotonic function, the value of β that maximizes $L(\beta)$ will also maximize the log of $L(\beta)$.) This log likelihood function, designated LL, is written as

$$\mathrm{LL}(\beta) = \sum_{n\in N} \sum_{i\in J_n} \delta_{in} \log P_{in}. \tag{2.11}$$

Recalling that δ_{in} is zero for nonchosen alternatives, $\mathrm{LL}(\beta)$ is simply the log of the probability of the chosen alternative of each person, summed over all sampled decisionmakers. The estimate of β is that which maximizes this sum. (Note that, since L is a probability and consequently cannot exceed one, LL is always negative, since the log of one is zero. Therefore, maximizing LL is the same as minimizing its magnitude, a point that often causes confusion.) Several computer software packages are available that perform this maximization specifically in the context of logit models.

Given that the estimated value of β is that which maximizes $\mathrm{LL}(\beta)$, we can easily demonstrate that the estimated values of alternative-specific constants are those that equate the average probability for each alternative with the share of decisionmakers in the sample who actually chose that alternative. For ease of notation, consider a binary choice situation in which the representative utility of choosing the first alternative is zero by normalization and the representative utility of the second alternative is $\alpha + V_n$, where α is a constant and V_n varies over n (with its dependence on parameters suppressed in this notation). Then

$$P_{1n} = \frac{1}{1 + e^{\alpha+V_n}} \qquad \text{and} \qquad P_{2n} = \frac{e^{\alpha+V_n}}{1 + e^{\alpha+V_n}}.$$

Let δ_n equal one if person n chose alternative one and zero if person n chose alternative two (hence, $1 - \delta_n$ equals one if alternative two is chosen). The log likelihood function in this case is

$$\begin{aligned}
\mathrm{LL} &= \sum_n [\delta_n \log P_{1n} + (1 - \delta_n) \log P_{2n}] \\
&= \sum_n \left[\delta_n \log\left(\frac{1}{1 + e^{\alpha+V_n}}\right) + (1 - \delta_n) \log\left(\frac{e^{\alpha+V_n}}{1 + e^{\alpha+V_n}}\right)\right] \\
&= \sum_n [-\delta_n \log(1 + e^{\alpha+V_n}) + (1 - \delta_n)(\alpha + V_n) \\
&\qquad - (1 - \delta_n) \log(1 + e^{\alpha+V_n})] \\
&= \sum_n [(1 - \delta_n)(\alpha + V_n) - \log(1 + e^{\alpha+V_n})].
\end{aligned}$$

To maximize LL with respect to α, we take the derivative of LL with respect to α, equate the derivative to zero, and solve for α:

$$\begin{aligned}
\frac{\partial \mathrm{LL}}{\partial \alpha} &= \sum_n \left[(1 - \delta_n) - \frac{e^{\alpha+V_n}}{1 + e^{\alpha+V_n}} \right] \\
&= \sum_n [(1 - \delta_n) - P_{2n}] \\
&= \sum_n [(1 - \delta_n) - (1 - P_{1n})] = \sum_n (P_{1n} - \delta_n) = 0.
\end{aligned}$$

Therefore,

$$\sum_n \delta_n = \sum_n P_{1n},$$

and

$$\frac{1}{N}\sum_n \delta_n = \frac{1}{N}\sum_n P_{1n},$$

where N is the total sample size. That is, the value of α that maximizes the log likelihood function (that is, the estimated value of α) is that which equates the average probability for each alternative $((1/N)\sum P_{1n})$ with the share of sampled decisionmakers who chose that alternative $((1/N)\sum \delta_n)$.

Though the proof is more tedious, the same result applies in the multinomial case. That is, if a constant is included in the representative utility of each alternative (except, of course, for one alternative, whose constant is normalized to zero; see section 2.3), then the estimated values of these constants are such that the average probability for each alternative equals the share of sampled decisionmakers who actually chose that alternative. This fact has considerable practical importance—for example (see in section 2.2), the mitigation of the ill effects of the independence from irrelevant alternatives property not holding for the choice situation being examined.

Estimation on a Subset of Alternatives

In some situations, the number of alternatives facing the decisionmaker is so large that estimating model parameters is very expensive or even impossible (due perhaps to core capacity limitations of the researcher's computer). As mentioned in the discussion of the independence from irrelevant alternatives property, estimation can be performed on a subset of alternatives and not lose consistency. For example, a researcher examining a choice situation that involves 100 alternatives can estimate on a subset of 10 alternatives for each sampled decisionmaker, with the person's chosen alternative included as well as 9 alternatives randomly selected from the remaining 99.

