2 Properties of Discrete Choice Models

2.1 Overview

This chapter describes the features that are common to all discrete choice models. We start by discussing the choice set, which is the set of options that are available to the decision maker. We then define choice probabilities and derive them from utility-maximizing behavior. The most prominent types of discrete choice models, namely logit, generalized extreme value (GEV), probit, and mixed logit, are introduced and compared within the context of this general derivation. Utility, as a constructed measure of well-being, has no natural level or scale. This fact has important implications for the specification and normalization of discrete choice models, which we explore. We then show how individual-level models are aggregated to obtain market-level predictions, and how the models are used for forecasting over time.

2.2 The Choice Set

Discrete choice models describe decision makers' choices among alternatives. The decision makers can be people, households, firms, or any other decision-making unit, and the alternatives might represent competing products, courses of action, or any other options or items over which choices must be made. To fit within a discrete choice framework, the set of alternatives, called the *choice set*, needs to exhibit three characteristics. First, the alternatives must be *mutually exclusive* from the decision maker's perspective. Choosing one alternative necessarily implies not choosing any of the other alternatives. The decision maker chooses only one alternative from the choice set. Second, the choice set must be *exhaustive*, in that all possible alternatives are included. The decision maker necessarily chooses one of the alternatives. Third, the number of alternatives must be finite. The researcher can count the alternatives and eventually be finished counting.

The first and second criteria are not restrictive. Appropriate definition of alternatives can nearly always assure that the alternatives are mutually exclusive and the choice set is exhaustive. For example, suppose two alternatives labeled *A* and *B* are not mutually exclusive because the decision maker can choose both of the alternatives. The alternatives can be redefined to be "*A* only," "*B* only," and "both *A* and *B*," which are necessarily mutually exclusive. Similarly, a set of alternatives might not be exhaustive because the decision maker has the option of not choosing any of them. In this case, an extra alternative can be defined as "none of the other alternatives." The expanded choice set, consisting of the original alternatives plus this new one, is clearly exhaustive.

Often the researcher can satisfy these two conditions in several different ways. The appropriate specification of the choice set in these situations is governed largely by the goals of the research and the data that are available to the researcher. Consider households' choice among heating fuels, a topic which has been studied extensively in efforts to forecast energy use and to develop effective fuel-switching and energy conservation programs. The available fuels are usually natural gas, electricity, oil, and wood. These four alternatives, as listed, violate both mutual exclusivity and exhaustiveness. The alternatives are not mutually exclusive because a household can (and many do) have two types of heating, e.g., a natural gas central heater and electric room heaters, or a wood stove along with electric baseboard heating. And the set is not exhaustive because the household can have no heating (which, unfortunately, is not as rare as one might hope). The researcher can handle each of these issues in several ways. To obtain mutually exclusive alternatives, one approach is to list every possible combination of heating fuels as an alternative. The alternatives are then defined as: "electricity alone," "electricity and natural gas, but no other fuels," and so on. Another approach is to define the choice as the choice among fuels for the "primary" heating source. Under this procedure, the researcher develops a rule for determining which heating fuel is primary when a household uses multiple heating fuels. By definition, only one fuel (electricity, natural gas, oil, or wood) is primary. The advantage of listing every possible combination of fuels is that it avoids the need to define a "primary" fuel, which is a difficult and somewhat arbitrary distinction. Also, with all combinations considered, the researcher has the ability to examine the factors that determine households' use of multiple fuels. However, to implement this approach, the researcher needs data that distinguish the alternatives, for example, the cost of heating a house with natural gas and electricity versus the cost with natural gas alone. If the researcher restricts the analysis to choice of primary fuel, then the data requirements

are less severe. Only the costs associated with each fuel are needed. Also, a model with four alternatives is inherently easier to estimate and forecast with than a model with the large number of alternatives that arises when every possible combination of fuels is considered. The researcher will need to take these trade-offs into consideration when specifying the choice set.

The same type of issue arises with regard to exhaustiveness. In our case of heating-fuel choice, the researcher can either include "no heating" as an alternative or can redefine the choice situation as being the choice of heating fuel conditional on having heating. The first approach allows the researcher to examine the factors that relate to whether a household has heating. However, this ability is only realized if the researcher has data that meaningfully relate to whether or not a household has heating. Under the second approach, the researcher excludes from the analysis households without heating, and, by doing so, is relieved of the need for data that relate to these households.

As we have just described, the conditions of mutual exclusivity and exhaustiveness can usually be satisfied, and the researcher often has several approaches for doing so. In contrast, the third condition, namely, that the number of alternatives is finite, is actually restrictive. This condition is the defining characteristic of discrete choice models and distinguishes their realm of application from that for regression models. With regression models, the dependent variable is continuous, which means that there is an infinite number of possible outcomes. The outcome might be chosen by a decision maker, such as the decision of how much money to hold in savings accounts. However, the alternatives available to the decision maker, which are every possible monetary value above zero, is not finite (at least not if all fractions are considered, which is an issue we return to later.) When there is an infinite number of alternatives, discrete choice models cannot be applied.

Often regression models and discrete choice models are distinguished by saying that regressions examine choices of "how much" and discrete choice models examine choice of "which." This distinction, while perhaps illustrative, is not actually accurate. Discrete choice models can be and have been used to examine choices of "how much." A prominent example is households' choice of how many cars to own. The alternatives are 0, 1, 2, and so on, up to the largest number that the researcher considers possible (or observes). This choice set contains a finite number of mutually exclusive and exhaustive alternatives, appropriate for analysis via discrete choice models. The researcher can also define the choice set more succinctly as 0, 1, and 2 or more vehicles, if the goals of the research can be met with this specification.

When considered in this way, most choices involving "how many" can be represented in a discrete choice framework. In the case of savings accounts, every one-dollar increment (or even every one-cent increment) can be considered an alternative, and as long as some finite maximum exists, then the choice set fits the criteria for discrete choice. Whether to use regression or discrete choice models in these situations is a specification issue that the researcher must consider. Usually a regression model is more natural and easier. A discrete choice model would be used in these situations only if there were compelling reasons for doing so. As an example, Train *et al.* (1987a) analyzed the number and duration of phone calls that households make, using a discrete choice model instead of a regression model because the discrete choice model allowed greater flexibility in handling the nonlinear price schedules that households face. In general, the researcher needs to consider the goals of the research and the capabilities of alternative methods when deciding whether to apply a discrete choice model.

