Fundamentals of Economic Demand Modeling: Lessons from Travel Demand Analysis

Kenneth A. Small
University of California at Irvine

REVISED July 4, 2005

Forthcoming as Chapter 9 of:

Decision-Based Design: Making Effective Decisions in Product and Systems Design, Wei Chen, Kemper Lewis, and Linda C. Schmidt, editors, ASME Press.

Abstract

This chapter presents essential methods developed in transportation economics and travel demand analysis, and describes how they are used to measure the economic value that consumers place on product characteristics. The chapter focuses primarily on discrete-choice models estimated using data on individual consumers. It develops such models from a random utility framework, derives the familiar probit and logit models including multinomial logit, and provides examples of models developed for real transportation analysis. The chapter also presents more advanced discrete-choice models including generalized extreme value and mixed logit. Recent developments in stated-preference data and panel data are described. The use of such models for design is illustrated by a theoretical and empirical review of the value of time and reliability in urban travel.

Acknowledgment

This chapter is adapted from Chapter 2 of *Urban Transportation Economics*, by Kenneth A. Small and Erik Verhoef, 2nd edition, Routledge, forthcoming 2005. (First edition © 1992 by Harwood Academic Publishers; reprinted with permission.) The work has benefited from comments by Alex Anas, Richard Arnott, David Brownstone, Marc Gaudry, Amihai Glazer, David Hensher, Sergio Jara-Díaz, Charles Lave, Kenneth Train, and Clifford Winston. All responsibility for accuracy and interpretation lies with the author.

Nomenclature

cdf	Cumulative distribution function
D^{i}	Alternative-specific dummy variable for alternative *i*
d_{jn}	Choice variable (=1 if decision-maker *n* chooses alternative *j*)
E	Expectation
G	Function for generating GEV models of discrete choice
GEV	Generalized extreme value
I_{r}	Inclusive value for alternative group *r*
iid	Identically and independently distributed
J	Number of dependent variables (aggregate models) or alternatives (disaggregate models)
J_{r}	Number of alternatives in alternative group *r*
L	Leisure
$L(\cdot)$	Log-likelihood function
log	Natural logarithm
N	Number of people; Number of vehicles in queue
n	indexes single individual consumer
P	Choice probability
P_L	Probability of being late
R	Number of "replications" (random draws) in simulated probability
s_{n}	Vector of socioeconomic or other characteristics of decision-maker *n*
SDE	Schedule delay early = work start time minus early actual arrival time
SDL	Schedule delay late = late actual arrival time minus work start time
T	Time spent in activities; Travel Time (usually in-vehicle) if used as scalar without sub- or super-scripts
T^0	Out-of-vehicle travel time
T_F	Free-flow travel time
T_R	Random component of travel time
T_{w}	Time spent at work (in value-of-time analysis)
t	Time of arrival
t^*	Desired time of arrival
t_d	Departure time
U	Utility function
V	Indirect utility function
v_R	Value of reliability
v_{T}	Value of time (usually in-vehicle time)

W	Welfare measure
w	Wage rate (in travel-demand analysis);
X	Generalized consumption good (numeraire)
x	Consumption vector
Y	Unearned income
y	Generalized argument for function G generating GEV models
z	Independent variables for travel-demand models
α_i	Alternative-specific constant for alternative i in discrete-choice indirect utility function; value of travel time in reliability analysis
β	Parameter vector in discrete-choice indirect utility function (in travel-demand analysis); value of early-arrival time (in reliability analysis)
γ	Value of late-arrival time (in reliability analysis)
γ_i	Coefficient of an independent variable interacted with an alternative-specific constant for alternative i in discrete-choice utility function
ε_i	Stochastic term for alternative i in discrete-choice indirect utility function
θ	Value of late arrival (in reliability analysis)
λ	Marginal utility of income
μ	Scale parameter for probability density function (in discrete-choice analysis)
ρ	Parameter of GEV functions (in discrete-choice analysis)

Fundamentals of Economic Demand Modeling: Lessons from Travel Demand Analysis

Kenneth A. Small
University of California at Irvine

In order to design facilities or products, one must know how and under what circumstances they will be used. In order to design them cost-effectively, one must also know how specific features are valued by users. These requirements can be met by applying tools of economic demand analysis.

This chapter illustrates the use of economic demand tools by providing a detailed account of their use in the field of urban transportation, where many of them were in fact first developed. Transportation applications include many design decisions, often at the level of an entire system such as a public mass transit network. The design elements that can be addressed using transportation demand analysis include speed, frequency, reliability, comfort, and ability to match people's desired schedules.