In general, estimation with a subset of alternatives for each sampled decisionmaker proceeds as if each decisionmaker actually faced only the alternatives in the subset. Denote the subset of alternatives selected for person n as K_n, which can be the same or different for different persons. Label the set of sampled individuals who actually chose an alternative within their subset as M. A "quasi" log likelihood function is constructed as

$$\mathrm{QLL}(\beta) = \sum_{n \in M} \sum_{i \in K_n} \delta_{in} \log \tilde{P}_{in},$$

where

$$\tilde{P}_{in} = \frac{e^{V_{in}}}{\sum_{j \in K_n} e^{V_{jn}}}.$$

This is the same as the log likelihood function given in equation (2.11) except (1) the subset of alternatives K_n replaces, for each sampled person, the complete set J_n in both the calculation of the probabilities and in the summation within the function, and (2) only the sampled persons in subset M are included in the summation rather than all sampled persons (that is, those whose chosen alternative is not in their subset of alternatives are excluded). Since, in accordance with the independence from irrelevant alternatives property, relative probabilities within a subset of alternatives are unaffected by exclusion of aiternatives not in the subset, maximization of QLL(β) provides a consistent estimate of β. However, since information is excluded from QLL(β) that LL(β) incorporates (i.e., information on alternatives not in each subset and on decisionmakers whose chosen alternatives are not in their subsets), the value of β that maximizes QLL(β) is not an efficient estimator.

Estimation with 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 housholds in most areas would include only a small proportion who had chosen solar water heating systems. If the researcher is particularly interested in factors that affect the penetration of solar devices, estimation on a random sample of households would require a very large total sample size.

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 heater systems.

Samples selected on the basis of decisionmakers' choices can be purely choice based or a hybrid of choice based and exogenous. For a purely choice based sample, the population is divided into those that choose each alternative and decisionmakers within each group are drawn randomly, 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 N households, and draw randomly from all other communities at a rate of one out of M, where M is larger than N. 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 decisionmaker's choice is fairly complex in general, and varies with the exact form of the sampling procedure. For interested readers, details are given by Ben-Akiva and Lerman (1985).

One result, however, is particularly significant, since it allows researchers to use choice based samples without becoming involved in complex estimation procedures. This result 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 the model parameters **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 i, labeled $\hat{\alpha}_i$, is related to the true constant α_i^*,

$$E(\hat{\alpha}_i) = \alpha_i^* - \ln(A_i/S_i),$$

where A_i is the proportion of decisionmakers in the population that choose alternative i and S_i is the proportion in the choice based sample that choose alternative i. Consequently, if A_i is known (that is, if population shares are known for each alternative), then a consistent estimate of the alternative-specific constant is the estimated constant $\hat{\alpha}_i$ plus $\ln(A_i/S_i)$.

2.7 Goodness of Fit

A statistic, called the likelihood ratio index, is often used with qualitative choice models to measure how well the model fits 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

$$\rho = 1 - (\mathrm{LL}(\beta^*)/\mathrm{LL}(0)),$$

where $\mathrm{LL}(\beta^*)$ 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 $\mathrm{LL}(\beta^*) = \mathrm{LL}(0)$ and so $\rho = 0$. This is the lowest value that ρ can take (since if $\mathrm{LL}(\beta^*)$ is less than LL(0), then β^* would not be the maximum likelihood estimate).

At the other extreme, suppose the estimated model were so good that each sampled decisionmaker'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 $\mathrm{LL}(\beta^*) = 0$, $\rho = 1$. 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 allow for perfectly predicting the choices of the sampled decisionmakers.

It is important to note that the likelihood ratio index is not at all similar in its interpretation to the *R*-squared used in regression, despite both statistics having the same range. *R*-squared indicates the percent 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 percent increase in the log likelihood function above the value taken at zero parameters (since $\rho = 1 - (\mathrm{LL}(\beta^*)/\mathrm{LL}(0)) = (\mathrm{LL}(0) - \mathrm{LL}(\beta^*))/\mathrm{LL}(0))$. However, the meaning of such a percent 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 decisionmaker cannot be compared via their likelihood ratio index values.

Another goodness-of-fit statistic that is sometimes used, but is of even less value than the likelihood ratio index, is the "percent correctly predicted." This statistic is calculated by identifying for each sampled decisionmaker the alternative with the highest probability, based on the estimated model, and determining whether or not this was the alternative that the decisionmaker actually chose; the percent of sampled decisionmakers for which the highest probability alternative and the chosen alternative are the same is called the percent correctly predicted.

This statistic, while popular in the early applications of qualitative choice models, 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 decisionmaker is predicted by the researcher to choose the alternative for which the model gives the highest probability. Recall from section 1.3, however, that the researcher does not have enough information to predict the decisionmaker's choice; he has only enough information to state the probability that the decisionmaker will choose each alternative. In stating choice probabilities, the researcher is saying that if the choice situation were repeated numerous times, 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 might be useful. Suppose an estimated model predicts choice probabilities of .75 and .25 in a two-alternative situation. Those probabilities mean that if the situation were repeated 100 times, the researcher's best predictions of how many times each alternative would be chosen are 75 and 25. However, the percent correctly predicted statistic is based on the notion that the best prediction in each situation is the alternative with the highest probability. With 100 repetitions, this notion would predict that one alternative would be chosen all 100 times and the other alternative never chosen. This misses the point of probabilities and seems to imply that the researcher has perfect information.