2.3 Derivation of Choice Probabilities

Discrete choice models are usually derived under an assumption of utility-maximizing behavior by the decision maker. Thurstone (1927) originally developed the concepts in terms of psychological stimuli, leading to a binary probit model of whether respondents can differentiate the level of stimulus. Marschak (1960) interpreted the stimuli as utility and provided a derivation from utility maximization. Following Marschak, models that can be derived in this way are called random utility models (RUMs). It is important to note, however, that models derived from utility maximization can also be used to represent decision making that does not entail utility maximization. The derivation assures that the model is consistent with utility maximization; it does not preclude the model from being consistent with other forms of behavior. The models can also be seen as simply describing the relation of explanatory variables to the outcome of a choice, without reference to exactly how the choice is made.

Random utility models (RUMs) are derived as follows. A decision maker, labeled n, faces a choice among J alternatives. The decision maker would obtain a certain level of utility (or profit) from each alternative. The utility that decision maker n obtains from alternative j is U_{nj}, $j = 1, \ldots, J$. This utility is known to the decision maker but not, as we see in the following, by the researcher. The decision maker chooses the alternative that provides the greatest utility. The behavioral model is therefore: choose alternative i if and only if $U_{ni} > U_{nj}\ \forall j \neq i$.

Consider now the researcher. The researcher does not observe the decision maker's utility. The researcher observes some attributes of the alternatives as faced by the decision maker, labeled $x_{nj}\ \forall j$, and some attributes of the decision maker, labeled s_n, and can specify a function that relates these observed factors to the decision maker's utility. The function is denoted $V_{nj} = V(x_{nj}, s_n)\ \forall j$ and is often called *representative utility*. Usually, V depends on parameters that are unknown to the researcher and therefore estimated statistically; however, this dependence is suppressed for the moment.

Since there are aspects of utility that the researcher does not or cannot observe, $V_{nj} \neq U_{nj}$. Utility is decomposed as $U_{nj} = V_{nj} + \varepsilon_{nj}$, where ε_{nj} captures the factors that affect utility but are not included in V_{nj}. This decomposition is fully general, since ε_{nj} is defined as simply the difference between true utility U_{nj} and the part of utility that the researcher captures in V_{nj}. Given its definition, the characteristics of ε_{nj}, such as its distribution, depend critically on the researcher's specification of V_{nj}. In particular, ε_{nj} is not defined for a choice situation *per se*. Rather, it is defined relative to a researcher's representation of that choice situation. This distinction becomes relevant when evaluating the appropriateness of various specific discrete choice models.

The researcher does not know $\varepsilon_{nj}\ \forall j$ and therefore treats these terms as random. The joint density of the random vector $\varepsilon_n = \langle \varepsilon_{n1}, \ldots, \varepsilon_{nJ} \rangle$ is denoted $f(\varepsilon_n)$. With this density, the researcher can make probabilistic statements about the decision maker's choice. The probability that decision maker n chooses alternative i is

$$\begin{aligned} P_{ni} &= \text{Prob}(U_{ni} > U_{nj}\ \forall j \neq i) \\ &= \text{Prob}(V_{ni} + \varepsilon_{ni} > V_{nj} + \varepsilon_{nj}\ \forall j \neq i) \\ &= \text{Prob}(\varepsilon_{nj} - \varepsilon_{ni} < V_{ni} - V_{nj}\ \forall j \neq i). \end{aligned} \tag{2.1}$$

This probability is a cumulative distribution, namely, the probability that each random term $\varepsilon_{nj} - \varepsilon_{ni}$ is below the observed quantity $V_{ni} - V_{nj}$. Using the density $f(\varepsilon_n)$, this cumulative probability can be rewritten as

$$\begin{aligned} P_{ni} &= \text{Prob}(\varepsilon_{nj} - \varepsilon_{ni} < V_{ni} - V_{nj}\ \forall j \neq i) \\ &= \int_{\varepsilon} I(\varepsilon_{nj} - \varepsilon_{ni} < V_{ni} - V_{nj}\ \forall j \neq i) f(\varepsilon_n)\, d\varepsilon_n, \end{aligned} \tag{2.2}$$

where $I(\cdot)$ is the indicator function, equaling 1 when the expression in parentheses is true and 0 otherwise. This is a multidimensional integral over the density of the unobserved portion of utility, $f(\varepsilon_n)$. Different discrete choice models are obtained from different specifications of this density, that is, from different assumptions about the distribution of the

unobserved portion of utility. The integral takes a closed form only for certain specifications of $f(\cdot)$. Logit and nested logit have closed-form expressions for this integral. They are derived under the assumption that the unobserved portion of utility is distributed iid extreme value and a type of generalized extreme value, respectively. Probit is derived under the assumption that $f(\cdot)$ is a multivariate normal, and mixed logit is based on the assumption that the unobserved portion of utility consists of a part that follows any distribution specified by the researcher plus a part that is iid extreme value. With probit and mixed logit, the resulting integral does not have a closed form and is evaluated numerically through simulation. Each of these models is discussed in detail in subsequent chapters.

The meaning of choice probabilities is more subtle, and more revealing, than it might at first appear. An example serves as illustration. Consider a person who can take either a car or a bus to work. The researcher observes the time and cost that the person would incur under each mode. However, the researcher realizes that there are factors other than time and cost that affect the person's utility and hence his choice. The researcher specifies

$$V_c = \alpha T_c + \beta M_c,$$
$$V_b = \alpha T_b + \beta M_b,$$

where T_c and M_c are the time and cost (in money) that the person incurs traveling to work by car, T_b and M_b are defined analogously for bus, and the subscript n denoting the person is omitted for convenience. The coefficients α and β are either known or estimated by the researcher.

Suppose that, given α and β and the researcher's measures of the time and cost by car and bus, it turns out that $V_c = 4$ and $V_b = 3$. This means that, on observed factors, car is better for this person than bus by 1 unit. (We discuss in following text the normalization of utility that sets the dimension of these units.) It does not mean, however, that the person necessarily chooses car, since there are other factors not observed by the researcher that affect the person. The probability that the person chooses bus instead of car is the probability that the unobserved factors for bus are sufficiently better than those for car to overcome the advantage that car has on observed factors. Specifically, the person will choose bus if the unobserved portion of utility is higher than that for car by at least 1 unit, thus overcoming the 1-unit advantage that car has on observed factors. The probability of this person choosing bus is therefore the probability that $\varepsilon_b - \varepsilon_c > 1$. Likewise, the person will choose car if the unobserved utility for bus is *not* better than that for car by at least 1 unit, that is, if $\varepsilon_b - \varepsilon_c < 1$. Since 1 is the difference between V_c and V_b in our example,

the probabilities can be stated more explicitly as

$$P_c = \text{Prob}(\varepsilon_b - \varepsilon_c < V_c - V_b)$$

and

$$\begin{aligned} P_b &= \text{Prob}(\varepsilon_b - \varepsilon_c > V_c - V_b) \\ &= \text{Prob}(\varepsilon_c - \varepsilon_b < V_b - V_c). \end{aligned}$$

These equations are the same as equation (2.1), re-expressed for our car–bus example.