The chapter begins (Section 9.1) with conventional aggregate approach to economic demand, and then moves to disaggregate models (Section 9.2), also known as "behavioral" because they depict decision-making process by individual consumers. Some specific examples (Section 9.3) and more advanced topics (Sections 2.4) are then discussed. Finally, Section 9.5 analyzes two behavioral results of travel-demand studies that are especially important for design, namely travelers' willingness to pay for travel-time savings and improved reliability.

9.1 Aggregate Models

In standard economic analysis of consumer demand, the aggregate demand for some product is explained as a function of variables that describe the product or its consumers. For example, total transit demand in a city might be related to the amounts of residential and industrial development, the average transit fare, the costs of alternative modes, some simple measures of service quality, and average income.

In one type of study, known as *cross-sectional*, one examines influences on travel behavior by looking at variations across cities or across locations within a city. An example is the analysis by Kain and Liu (2002) of mode share to work in Santiago, Chile. The share is measured for each of 34 districts and its logarithm is regressed on variables such as travel time, transit availability, and household income.

Sometimes there are many cases of zero reported trips by a given mode between a pair of zones, making ordinary regression analysis invalid. This illustrates a pervasive feature of travel-demand analysis: many of the variables to be explained have a limited range. For this reason, travel-demand researchers have contributed importantly to the development of techniques to handle limited dependent variables (McFadden, 2001). We note here one such technique that is applicable to aggregate data.

Suppose the dependent variable of a model can logically take values only within a certain range. For example, if the dependent variable x is the modal share of transit, it must lie between zero and one. Instead of explaining x directly, we can explain the logistic transformation of x as follows:

$$\log\left(\frac{x}{1-x}\right) = \beta' z + \varepsilon \tag{9.1}$$

where β is a vector of parameters, z is a vector of independent variables, and ε is an error term with infinite range. Equivalently,

$$x = \frac{\exp(\beta' z + \varepsilon)}{1 + \exp(\beta' z + \varepsilon)}. \tag{9.2}$$

This is an *aggregate logit* model for a single dependent variable.

In many applications, several dependent variables x_i are related to each other, each associated with particular values z_i of some independent variables. For example, x_i might be the share of trips made by mode i, and z_i a vector of service characteristics of mode i. A simple extension of equation (9.2) ensures that the shares sum to one:

$$x_i = \frac{\exp(\beta' z_i + \varepsilon_i)}{\sum_{j=1}^{J} \exp(\beta' z_j + \varepsilon_j)} \tag{9.3}$$

where J is the number of modes.

In another type of study, known as *time-series*, one looks instead at variations over time within a single area. Several studies have examined transit ridership using data over time from a single metropolitan area or even a single transit corridor – for example Gómez-Ibáñez (1996) for Boston. Time-series studies are sensitive to the tendency for unobserved influences to persist over time, a situation known as auto-correlation in the error term. One may also postulate "inertia" by including among the explanatory variables one or more lagged values of the variable being explained. For example, Greene (1992), using US nationwide data, considers the possibility that once people have established the travel patterns resulting in a particular level of vehicle-miles traveled, they change them only gradually if conditions such as fuel prices suddenly change. From the coefficients on the lagged dependent variables, one can ascertain the difference between short- and long-run responses.

It is common to combine cross-sectional and time-series variation, so that individual consumers analysis units are observed repeatedly over time. The resulting data are called *panel data* or *longitudinal data* (Kitamura, 2000). For example, Voith (1997) analyzes ridership data from 118 commuter-rail stations in metropolitan Philadelphia over the years 1978–91 to ascertain the effects of level of service and of demographics on rail ridership. Studies using panel data need to account for the fact that, even aside from autocorrelation, the error terms for observations from the same location at different points in time cannot plausibly be assumed to be independent. Neglecting this fact will result in unnecessarily imprecise and possibly biased estimates. Several approaches are available to account for this panel structure, the most popular being to estimate a "fixed effects" model, in which a separate constant is estimated for every location.

9.2 Disaggregate Models

An alternative approach, known as *disaggregate* or *behavioral* travel-demand modeling, is now far more common for travel demand research. Made possible by micro data (data on individual consumers), this approach explains behavior directly at the level of a person, household, or firm. When survey data are available, disaggregate models are statistically more efficient in using such data because disaggregate models take account of every observed choice rather than just aggregate shares; this enables them to take advantage of variation in behavior across individuals that may be correlated with variation in individual conditions, whereas such variations are obscured in aggregate statistics. Disaggregate models are also based on a more satisfactory microeconomic theory of demand. Most such models analyze choices among discrete rather than continuous alternatives and so are called *discrete-choice models*. Train (2003) provides a thorough treatment.