2.8 Hypothesis Testing

A likelihood ratio test is a very general test that is used in nearly all contexts. (The one major exception is for testing hypotheses on individual parameters for which standard t-tests are performed.) Consider a null hypothesis H that can be expressed as constraints on the values of the parameters. Two of the most common such hypotheses are (1) several parameters being zero, and (2) two or more parameters being equal to each other. The constrained maximum likelihood estimate of the parameters (labeled β^{H}) is that value of β that gives the highest value of LL without violating the constraints of the null hypothesis H. For example, if H is the hypothesis that the first two elements of N-tuple β are equal, then β^{H} is the value of β that, out of the set of all N-tuples whose first two elements are equal, results in the highest value of the likelihood function.

Define the ratio of likelihoods, $R = L(\beta^{\mathrm{H}})/L(\beta^*)$, where $L(\beta^{\mathrm{H}})$ is the (constrained) maximum value of the likelihood function under the null hypothesis H and $L(\beta^*)$ is the unconstrained maximum of the likelihood function. As in likelihood ratio tests for models other than those of qualitative 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 $-2(\mathrm{LL}(\beta^{\mathrm{H}}) - \mathrm{LL}(\beta^*))$. 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.

Examples

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 summing 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.

2.9 Derivation of Logit Probabilities

It was stated without proof in section 2.1 that if the utility of alternative i is decomposed into observed and unobserved parts $U_{in} = V_{in} + e_{in}$ and each e_{in} is independently identically distributed in accordance with the extreme value distribution, then the choice probabilities have the logit form

$$P_{in} = \frac{e^{V_{in}}}{\sum_j e^{V_{jn}}}$$

This statement is demonstrated as follows.

Under the extreme value distribution, the density function for each e_{in} is

$$\exp(-e_{in}) \cdot \exp(-e^{-e_{in}}),$$

with a cumulative distribution of

$$\exp(-e^{-e_{in}}).$$

The probability that alternative i is chosen is

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

Rearranging terms within the parentheses, we can write

$$P_{in} = \text{Prob}(e_{jn} < e_{in} + V_{in} - V_{jn}, \text{ for all } j \text{ in } J_n, j \neq i). \tag{2.12}$$

Suppose, for the moment, that e_{in} takes a particular value, say, s. The probability that alternative i is chosen is then the probability that each e_{jn} is less than $s + V_{in} - V_{jn}$, respectively, for all j in $J_n, j \neq i$. The probability that $e_{in} = s$ and, simultaneously, that $e_{jn} < s + V_{in} - V_{jn}$, for all j in J_n, $j \neq i$, is the density of e_{in} evaluated at s times the cumulative distribution for each e_{jn} except e_{in} evaluated at $s + V_{in} - V_{jn}$. From the extreme value distribution, this is

$$e^{-s} \exp(-e^{-s}) \prod_{\substack{j \in J_n \\ j \neq i}} \exp(-e^{-(s+V_{in}-V_{jn})}).$$

Since $V_{in} - V_{in} = 0$, this expression can be rewritten as

$$e^{-s} \prod_{j \in J_n} \exp(-e^{-(s+V_{in}-V_{jn})}). \tag{2.13}$$

The random variable e_{in} need not equal s, however; it can take any value within its range. The right-hand side of equation (2.12) is, therefore, the sum of expression (2.13) over all possible values of s. That is, since e_{in} is continuous, equation (2.12) becomes

$$P_{in} = \int_{s=-\infty}^{\infty} e^{-s} \prod \exp(-e^{-(s+V_{in}-V_{jn})})\, ds.$$

Our task in deriving the choice probabilities is to evaluate this integral. Collecting terms in the exponent of e,

$$P_{in} = \int_{s=-\infty}^{\infty} e^{-s} \exp\left\{-\sum_{j \in J_n} e^{-(s+V_{in}-V_{jn})}\right\} ds$$

$$= \int_{s=-\infty}^{\infty} e^{-s} \exp\left\{-e^{-s} \sum_{j \in J_n} e^{-(V_{in}-V_{jn})}\right\} ds.$$

Let $e^{-s} = t$. Then $-e^{-s}ds = dt$ and $ds = -(dt/t)$. Note that as s approaches infinity, t approaches zero, and as s approaches negative infinity, t becomes infinitely large. Using these new terms,

$$P_{in} = \int_{\infty}^{0} t \exp\left\{-t \cdot \sum_{j \in J_n} e^{-(V_{in}-V_{jn})}\right\} (-dt/t)$$

$$= \int_{0}^{\infty} \exp\left\{-t \cdot \sum_{j \in J_n} e^{-(V_{in}-V_{jn})}\right\} dt$$

$$= \left.\frac{\exp\{-t \cdot \sum_{j \in J_n} e^{-(V_{in}-V_{jn})}\}}{-\sum_{j \in J_n} e^{-(V_{in}-V_{jn})}}\right|_{0}^{\infty}$$

$$= \frac{1}{\sum_{j \in J_n} e^{-(V_{in}-V_{jn})}} = \frac{e^{V_{in}}}{\sum_{j \in J_n} e^{V_{jn}}},$$

as required.