The question arises in the derivation of the choice probabilities: what is meant by the distribution of ε_n? The interpretation that the researcher places on this density affects the researcher's interpretation of the choice probabilities. The most prominent way to think about this distribution is as follows. Consider a population of people who face the same observed utility $V_{nj}\ \forall j$ as person n. Among these people, the values of the unobserved factors differ. The density $f(\varepsilon_n)$ is the distribution of the unobserved portion of utility within the population of people who face the same observed portion of utility. Under this interpretation, the probability P_{ni} is the share of people who choose alternative i within the population of people who face the same observed utility for each alternative as person n. The distribution can also be considered in subjective terms, as representing the researcher's subjective probability that the person's unobserved utility will take given values. In this case, P_{ni} is the probability that the researcher ascribes to the person's choosing alternative i given the researcher's ideas about the unobserved portions of the person's utility. As a third possibility, the distribution can represent the effect of factors that are quixotic to the decision maker himself (representing, e.g., aspects of bounded rationality), so that P_{ni} is the probability that these quixotic factors induce the person to choose alternative i given the observed, nonquixotic factors.

2.4 Specific Models

Logit, GEV, probit, and mixed logit are discussed at length in the subsequent chapters. However, a quick preview of these models is useful at this point, to show how they relate to the general derivation of all choice models and how they differ within this derivation. As stated earlier, different choice models are derived under different specifications of the density of unobserved factors, $f(\varepsilon_n)$. The issues therefore are what distribution is assumed for each model, and what is the motivation for these different assumptions.

Logit (discussed in Chapter 3) is by far the most widely used discrete choice model. It is derived under the assumption that ε_{ni} is iid extreme value for all i. The critical part of the assumption is that the unobserved factors are uncorrelated over alternatives, as well as having the same variance for all alternatives. This assumption, while restrictive, provides a very convenient form for the choice probability. The popularity of the logit model is due to this convenience. However, the assumption of independence can be inappropriate in some situations. Unobserved factors related to one alternative might be similar to those related to another alternative. For example, a person who dislikes travel by bus because of the presence of other riders might have a similar reaction to rail travel; if so, then the unobserved factors affecting bus and rail are corrrelated rather than independent. The assumption of independence also enters when a logit model is applied to sequences of choices over time. The logit model assumes that each choice is independent of the others. In many cases, one would expect that unobserved factors that affect the choice in one period would persist, at least somewhat, into the next period, inducing dependence among the choices over time.

The development of other models has arisen largely to avoid the independence assumption within a logit. Generalized extreme-value models (GEV, discussed in Chapter 4) are based, as the name implies, on a generalization of the extreme-value distribution. The generalization can take many forms, but the common element is that it allows correlation in unobserved factors over alternatives and collapses to the logit model when this correlation is zero. Depending on the type of GEV model, the correlations can be more or less flexible. For example, a comparatively simple GEV model places the alternatives into several groups called nests, with unobserved factors having the same correlation for all alternatives within a nest and no correlation for alternatives in different nests. More complex forms allow essentially any pattern of correlation. GEV models usually have closed forms for the choice probabilities, so that simulation is not required for their estimation.

Probits (Chapter 5) are based on the assumption that the unobserved factors are distributed jointly normal: $\varepsilon_n' = \langle \varepsilon_{n1}, \ldots, \varepsilon_{nJ} \rangle \sim N(0, \Omega)$. With full covariance matrix Ω, any pattern of correlation and heteroskedasticity can be accommodated. When applied to sequences of choices over time, the unobserved factors are assumed to be jointly normal over time as well as over alternatives, with any temporal correlation pattern. The flexibility of the probit model in handling correlations over alternatives and time is its main advantage. Its only functional limitation arises from its reliance on the normal distribution. In some situations, unobserved factors may not be normally distributed. For example, a customer's willingness to pay for a desirable attribute of a product is

necessary positive. Assuming that this unobserved factor is normally distributed contradicts the fact that it is positive, since the normal distribution has density on both sides of zero.

Mixed logit (Chapter 6) allows the unobserved factors to follow any distribution. The defining characteristic of a mixed logit is that the unobserved factors can be decomposed into a part that contains all the correlation and heteroskedasticity, and another part that is iid extreme value. The first part can follow any distribution, including non-normal distributions. We will show that mixed logit can approximate any discrete choice model and thus is fully general.

Other discrete choice models (Chapter 7) have been specified by researchers for specific purposes. Often these models are obtained by combining concepts from other models. For example, a mixed probit is obtained by decomposing the unobserved factors into two parts, as in mixed logit, but giving the second part a normal distribution instead of extreme value. This model has the generality of mixed logit and yet for some situations can be easier to estimate. By understanding the derivation and motivation for all the models, each researcher can specify a model that is tailor-made for the situation and goals of her research.

2.5 Identification of Choice Models

Several aspects of the behavioral decision process affect the specification and estimation of any discrete choice model. The issues can be summarized easily in two statements: "Only differences in utility matter" and "The scale of utility is arbitrary." The implications of these statements are far-reaching, subtle, and, in many cases, quite complex. We discuss them below.

2.5.1. Only Differences in Utility Matter

The absolute level of utility is irrelevant to both the decision maker's behavior and the researcher's model. If a constant is added to the utility of all alternatives, the alternative with the highest utility doesn't change. The decision maker chooses the same alternative with $U_{nj}\ \forall j$ as with $U_{nj} + k\ \forall j$ for any constant k. A colloquial way to express this fact is, "A rising tide raises all boats."

The level of utility doesn't matter from the researcher's perspective either. The choice probability is $P_{ni} = \text{Prob}(U_{ni} > U_{nj}\ \forall j \neq i) = \text{Prob}(U_{ni} - U_{nj} > 0\ \forall j \neq i)$, which depends only on the difference in utility, not its absolute level. When utility is decomposed into the observed and unobserved parts, equation (2.1) expresses the choice

probability as $P_{ni} = \text{Prob}(\varepsilon_{nj} - \varepsilon_{ni} < V_{ni} - V_{nj} \; \forall j \neq i)$, which also depends only on differences.

The fact that only differences in utility matter has several implications for the identification and specification of discrete choice models. In general it means that the only parameters that can be estimated (that is, are identified) are those that capture differences across alternatives. This general statement takes several forms.