9.2.1 Basic Discrete-Choice Models

The most widely used theoretical foundation for these models is the additive random-utility model of McFadden (1973). Suppose a consumer *n* facing discrete alternatives *j*=1,...,*J* chooses the one that maximizes utility as given by

$$U_{jn} = V\left(z_{jn}, s_n, \beta\right) + \varepsilon_{jn} \tag{9.4}$$

where $V(\cdot)$ is a function known as the *systematic utility*, z_{jn} is a vector of attributes of the alternatives as they apply to this consumer, s_n is a vector of characteristics of the consumer (effectively allowing different utility structures for different groups of consumers), β is a vector of unknown parameters, and ε_{jn} is an unobservable component of utility which captures idiosyncratic preferences. U_{jn} and $V(\cdot)$ implicitly incorporate a budget constraint, and thus are functions of income and prices as well as product quantities and attributes; in economics terminology, such a utility function is called *indirect* to distinguish it from the direct or primal dependence of preferences on those quantities and attributes. U_{jn} and $V(\cdot)$ are also conditional on choice *j*. For these reasons they are known as *conditional indirect utility* functions.

The choice is probabilistic because the measured variables do not include everything relevant to the individual's decision. This fact is represented by the random terms ε_{jn}. Once a functional form for *V* is specified, the model becomes complete by specifying a joint cumulative distribution function (cdf) for these random terms, $F(\varepsilon_{1n},...,\varepsilon_{Jn})$. Denoting $V(z_{jn},s_n,\beta)$ by V_{jn}, the choice probability for alternative *i* is then

$$\begin{aligned} P_{in} &= \Pr[U_{in} > U_{jn} \quad \text{for all } j \neq i] \\ &= \Pr[\varepsilon_{jn} - \varepsilon_{in} < V_{in} - V_{jn} \quad \text{for all } j \neq i] \\ &= \int_{-\infty}^{\infty} F_i(V_{in} - V_{1n} + \varepsilon_{in}, \ldots, V_{in} - V_{Jn} + \varepsilon_{in}) d\varepsilon_{in} \end{aligned} \tag{9.5}$$

where F_i is the partial derivative of *F* with respect to its *i*-th argument. (F_i is thus the probability density function of ε_{in} conditional on the inequalities in the middle row of (9.5).)

Suppose $F(\cdot)$ is multivariate normal. Then (9.5) is the *multinomial probit* model with general covariance structure. However, neither F nor F_i can be expressed in closed form; instead, equation (9.5) is usually written as a (J-1)-dimensional integral of the normal density function. In the special case where the random terms are identically and independently distributed (iid) with the univariate normal distribution, F is the product of J univariate normal cdfs, and we have the *iid probit* model, which still requires computation of a (J-1)-dimensional integral. For example, in the iid probit model for binary choice (J=2), (9.5) becomes

$$P_{1n} = \Phi\left(\frac{V_{1n} - V_{2n}}{\sigma}\right) \tag{9.6}$$

where Φ is the cumulative standard normal distribution function (a one-dimensional integral) and σ is the standard deviation of ε_{1n}-ε_{2n}. In equation (9.6), σ cannot be distinguished empirically from the scale of utility, which is arbitrary; for example, doubling σ has the same effect as doubling both V_1 and V_2. Hence it is conventional to normalize by setting σ=1.

The *logit* model (also known as multinomial logit or conditional logit) arises when the J random terms are iid with the extreme-value distribution (also known as Gumbel, Weibull, or double-exponential). This distribution is defined by

$$\Pr\left[\varepsilon_{jn} < x\right] = \exp\left(-e^{-\mu x}\right) \tag{9.7}$$

for all real numbers x, where μ is a scale parameter. Here the convention is to normalize by setting μ=1. With this normalization, McFadden (1973) shows that the resulting probabilities calculated from (9.5) have the logit form:

$$P_{in} = \frac{\exp(V_{in})}{\sum_{j=1}^{J} \exp(V_{jn})}. \tag{9.8}$$

This formula is easily seen to have the celebrated and restrictive property of *independence from irrelevant alternatives*: namely, that the odds ratio (P_{in}/P_{jn}) depends on the utilities V_{in} and V_{jn} but not on the utilities for any other alternatives. This property implies, for example, that adding a new alternative k (equivalent to increasing its systematic utility V_{kn} from -∞ to some finite value) will not affect the relative proportions of people using previously existing alternatives. It also implies that for a given alternative k, the cross-elasticities $\partial \log P_{jn} / \partial \log V_{kn}$ are identical for all $j \neq k$: hence if the attractiveness of alternative k is increased, the probabilities of all the other alternatives $j \neq k$ will be reduced by identical percentages. The binary form of (9.8) is:

$$P_{in} = \{1 + \exp[-(V_{1n} - V_{2n})]\}^{-1}.$$

It is really the iid assumption (identically and independently distributed error terms) that is restrictive, whether or not it entails independence of irrelevant alternatives. Hence there is no basis for the widespread belief that iid probit is more general than logit. In fact, the logit and iid probit models have been found empirically to give virtually identical results when normalized comparably (Horowitz, 1980).[1] Furthermore, both probit and logit may be generalized by defining non-iid distributions. In the probit case the generalization uses the multivariate normal distribution, whereas in the logit case it can take a number of forms to be discussed in Section 9.4.

As for the functional form of V, by far the most common is linear in unknown parameters β. Note that this form can easily be made nonlinear in *variables* just by specifying new variables equal to nonlinear functions of the original ones. For example, the utility on mode i of a traveler n with wage w_n facing travel costs c_{in} and times T_{in} could be:

$$V_{in} = \beta_1 \cdot (c_{in} / w_n) + \beta_2 T_{in} + \beta_3 T_{in}^2. \qquad (9.9)$$

This is non-linear in travel time and in wage rate. If we redefine z_{in} as the vector of all such combinations of the original variables (z_{in} and s_n in eqn 9.4), the linear-in-parameters specification is simply written as

$$V_{in} = \beta' z_{in} \qquad (9.10)$$

where β' is the transpose of column vector β.

[1] Comparable normalization is accomplished by dividing the logit coefficients by $\pi/\surd 3$ in order to give the utilities the same standard deviations in the two models. In both models, the choice probabilities depend on $(\beta/\sigma_\varepsilon)$, where σ_ε^2 is the variance of each of the random terms ε_{in}. In the case of probit, the variance of ε_{1n}-ε_{2n}, $2\sigma_\varepsilon^2$, is set to one by the conventional normalization; hence $\sigma_\varepsilon^{PROBIT}$ =1/√2. In the case of logit, the normalization μ=1 in equation (9.7) implies that ε_{in} has standard deviation $\sigma_\varepsilon^{LOGIT}$ =π/√6 (Hastings and Peacock, 1975, p. 60). Hence to make logit and iid probit comparable, the logit coefficients must be divided by $\sigma_\varepsilon^{LOGIT} / \sigma_\varepsilon^{PROBIT}$ = π/√3 = 1.814.

9.2.2 Estimation

For a given model, data on actual choices, along with traits z_{jn}, can be used to estimate the parameter vector β in (9.10) and to carry out statistical tests of the assumed error distribution and the assumed functional form of V. Parameters are usually estimated by maximizing the log-likelihood function:

$$L(\beta) = \sum_{n=1}^{N} \sum_{i=1}^{J} d_{in} \log P_{in}(\beta) \tag{9.11}$$

where N is the sample size. In this equation, d_{in} is the choice variable, defined as 1 if consumer n chooses alternative i and 0 otherwise, and $P_{in}(\beta)$ is the choice probability.

A correction to (9.11) is available for choice-based samples, i.e., those in which the sampling frequencies depend on the choices made. (For example, transportation mode choice might be estimated from data arising from roadside surveys and surveys taken on transit vehicles.) The correction simply multiplies each term in the second summation by the inverse of the sampling probability for that sample member (Manski and Lerman, 1977).

One of the major attractions of logit is the computational simplicity of its log-likelihood function, due to taking the logarithm of the numerator in equation (9.8). With V linear in β, the logit log-likelihood function is globally concave in β, so finding a local maximum assures finding the global maximum. Fast computer routines to do this are widely available.

It is possible that the likelihood function is unbounded in one of the coefficients, making it impossible to maximize. This happens if one includes a variable that is a perfect predictor of choice within the sample. For example, suppose one is predicting car ownership (yes or no) and wants to include among variables s_n in (9.4) a dummy variable for high income. If it happens that within the sample everyone with high income owns a car, the likelihood function increases without limit in the coefficient of this dummy variable. We might solve the problem by respecifying the model with more broadly defined income groups or more narrowly defined alternatives. Alternatively, we could postulate a *linear probability model*, in which probability rather than utility is a linear function of coefficients; this model has certain statistical disadvantages but is simple and may be adequate with large samples.