Alternative-Specific Constants

It is often reasonable to specify the observed part of utility to be linear in parameters with a constant: $V_{nj} = x'_{nj}\beta + k_j \; \forall j$, where x_{nj} is a vector of variables that relate to alternative j as faced by decision maker n, β are coefficients of these variables, and k_j is a constant that is specific to alternative j. The alternative-specific constant for an alternative captures the average effect on utility of all factors that are not included in the model. Thus they serve a similar function to the constant in a regression model, which also captures the average effect of all unincluded factors.

When alternative-specific constants are included, the unobserved portion of utility, ε_{nj}, has zero mean by construction. If ε_{nj} has a nonzero mean when the constants are not included, then adding the constants makes the remaining error have zero mean: that is, if $U_{nj} = x'_{nj}\beta + \varepsilon^*_{nj}$ with $E(\varepsilon_{nj})^* = k_j \neq 0$, then $U_{nj} = x'_{nj}\beta + k_j + \varepsilon_{nj}$ with $E(\varepsilon_{nj}) = 0$. It is reasonable, therefore, to include a constant in V_{nj} for each alternative. However, since only differences in utility matter, only differences in the alternative-specific constants are relevant, not their absolute levels. To reflect this fact, the researcher must set the overall level of these constants.

The concept is readily apparent in the car–bus example. A specification of utility that takes the form

$$U_c = \alpha T_c + \beta M_c + k^0_c + \varepsilon_c,$$
$$U_b = \alpha T_b + \beta M_b + k^0_b + \varepsilon_b,$$

with $k^0_b - k^0_c = d$, is equivalent to a model with

$$U_c = \alpha T_c + \beta M_c + k^1_c + \varepsilon_c,$$
$$U_b = \alpha T_b + \beta M_b + k^1_b + \varepsilon_b,$$

where the difference in the new constants is the same as the difference in the old constants, namely, $k^1_b - k^1_c = d = k^0_b - k^0_c$. Any model with the same difference in constants is equivalent. In terms of estimation, it is impossible to estimate the two constants themselves, since an infinite

number of values of the two constants (any values that have the same difference) result in the same choice probabilities.

To account for this fact, the researcher must normalize the absolute levels of the constants. The standard procedure is to normalize one of the constants to zero. For example, the researcher might normalize the constant for the car alternative to zero:

$$U_c = \alpha T_c + \beta M_c + \varepsilon_c,$$
$$U_b = \alpha T_b + \beta M_b + k_b + \varepsilon_b.$$

Under this normalization, the value of k_b is d, which is the difference in the original (unnormalized) constants. The bus constant is interpreted as the average effect of unincluded factors on the utility of bus *relative* to car.

With J alternatives, at most $J - 1$ alternative-specific constants can enter the model, with one of the constants normalized to zero. It is irrelevant which constant is normalized to zero: the other constants are interpreted as being relative to whichever one is set to zero. The researcher could normalize to some value other than zero, of course; however, there would be no point in doing so, since normalizing to zero is easier (the constant is simply left out of the model) and has the same effect.

Sociodemographic Variables

The same issue affects the way that socio-demographic variables enter a model. Attributes of the alternatives, such as the time and cost of travel on different modes, generally vary over alternatives. However, attributes of the decision maker do not vary over alternatives. They can only enter the model if they are specified in ways that create differences in utility over alternatives.

Consider for example the effect of a person's income on the decision whether to take bus or car to work. It is reasonable to suppose that a person's utility is higher with higher income, whether the person takes bus or car. Utility is specified as

$$U_c = \alpha T_c + \beta M_c + \theta_c^0 Y + \varepsilon_c,$$
$$U_b = \alpha T_b + \beta M_b + \theta_b^0 Y + k_b + \varepsilon_b,$$

where Y is income and θ_c^0 and θ_b^0 capture the effects of changes in income on the utility of taking car and bus, respectively. We expect that $\theta_c^0 > 0$ and $\theta_b^0 > 0$, since greater income makes people happier no matter what mode they take. However, $\theta_c^0 \neq \theta_b^0$, since income probably has a different effect on the person depending on his mode of travel. Since only differences in utility matter, the absolute levels of θ_c^0 and θ_b^0 cannot be estimated, only their difference. To set the level, one of these

parameters is normalized to zero. The model becomes

$$U_c = \alpha T_c + \beta M_c + \varepsilon_c,$$
$$U_b = \alpha T_b + \beta M_b + \theta_b Y + k_b + \varepsilon_b,$$

where $\theta_b = \theta_b^0 - \theta_c^0$ and is interpreted as the differential effect of income on the utility of bus compared to car. The value of θ_b can be either positive or negative.

Sociodemographic variables can enter utility in other ways. For example, cost is often divided by income:

$$U_c = \alpha T_c + \beta M_c / Y + \varepsilon_c,$$
$$U_b = \alpha T_b + \beta M_b / Y + \theta_b Y + k_b + \varepsilon_b.$$

The coefficient of cost in this specification is β / Y. Since this coefficient decreases in Y, the model reflects the concept that cost becomes less important in a person's decision making, relative to other issues, when income rises.

When sociodemographic variables are interacted with attributes of the alternatives, there is no need to normalize the coefficients. The sociodemographic variables affect the differences in utility through their interaction with the attributes of the alternatives. The difference $U_c - U_b = \ldots \beta(M_c - M_b)/Y \ldots$ varies with income, since costs differ over alternatives.

Number of Independent Error Terms

As given by equation (2.2), the choice probabilities take the form

$$P_{ni} = \int_{\varepsilon} I(\varepsilon_{nj} - \varepsilon_{ni} < V_{ni} - V_{nj} \; \forall j \neq i) f(\varepsilon_n) \, d\varepsilon_n.$$

This probability is a J-dimensional integral over the density of the J error terms in $\varepsilon_n = \langle \varepsilon_{n1}, \ldots, \varepsilon_{nJ} \rangle$. The dimension can be reduced, however, through recognizing that only differences in utility matter. With J errors (one for each alternative), there are $J - 1$ error differences. The choice probability can be expressed as a $(J-1)$-dimensional integral over the density of these error differences:

$$\begin{aligned} P_{ni} &= \text{Prob}(U_{ni} > U_{nj} \; \forall j \neq i) \\ &= \text{Prob}(\varepsilon_{nj} - \varepsilon_{ni} < V_{ni} - V_{nj} \; \forall j \neq i) \\ &= \text{Prob}(\tilde{\varepsilon}_{nji} < V_{ni} - V_{nj} \; \forall j \neq i) \\ &= \int I(\tilde{\varepsilon}_{nji} < V_{ni} - V_{nj} \; \forall j \neq i) g(\tilde{\varepsilon}_{ni}) \, d\tilde{\varepsilon}_{ni} \end{aligned}$$

where $\tilde{\varepsilon}_{nji} = \varepsilon_{nj} - \varepsilon_{ni}$ is the difference in errors for alternatives i and j; $\tilde{\varepsilon}_{ni} = \langle \tilde{\varepsilon}_{n1i}, \ldots, \tilde{\varepsilon}_{nJi} \rangle$ is the $(J-1)$-dimensional vector of error differences, with the $\ldots$ over all alternatives except i; and $g(\cdot)$ is the density of these error differences. Expressed in this way, the choice probability is a $(J-1)$-dimensional integral.