9.2.3 Data

Some of the most important variables for travel demand modeling are determined endogenously within a larger model of which the demand model is just one component. With aggregate data, the endogeneity of travel characteristics is an important issue for obtaining valid statistical estimates. Fortunately, endogeneity can usually be ignored when using disaggregate data because, from the point of view of the individual consumer, the travel environment does not depend appreciably on that one individual's choices.

Nevertheless, measuring the values of attributes z_{in}, which typically vary by alternative, is more difficult than it may first appear. How does one know the traits that a traveler would have encountered on an alternative that was not in fact used? One possibility is to use objective estimates, such as the *engineering values* produced by network models of the transportation system. Another is to use *reported values* obtained directly from survey respondents. Each is subject to problems. Reported values measure people's perceptions of travel conditions, which, even for alternatives they choose regularly, may differ from the measures employed in policy analysis or forecasting. Worse still, reported values may be systematically biased so as to justify the choice, thereby exaggerating the advantages of the alternative chosen and the disadvantages of other alternatives.

The data described thus far measure information about *revealed preferences* (RP), those reflected in actual choices. There is growing interest in using *stated preference* (SP) data, based on responses to hypothetical situations (Hensher, 1994). SP data permit more control over the ranges of and correlations among the independent variables, and they also can elicit information about potential travel options not now available. It is still an open question how accurately they described what people really do. This is a very common dilemma in studies intended for use in engineering design, which have no choice but to rely on SP data if they concern product characteristics not available in actual situations.

It is possible combine data from both revealed and stated preferences in a single estimation procedure in order to take advantage of the strengths of each (Louviere and Hensher, 2001). So long as observations are independent of each other, the log-likelihood functions simply add. To prevent SP survey bias from contaminating inferences from RP, it is recommended to estimate certain parameters separately in the two portions of the data: namely, the scale factors μ for the two parts of the sample (with one but not both normalized), any alternative-specific constants, and any critical behavioral coefficients that may differ. The log-likelihood function

(9.11) then breaks into two terms, one for RP observations and one for SP observations, with appropriate constraints among the coefficients in the two parts and with one part multiplied by a relative scale factor to be estimated.

9.2.4 Interpreting Coefficient Estimates

It is useful for interpreting empirical results to note that a change in $\beta' z_{in}$ in (9.10) by an amount of ±1 increases or decreases the relative odds of alternative *i*, compared to each other alternative, by a factor exp(1)=2.72. Thus a quick gauge of the behavioral significance of any particular variable can be obtained by considering the size of typical variations in that variable, multiplying it by the relevant coefficient, and comparing with 1.0.

The parameter vector may contain *alternative-specific constants* for one or more alternatives *i*. That is, the systematic utility may be of the form

$$V_{in} = \alpha_i + \beta' z_{in} . \tag{9.12}$$

Since only utility differences matter, at least one of the alternative-specific constants must be normalized (usually to zero); that alternative then serves as a "base alternative" for comparisons. Of course, using alternative-specific constants makes it impossible to forecast the result of adding a new alternative unless there is some basis for a guess as to what its alternative-specific constant would be.

Equation (9.12) is really a special case of (9.10) in which one or more of the variables Z are *alternative-specific dummy variables*, D^k, defined by $D^k_{jn}=1$ if $j=k$ and 0 otherwise (for each $j=1,\ldots,J$). (Such a variable does not depend on n.) In this notation, parameter α_i in (9.12) is viewed as the coefficient of variable D^i included among the z variables in (9.10). Such dummy variables can also be interacted with (i.e., multiplied by) any other variable, making it possible for the latter variable to affect utility in a different way for each alternative. All such variables and interactions may be included in z, and their coefficients in β, thus allowing (9.10) still to represent the linear-in-parameters specification.

The most economically meaningful quantities obtained from estimating a discrete-choice model are often ratios of coefficients. By interacting the variables of interest with socioeconomic characteristics or alternative-specific constants, these ratios can be specified quite flexibly so as to vary in a manner thought to be *a priori* plausible. A particularly important example in transportation is the ratio of coefficients of time and money, often called the *value of travel-time savings*, or *value of time* for short. It represents the monetary value that the traveler places on an incremental time saving. Similarly, a per-unit value can be placed any product attribute that consumers care about: for example, interior capacity of a vehicle, throughput rate of a communications device, or resolution of a visual display.

The value of time in the model (9.9) is

$$(v_T)_{in} \equiv -\left(\frac{dc_{in}}{dT_{in}}\right)_{V_{in}} \equiv \frac{\partial V_{in} / \partial T_{in}}{\partial V_{in} / \partial c_{in}} = \left(\frac{\beta_2 + 2\beta_3 T_{in}}{\beta_1}\right) \cdot w_n , \tag{9.13}$$

which varies across individuals since it depends on w_n and T_{in}.