The density of the error differences $g(\cdot)$, and the density of the original errors, $f(\cdot)$, are related in a particular way. Suppose a model is specified with an error for each alternative: $\varepsilon_n = \langle \varepsilon_{n1}, \ldots, \varepsilon_{nJ} \rangle$ with density $f(\varepsilon_n)$. This model is equivalent to a model with $J-1$ errors defined as $\tilde{\varepsilon}_{njk} = \varepsilon_{nj} - \varepsilon_{nk}$ for any k and density $g(\tilde{\varepsilon}_{nk})$ derived from $f(\varepsilon_n)$. For any $f(\varepsilon_n)$, the corresponding $g(\tilde{\varepsilon}_{nk})$ can be derived. However, since ε_n has more elements than $\tilde{\varepsilon}_{nk}$, there is an infinite number of densities for the J error terms that give the same density for the $J-1$ error differences. Stated equivalently, any $g(\tilde{\varepsilon}_{nk})$ is consistent with an infinite number of different $f(\varepsilon_n)$'s. Since choice probabilities can always be expressed as depending only on $g(\tilde{\varepsilon}_{nk})$, one dimension of the density of $f(\varepsilon_n)$ is not identified and must be normalized by the researcher.

The normalization of $f(\varepsilon_n)$ can be handled in various ways. For some models, such as logit, the distribution of the error terms is sufficiently restrictive that the normalization occurs automatically with the assumptions on the distribution. For other models, such as probit, identification is often obtained by specifying the model only in terms of error differences, that is, by parameterizing $g(\cdot)$ without reference to $f(\cdot)$. In all but the simplest models, the researcher needs to consider the fact that only the density of error differences affects the probabilities and therefore is identified. In discussing the various models in subsequent chapters, we will return to this issue and how to handle it.

2.5.2. The Overall Scale of Utility Is Irrelevant

Just as adding a constant to the utility of all alternatives does not change the decision maker's choice, neither does multiplying each alternative's utility by a constant. The alternative with the highest utility is the same no matter how utility is scaled. The model $U_{nj}^0 = V_{nj} + \varepsilon_{nj} \; \forall j$ is equivalent to $U_{nj}^1 = \lambda V_{nj} + \lambda \varepsilon_{nj} \; \forall j$ for any $\lambda > 0$. To take account of this fact, the researcher must normalize the scale of utility.

The standard way to normalize the scale of utility is to normalize the variance of the error terms. The scale of utility and the variance of the error terms are definitionally linked. When utility is multiplied by λ, the variance of each ε_{nj} changes by λ^2: $\text{Var}(\lambda \varepsilon_{nj}) = \lambda^2 \text{ Var}(\varepsilon_{nj})$. Therefore normalizing the variance of the error terms is equivalent to normalizing the scale of utility.

Normalization with iid Errors

If the error terms are assumed to be independently, identically distributed (iid), then the normalization for scale is straightforward. The researcher normalizes the error variance to some number, which is usually chosen for convenience. Since all the errors have the same variance by assumption, normalizing the variance of any of them sets the variance for them all.

When the observed portion of utility is linear in parameters, the normalization provides a way of interpreting coefficients. Consider the model $U^0_{nj} = x'_{nj}\beta + \varepsilon^0_{nj}$ where the variance of the error terms is $\text{Var}(\varepsilon^0_{nj}) = \sigma^2$. Suppose the research normalizes the scale by setting the error variance to 1. The original model becomes the following equivalent specification: $U^1_{nj} = x'_{nj}(\beta/\sigma) + \varepsilon^1_{nj}$ with $\text{Var}(\varepsilon^1_{nj}) = 1$. The original coefficients β are divided by the standard deviation of the unobserved portion of utility. The new coefficients β/σ reflect, therefore, the effect of the observed variables *relative* to the standard deviation of the unobserved factors.

The same concepts apply for whatever number the researcher chooses for normalization. As we will see in the next chapter, the error variances in a standard logit model are traditionally normalized to $\pi^2/6$, which is about 1.6. In this case, the preceding model becomes $U_{nj} = x'_{nj}(\beta/\sigma)\sqrt{1.6} + \varepsilon_{nj}$ with $\text{Var}(\varepsilon_{nj}) = 1.6$. The coefficients still reflect the variance of the unobserved portion of utility. The only difference is that the coefficients are larger by a factor of $\sqrt{1.6}$.

While it is immaterial which number is used by the researcher for normalization, interpretation of model results must take the normalization into consideration. Suppose, for example, that a logit and an independent probit model were both estimated on the same data. As stated earlier, the error variance is normalized to 1.6 for logit. Suppose the researcher normalized the probit to have error variances of 1, which is traditional with independent probits. This difference in normalization must be kept in mind when comparing estimates from the two models. In particular, the coefficients in the logit model will be $\sqrt{1.6}$ times larger than those for the probit model, simply due to the difference in normalization. If the researcher does not take this scale difference into account when comparing the models, she might inadvertently think that the logit model implies that people care more about the attributes (since the coefficients are larger) than implied by the probit model. For example, in a mode choice model, suppose the estimated cost coefficient is -0.55 from a logit model and -0.45 from an independent probit model. It is incorrect to say that the logit model implies more sensitivity to costs than the probit model. The coefficients in one of the models must be

adjusted to account for the difference in scale. The logit coefficients can be divided by $\sqrt{1.6}$, so that the error variance is 1, just as in the probit model. With this adjustment, the comparable coefficients are -0.43 for the logit model and -0.45 for the probit model. The logit model implies less price sensitivity than the probit. Instead, the probit coefficients could be converted to the scale of the logit coefficients by multiplying them by $\sqrt{1.6}$, in which case the comparable coefficients would be -0.55 for logit and -0.57 for probit.

A similar issue of interpretation arises when the same model is estimated on different data sets. The relative scale of the estimates from the two data sets reflects the relative variance of unobserved factors in the data sets. Suppose mode choice models were estimated in Chicago and Boston. For Chicago, the estimated cost coefficient is -0.55 and the estimated coefficient of time is -1.78. For Boston, the estimates are -0.81 and -2.69. The ratio of the cost coefficient to the time coefficient is very similar in the two cities: 0.309 in Chicago and 0.301 in Boston. However, the scale of the coefficients is about fifty percent higher for Boston than for Chicago. This scale difference means that the unobserved portion of utility has less variance in Boston than in Chicago: since the coefficients are divided by the standard deviation of the unobserved portion of utility, lower coefficients mean higher standard deviation and hence variance. The models are revealing that factors other than time and cost have less effect on people in Boston than in Chicago. Stated more intuitively, time and cost have more importance, relative to unobserved factors, in Boston than in Chicago, which is consistent with the larger scale of the coefficients for Boston.