As a more complex example, suppose we extend equation (9.9) by adding alternative-specific dummies, both separately (with coefficients α_i) and interacted with travel time (with coefficients γ_i):

$$V_{in} = \alpha_i + \beta_1 \cdot (c_{in} / w_n) + \beta_2 T_{in} + \beta_3 T_{in}^2 + \gamma_i T_{in} \tag{9.14}$$

where one of the α_i and one of the γ_i are normalized to zero. This yields the following value of time applicable when individual n chooses alternative i:

$$(v_T)_{in} = \left(\frac{\beta_2 + 2\beta_3 T_{in} + \gamma_i}{\beta_1}\right) \cdot w_n . \tag{9.15}$$

Now the value of time varies across modes even with identical travel times, due to the presence of γ_i. In the same way, the value consumers place on a specified increase in resolution of a visual display could depend on the income (or any other characteristic) of the individual and on the particular model or display type selected.

Confidence bounds for a ratio of coefficients can be estimated by standard approximations for transformations of normal variates.[2] Or they can be estimated using a Monte Carlo procedure: take repeated random draws from the distribution of β (which is estimated along with β itself), and then examine the resulting values of the ratio in question. The empirical distribution of these values is an estimate of the actual distribution of the ratio, and one can

[2] Letting v_T=β_2/β_1, the standard deviation σ_v of v_T obeys the intuitive formula: $(\sigma_v/v_T)^2 \cong (\sigma_1/\beta_1)^2 + (\sigma_1/\beta_1)^2 - 2\sigma_{12}/(\beta_1\beta_2)$, where σ_1 and σ_2 are the standard deviations of β_1 and β_2 and σ_{12} is their covariance.

describe it in any number of ways including its standard deviation. As another example, the 5th and 95th percentile values of those values define a 90 percent confidence interval for β. See Train (2003, ch. 9) for how to take such random draws.

9.2.5 Randomness, Scale of Utility, Measures of Benefit, and Forecasting

The variance of the random utility term in equation (9.4) reflects randomness in behavior of individuals or, more likely, heterogeneity among observationally identical individuals. Hence it plays a key role in determining how sensitive travel behavior is to observable quantities such as price, service quality, and demographic traits. Little randomness implies a nearly deterministic model, one in which behavior suddenly changes at some crucial switching point (for example, when transit service becomes as fast as a car). Conversely, if there is a lot of randomness, behavior changes only gradually as the values of independent variables are varied.

When the variance of the random component is normalized, however, the degree of randomness becomes represented by the inverse of the scale of the systematic utility function. For example, in the logit model (9.8), suppose systematic utility is linear in parameter vector β as in (9.10). If all the elements of β are small in magnitude, the corresponding variables have little effect on probabilities so choices are dominated by randomness. If the elements of β are large, most of the variation in choice behavior is explained by variation in the observable variables. Randomness in individual behavior can also be viewed as producing variety, in aggregate behavior.

It is sometimes useful to have a measure of the overall desirability of the choice set being offered to a consumer. Such a measure must account both for the utility of the individual choices being offered and for the variety of choices offered. The value of variety is directly related to randomness because both arise from unobserved idiosyncrasies in preferences. If choice were deterministic, the consumer would care only about the traits of the best alternative; improving or offering inferior alternatives would have no value. But with random utilities, there is some chance that an alternative with a low value of V_{in} will nevertheless be chosen; so it is desirable for such an alternative to be offered and to be made as attractive as possible. A natural measure of the desirability of choice set J is the expected maximum utility of that set, which for the logit model has the convenient form:

$$E \max_{\mathrm{j}} (V_j + \varepsilon_j) = \mu^{-1} \log \sum_{j=1}^{J} \exp(\mu V_j) + \gamma \tag{9.16}$$

where γ=0.5772 is Euler's constant (it accounts for the nonzero mean of the error terms ε_j in the standard normalization). (Here we have retained the parameter μ, rather than normalizing it, to make clear how randomness affects expected utility.) When the amount of randomness is small (large μ), the summation on the right-hand side is dominated by its largest term (let's denote its index by $j*$); expected utility is then approximately $\rho\cdot\log[\exp(V_{j*}/\rho)] = V_{j*}$, the utility of the dominating alternative. When randomness dominates (small μ), all terms contribute more or less equally (let's denote their average utility value by V); then expected utility is approximately $\mu^{-1}\cdot\log[J\cdot\exp(\mu V)] = V+\mu^{-1}\cdot\log(J)$, which is the average utility plus a term reflecting the desirability of having many choices.

Expected utility is, naturally enough, directly related to measures of consumer welfare. Small and Rosen (1981) show that, in the absence of income effects, changes in aggregate consumer surplus (the area to the left of the demand curve and above the current price) are appropriate measures of welfare even when the demand curve is generated by a set of individuals making discrete choices. For a set of individuals n characterized by systematic utilities V_{jn}, changes in consumer surplus are proportional to changes in this expected maximum utility. The proportionality constant is the inverse of λ_n, the marginal utility of income. Thus a useful welfare measure for such a set of individuals, with normalization μ=1, is:

$$W = \frac{1}{\lambda_n} \log \sum_{j=1}^{J} \exp(V_{jn}) \,. \tag{9.17}$$

(The constant γ drops out of welfare comparisons so is omitted here.) Because portions of the utility V_i that are common to all alternatives cannot be estimated from the choice model, λ_n cannot be estimated directly. However, typically it can be determined from Roy's Identity:

$$\lambda_n = -\frac{1}{x_{in}} \cdot \frac{\partial V_{in}}{\partial c_{in}} \tag{9.18}$$

where x_{in} is consumption of good i conditional on choosing it among the discrete alternatives. In the case of commuting-mode choice, for example, x_{in} is just the individual's number of work trips per year (assuming income and hence welfare are measured in annual units).

Once we have estimated a disaggregate travel-demand model, we face the question of how to predict aggregate quantities such as total transit ridership or total travel flows between zones. Ben-Akiva and Lerman (1985, chap. 6) discuss several methods. The most straightforward and common is *sample enumeration*. A sample of consumers is drawn, each assumed to represent a subpopulation with identical observable characteristics. (The estimation sample itself may satisfy this criterion and hence be usable as an enumeration sample.) Each individual's choice probabilities, computed using the estimated parameters, predict the shares of that subpopulation choosing the various alternatives. These predictions can then simply be added, weighting each sample member according to the corresponding subpopulation size. Standard deviations of forecast values can be estimated by Monte Carlo simulation methods.

9.2.6 Ordered and Rank-Ordered Models

Sometimes there is a natural ordering to the alternatives that can be exploited to guide specification. For example, suppose one wants to explain a household's choice among owning no vehicle, one vehicle, or two or more vehicles. It is perhaps plausible that there is a single index of propensity to own many vehicles, and that this index is determined in part by observable variables like household size and employment status.

In such a case, an *ordered response* model might be assumed. In this model, the choice of individual n is determined by the size of a "latent variable" $y_n^* = \beta' z_n + \varepsilon_n$, with choice j occurring if this latent variable falls in a particular interval $[\mu_{j-1},\mu_j]$ of the real line, where μ_0=−∞ and μ_J=∞. The interval boundaries $\mu_1,...,\mu_{J-1}$ are estimated along with β, except that one of them can be normalized arbitrarily if $\beta' z_n$ contains a constant term. The probability of choice j is then

$$P_{jn} = \Pr[\mu_{j-1} < \beta' z_n + \varepsilon_n < \mu_j] = F(\mu_j - \beta' z_n) - F(\mu_{j-1} - \beta' z_n) \qquad (9.19)$$

where $F(\cdot)$ is the cumulative distribution function assumed for ε_n. In the *ordered probit* model $F(\cdot)$ is standard normal, while in the *ordered logit* model it is logistic, i.e. $F(x) = [1+\exp(-x)]^{-1}$. Note that all the variables in this model are characteristics of individuals, not of the alternatives, and thus if the latter information is available this model cannot easily take advantage of it.

In some cases the alternatives are integers indicating the number of times some random event occurs. An example would be the number of trips per month by a given household to a particular destination. For such cases, a set of models based on Poisson and negative binomial regressions is available (Washington, Karlaftis, and Mannering, 2003). In other cases, information is available not only on the most preferred alternative, but on the individual's ranking of other alternatives. Efficient use can be made of such data through the *rank-ordered logit* model, also called "expanded logit" (Hausman and Ruud, 1987).

9.3 Examples of Disaggregate Models

Discrete-choice models have been estimated for nearly every conceivable travel situation. In this section we present two examples.