Normalization with Heteroskedastic Errors

In some situations, the variance of the error terms can be different for different segments of the population. The researcher cannot set the overall level of utility by normalizing the variance of the errors for all segments, since the variance is different in different segments. Instead, the researcher sets the overall scale of utility by normalizing the variance for one segment, and then estimates the variance (and hence scale) for each segment relative to this one segment.

For example, consider the situation described in the previous section, where the unobserved factors have greater variance in Chicago than in Boston. If separate models are estimated for Chicago and Boston, then the variance of the error term is normalized separately for each model. The scale of the parameters in each model reflects the variance of unincluded factors in that area. Suppose, however, that the researcher wants to estimate a model on data for both Chicago and Boston. She

cannot normalize the variance of the unobserved factors for all travelers to the same number, since the variance is different for travelers in Boston than for those in Chicago. Instead, the researcher sets the overall scale of utility by normalizing the variance in one area (say Boston) and then estimates the variance in the other area *relative* to that in the first area (the variance in Chicago relative to that in Boston).

The model in its original form is

$$U_{nj} = \alpha T_{nj} + \beta M_{nj} + \varepsilon^B_{nj} \ \forall n \text{ in Boston}$$
$$U_{nj} = \alpha T_{nj} + \beta M_{nj} + \varepsilon^C_{nj} \ \forall n \text{ in Chicago,}$$

where the variance of ε^B_{nj} is not the same as the variance of ε^C_{nj}. Label the ratio of variances as $k = \text{Var}(\varepsilon^C_{nj})/\text{Var}(\varepsilon^B_{nj})$. We can divide the utility for travelers in Chicago by $\sqrt{k}$; this division doesn't affect their choices, of course, since the scale of utility doesn't matter. However, doing so allows us to rewrite the model as

$$U_{nj} = \alpha T_{nj} + \beta M_{nj} + \varepsilon_{nj} \ \forall n \text{ in Boston}$$
$$U_{nj} = (\alpha/\sqrt{k})T_{nj} + (\beta/\sqrt{k})M_{nj} + \varepsilon_{nj} \ \forall n \text{ in Chicago,}$$

where now the variance of ε_{nj} is the same for all n in both cities (since $\text{Var}(\varepsilon^C_{nj}/\sqrt{k}) = (1/k)\text{Var}(\varepsilon^C_{nj}) = [\text{Var}(\varepsilon^B_{nj})/\text{Var}(\varepsilon^C_{nj})]\text{Var}(\varepsilon^C_{nj}) = \text{Var}(\varepsilon^B_{nj})$. The scale of utility is set by normalizing the variance of ε_{nj}. The parameter k, which is often called the scale parameter, is estimated along with β and α. The estimated value $\hat{k}$ of k tells the researcher the variance of unobserved factors in Chicago relative to that in Boston. For example, $\hat{k} = 1.2$ implies that the variance of unobserved factors is twenty percent greater in Chicago than in Boston.

The variance of the error term can differ over geographic regions, data sets, time, or other factors. In all cases, the researcher sets the overall scale of utility by normalizing one of the variances and then estimating the other variances relative to the normalized one. Swait and Louviere (1993) discuss the role of the scale parameter in discrete choice models, describing the variety of reasons that variances can differ over observations. As well as the traditional concept of variance in unobserved factors, psychological factors can come into play, depending on the choice situation and the interpretation of the researcher. For example, Bradley and Daly (1994) allow the scale parameter to vary over stated preference experiments in order to allow for respondents' fatigue in answering the survey questions. Ben-Akiva and Morikawa (1990) allow the scale parameter to differ for respondents' stated intentions versus their actual market choices.

Normalization with Correlated Errors

In the discussion so far we have assumed that ε_{nj} is independent over alternatives. When the errors are correlated over alternatives, normalizing for scale is more complex. We have talked in terms of setting the scale of utility. However, since only differences in utility matter, it is more appropriate to talk in terms of setting the scale of utility *differences*. When errors are correlated, normalizing the variance of the error for one alternative is not sufficient to set the scale of utility differences.

The issue is most readily described in terms of a four-alternative example. The utility for the four alternatives is $U_{nj} = V_{nj} + \varepsilon_{nj}$, $j = 1, \ldots, 4$. The error vector $\varepsilon_n = \langle \varepsilon_{n1}, \ldots, \varepsilon_{n4} \rangle$ has zero mean and covariance matrix

$$\Omega = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \cdot & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \cdot & \cdot & \sigma_{33} & \sigma_{34} \\ \cdot & \cdot & \cdot & \sigma_{44} \end{pmatrix}, \tag{2.3}$$

where the dots refer to the corresponding elements in the upper part of the symmetric matrix.

Since only differences in utility matter, this model is equivalent to one in which all utilities are differenced from, say, the first alternative. The equivalent model is $\tilde{U}_{nj1} = \tilde{V}_{nj1} - \tilde{\varepsilon}_{nj1}$ for $j = 2, 3, 4$, where $\tilde{U}_{nj1} = U_{nj} - U_{n1}$, $\tilde{V}_{nj1} = V_{nj} - V_{n1}$, and the vector of error differences is $\tilde{\varepsilon}_{n1} = \langle (\varepsilon_{n2} - \varepsilon_{n1}), (\varepsilon_{n3} - \varepsilon_{n1}), (\varepsilon_{n4} - \varepsilon_{n1}) \rangle$. The variance of each error difference depends on the variances and covariances of the original errors. For example, the variance of the difference between the first and second errors is $\text{Var}(\tilde{\varepsilon}_{n21}) = \text{Var}(\varepsilon_{n2} - \varepsilon_{n1}) = \text{Var}(\varepsilon_{n1}) + \text{Var}(\varepsilon_{n2}) - 2\,\text{Cov}(\varepsilon_{n1}, \varepsilon_{n2}) = \sigma_{11} + \sigma_{22} - 2\sigma_{12}$. We can similarly calculate the covariance between $\tilde{\varepsilon}_{n21}$, which is the difference between the first and second errors, and $\tilde{\varepsilon}_{n31}$, which is the difference between the first and third errors: $\text{Cov}(\tilde{\varepsilon}_{n21}, \tilde{\varepsilon}_{n31}) = E(\varepsilon_{n2} - \varepsilon_{n1})(\varepsilon_{n3} - \varepsilon_{n1}) = E(\varepsilon_{n2}\varepsilon_{n3} - \varepsilon_{n2}\varepsilon_{n1} - \varepsilon_{n3}\varepsilon_{n1} + \varepsilon_{n1}\varepsilon_{n1}) = \sigma_{23} - \sigma_{21} - \sigma_{31} + \sigma_{11}$. The covariance matrix for the vector of error differences becomes

$$\tilde{\Omega}_1 = \begin{pmatrix} \sigma_{11} + \sigma_{22} - 2\sigma_{12} & \sigma_{11} + \sigma_{23} - \sigma_{12} - \sigma_{13} & \sigma_{11} + \sigma_{24} - \sigma_{12} - \sigma_{14} \\ \cdot & \sigma_{11} + \sigma_{33} - 2\sigma_{13} & \sigma_{11} + \sigma_{34} - \sigma_{13} - \sigma_{14} \\ \cdot & \cdot & \sigma_{11} + \sigma_{44} - 2\sigma_{14} \end{pmatrix}.$$

Setting the variance of one of the original errors is not sufficient to set the variance of the error differences. For example, if the variance for the first alternative is set to some number $\sigma_{11} = k$, the variance of the difference between the errors for the first two alternatives becomes

$k + \sigma_{22} - \sigma_{12}$. An infinite number of values for $\sigma_{22} - \sigma_{12}$ provide equivalent models.

A common way to set the scale of utility when errors are not iid is to normalize the variance of one of the error differences to some number. Setting the variance of an error difference sets the scale of utility differences and hence of utility. Suppose we normalize the variance of $\tilde{\varepsilon}_{n21}$ to 1. The covariance matrix for the error differences, expressed in terms of the covariances of the original errors, becomes

$$(2.4) \quad \begin{pmatrix} 1 & (\sigma_{11}+\sigma_{23}-\sigma_{12}-\sigma_{13})/m & (\sigma_{11}+\sigma_{24}-\sigma_{12}-\sigma_{14})/m \\ \cdot & (\sigma_{11}+\sigma_{33}-2\sigma_{13})/m & (\sigma_{11}+\sigma_{34}-\sigma_{13}-\sigma_{14})/m \\ \cdot & \cdot & (\sigma_{11}+\sigma_{44}-2\sigma_{14})/m \end{pmatrix},$$

where $m = \sigma_{11} + \sigma_{22} - 2\sigma_{12}$. Utility is divided by $\sqrt{\sigma_{11} + \sigma_{22} - 2\sigma_{12}}$ to obtain this scaling.

Note that when the error terms are iid, normalizing the variance of one of these errors automatically normalizes the variance of the error differences. With iid errors, $\sigma_{jj} = \sigma_{ii}$ and $\sigma_{ij} = 0$ for $i \neq j$. Therefore, if σ_{11} is normalized to k, then the variance of the error difference becomes $\sigma_{11} + \sigma_{22} - 2\sigma_{12} = k + k - 0 = 2k$. The variance of the error difference is indeed being normalized, the same as with non-iid errors.

Normalization has implications for the number of parameters that can be estimated in the covariance matrix. The covariance of the original errors, Ω in equation (2.3), has ten elements in our four-alternative example. However, the covariance matrix of the error differences has six elements, one of which is normalized to set the scale of utility differences. The covariance matrix for error differences with the variance of the first error difference normalized to k takes the form

$$(2.5) \quad \tilde{\Omega}_1^* = \begin{pmatrix} k & \omega_{ab} & \omega_{ac} \\ \cdot & \omega_{bb} & \omega_{bc} \\ \cdot & \cdot & \omega_{cc} \end{pmatrix},$$

which has only five parameters. On recognizing that only differences matter and that the scale of utility is arbitrary, the number of covariance parameters drops from ten to five. A model with J alternatives has at most $J(J-1)/2 - 1$ covariance parameters after normalization.

Interpretation of the model is affected by the normalization. Suppose for example that the elements of matrix (2.5) were estimated. The parameter ω_{bb} is the variance of the difference between the errors for the first and third alternatives *relative* to the variance of the difference between the errors for the first and second alternatives. Complicating interpretation even further is the fact that the variance of the difference between

the errors for two alternatives reflects the variances of both as well as their covariance.

As we will see, the normalization of logit and nested logit models is automatic with the distributional assumptions that are placed on the error terms. Interpretation under these assumptions is relatively straightforward. For mixed logit and probit, fewer assumptions are placed on the distribution of error terms, so that normalization is not automatic. The researcher must keep the normalization issues in mind when specifying and interpreting a model. We return to this topic when discussing each discrete choice model in subsequent chapters.

2.6 Aggregation

Discrete choice models operate at the level of individual decision makers. However, the researcher is usually interested in some aggregate measure, such as the average probability within a population or the average response to a change in some factor.

In linear regression models, estimates of aggregate values of the dependent variable are obtained by inserting aggregate values of the explanatory variables into the model. For example, suppose h_n is housing expenditures of person n, y_n is the income of the person, and the model relating them is $h_n = \alpha + \beta y_n$. Since this model is linear, the average expenditure on housing is simply calculated as $\alpha + \beta \bar{y}$, where $\bar{y}$ is average income. Similarly, the average response to a one-unit change in income is simply β, since β is the response for each person.

Discrete choice models are not linear in explanatory variables, and consequently, inserting aggregate values of the explanatory variables into the models will not provide an unbiased estimate of the average probability or average response. The point can be made visually. Consider Figure 2.1, which gives the probabilities of choosing a particular alternative for two individuals with the observed portion of their utility (their *representative utility*) being a and b. The average probability is the average of the probabilities for the two people, namely, $(P_a + P_b)/2$. The average representative utility is $(a + b)/2$, and the probability evaluated at this average is the point on the curve above $(a + b)/2$. As shown for this case, the average probability is greater than the probability evaluated at the average representative utility. In general, the probability evaluated at the average representative utility underestimates the average probability when the individuals' choice probabilities are low and overestimates when they are high.

Estimating the average response by calculating derivatives and elasticities at the average of the explanatory variables is similarly problematic.

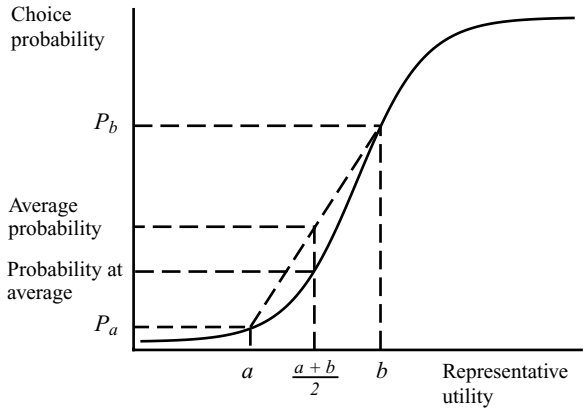

Figure 2.1. Difference between average probability and probability calculated at average representative utility.

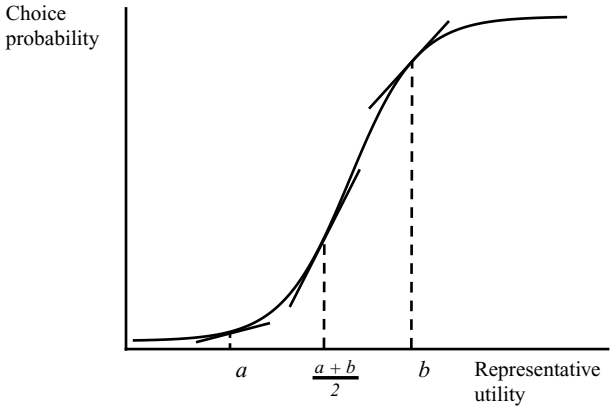

Figure 2.2. Difference between average response and response calculated at average representative utility.

Consider Figure 2.2, depicting two individuals with representative utilities a and b. The derivative of the choice probability for a change in representative utility is small for both of these people (the slope of the curve above a and b). Consequently, the average derivative is also small. However, the derivative at the average representative utility is very large (the slope above $(a + b)/2$). Estimating the average response in this way can be seriously misleading. In fact, Talvitie (1976) found, in a mode choice situation, that elasticities at the average representative utility can be as much as two or three times greater or less than the average of the individual elasticities.

Aggregate outcome variables can be obtained consistently from discrete choice models in two ways, by sample enumeration or segmentation. We discuss each approach in the following sections.

2.6.1. Sample Enumeration

The most straightforward, and by far the most popular, approach is sample enumeration, by which the choice probabilities of each decision maker in a sample are summed, or averaged, over decision makers. Consider a discrete choice model that gives probability P_{ni} that decision maker n will choose alternative i from a set of alternatives. Suppose a sample of N decision makers, labeled $n = 1, \ldots, N$, is drawn from the population for which aggregate statistics are required. (This sample might be the sample on which the model was estimated. However, it might also be a different sample, collected in a different area or at a later date than the estimation sample.) Each sampled decision maker n has some weight associated with him, w_n, representing the number of decision makers similar to him in the population. For samples based on exogenous factors, this weight is the reciprocal of the probability that the decision maker was selected into the sample. If the sample is purely random, then w_n is the same for all n; and if the sample is stratified random, then w_n is the same for all n within a stratum.

A consistent estimate of the total number of decision makers in the population who choose alternative i, labeled $\hat{N}_i$, is simply the weighted sum of the individual probabilities:

$$\hat{N}_i = \sum_n w_n P_{ni}.$$

The average probability, which is the estimated market share, is $\hat{N}_i/N$. Average derivatives and elasticities are similarly obtained by calculating the derivative and elasticity for each sampled person and taking the weighted average.

2.6.2. Segmentation

When the number of explanatory variables is small, and those variables take only a few values, it is possible to estimate aggregate outcomes without utilizing a sample of decision makers. Consider, for example, a model with only two variables entering the representative utility of each alternative: education level and gender. Suppose the education variable consists of four categories: did not complete high school, completed high school but did not attend college, attended college but

did not receive a degree, received a college degree. Then the total number of different types of decision makers (called *segments*) is eight: the four education levels for each of the two genders. Choice probabilities vary only over these eight segments, not over individuals within each segment.

If the researcher has data on the number of people in each segment, then the aggregate outcome variables can be estimated by calculating the choice probability for each segment and taking the weighted sum of these probabilities. The number of people estimated to choose alternative i is

$$\hat{N}_i = \sum_{s=1}^{8} w_s P_{si},$$

where P_{si} is the probability that a decision maker in segment s chooses alternative i, and w_s is the number of decision makers in segment s.

2.7 Forecasting

For forecasting into some future year, the procedures described earlier for aggregate variables are applied. However, the exogenous variables and/or the weights are adjusted to reflect changes that are anticipated over time. With sample enumeration, the sample is adjusted so that it *looks like* a sample that would be drawn in the future year. For example, to forecast the number of people who will choose a given alternative five years in the future, a sample drawn from the current year is adjusted to reflect changes in socioeconomic and other factors that are expected to occur over the next five years. The sample is adjusted by (1) changing the value of the variables associated with each sampled decision maker (e.g., increasing each decision maker's income to represent real income growth over time), and/or (2) changing the weight attached to each decision maker to reflect changes over time in the number of decision makers in the population that are similar to the sampled decision maker (e.g., increasing the weights for one-person households and decreasing weights for large households to reflect expected decreases in household size over time).

For the segmentation approach, changes in explanatory variables over time are represented by changes in the number of decision makers in each segment. The explanatory variables themselves cannot logically be adjusted, since the distinct values of the explanatory variables define the segments. Changing the variables associated with a decision maker in one segment simply shifts the decision maker to another segment.

2.8 Recalibration of Constants

As described in Section 2.5.1, alternative-specific constants are often included in a model to capture the average effect of unobserved factors. In forecasting, it is often useful to adjust these constants, to reflect the fact that unobserved factors are different for the forecast area or year compared to the estimation sample. Market-share data from the forecast area can be used to *recalibrate* the constants appropriately. The recalibrated model can then be used to predict changes in market shares due to changes in explanatory factors.

An iterative process is used to recalibrate the constants. Let α_j^0 be the estimated alternative-specific constant for alternative j. The superscript 0 is used to indicate that these are the starting values in the iterative process. Let S_i denote the share of decision makers in the forecast area that choose alternative j in the *base* year (usually, the latest year for which such data are available.) Using the discrete choice model with its original values of $\alpha_j^0 \; \forall j$, predict the share of decision makers in the forecast area who will choose each alternative. Label these predictions $\hat{S}_j^0 \; \forall j$. Compare the predicted shares with the actual shares. If the actual share for an alternative exceeds the predicted share, raise the constant for that alternative. Lower the constant if the actual share is below the predicted. An effective adjustment is

$$\alpha_j^1 = \alpha_j^0 + \ln\left(S_j / \hat{S}_j^0\right).$$

With the new constants, predict the share again, compare with the actual shares, and if needed adjust the constants again. The process is repeated until the forecasted shares are sufficiently close to the actual shares. The model with these recalibrated constants is then used to predict changes from base-year shares due to changes in observed factors that affect decision makers' choices.