9.3.1 Mode Choice

A series of models explaining choices of automobile ownership and commuting mode in the San Francisco Bay area were developed as part of planning for the Bay Area Rapid Transit System, which opened in 1975. One of the simplest explains only the choice among four modes: (1) auto alone, (2) bus with walk access, (3) bus with auto access, and (4) carpool (two or more occupants). The model's parameters are estimated from a sample of 771 commuters to San Francisco or Oakland who were surveyed prior to opening of the Bay Area Rapid Transit system.[3]

Mode choice is explained by three independent variables and three alternative-specific constants. The three variables are: c_{in}/w_n, the round-trip variable cost (in US \$) of mode i for traveler n, divided by the traveler's post-tax wage rate (in \$ per minute); T_{in}, the in-vehicle travel time (in minutes); and T^o_{in}, the out-of-vehicle travel time including walking, waiting, and transferring. Cost c_{in} includes parking, tolls, gasoline, and maintenance. The estimated utility function is:

$$\begin{array}{lllllllr} V = & -0.0412\cdot c/w & -0.0201\cdot T & -0.0531\cdot T^o & -0.89\cdot D^1 & -1.78\cdot D^3 & -2.15\cdot D^4 & (9.20) \\ & (0.0054) & (0.0072) & (0.0070) & (0.26) & (0.24) & (0.25) & \end{array}$$

where the subscripts denoting mode and individual have been omitted, and standard errors of coefficient estimates are given in parentheses. Variables D^j are alternative-specific dummies.

[3] This is the "naive model" reported by McFadden et al. (1977, pp. 121-123).

This utility function is a simplification of (9.14) (with $\beta^3=\gamma^j=0$), except that travel time is broken into two components, T and T^o. Adapting (9.15), we see that the "value of time" for each of these two components is proportional to the post-tax wage rate: specifically, the estimated values of in-vehicle and out-of-vehicle time are 49 percent and 129 percent of the after-tax wage. The negative alternative-specific constants indicate that the hypothetical traveler facing equal times and operating costs by all four modes will prefer bus with walk access (mode 2, the base mode); probably this is because each of the other three modes requires owning an automobile, which entails fixed costs not included in variable *c*. The strongly negative constants for bus with auto access (mode 3) and carpool (mode 4) probably reflect unmeasured inconvenience associated with getting from car to bus stop and with arranging carpools.

The model's fit could undoubtedly be greatly improved by including automobile ownership, perhaps interacted with $(D^1+D^3+D^4)$ to indicate a common effect on modes that use an automobile. However, there is good reason to exclude it because it is endogenous—people choosing one of those modes for other reasons are likely to buy an extra car as a result. This in fact is demonstrated by the more complete model of Train (1980), which considers both choices simultaneously. The way to interpret (9.20), then, is as a "reduced-form" model that incorporates implicitly the automobile ownership decision. It is thus applicable to a time frame long enough for automobile ownership to adjust to changes in the variables included in the model.

9.3.2 Choice of Free or Express Lanes

Lam and Small (2001) analyze data from commuters with an option of paying to travel in a set of express lanes on a very congested freeway. The data set contains cross-sectional variation in the cost of choosing the express lanes because the toll depends on time of day and on car occupancy, both of which differ across respondents. Travel time also varies by time of day, fortunately in a manner not too highly correlated with the toll. The authors construct a measure of the unreliability of travel time by obtaining data on travel times across many different days, all at the same time of day. After some experimentation, they choose the median travel time (across days) as the best measure of travel time, and the difference between 90th and 50th percentile travel times (also across days) as the best measure of unreliability. This latter choice is based on the idea that people are more averse to unexpected delays than to unexpected early arrivals.

The model explains a pair of related choices: (1) whether to acquire a transponder (required to ever use the express lanes), and (2) which lanes to take on the day in question. A natural way to view these choices is as a hierarchical set, in which the transponder choice is governed partly by the size of the perceived benefits of being able to use it to travel in the express lanes. As we will see in the next section, a model known as "nested logit" has been developed precisely for this type of situation, and indeed Lam and Small estimate such a model. As it happens, though, they obtain virtually identical results with a simpler "joint logit" model in which there are three alternatives: (1) no transponder; (2) have a transponder but travel in the free lanes on the day in question; and (3) have a transponder and travel in the express lanes on the day in question. The results of this model are:[4]

$$\begin{array}{lllll} V= & -0.862\cdot D^{\text{tag}} & +0.0239\cdot Inc\cdot D^{\text{tag}} & -0.766\cdot ForLang\cdot D^{\text{tag}} & -0.789\cdot D^{3} \\ & (0.411) & (0.0058) & (0.412) & (0.853) \end{array}$$

$$\begin{array}{llllll} -0.357c & -0.109\cdot T & -0.159\cdot R & +0.074\cdot Male\cdot R & + \text{(other terms).} & (9.21) \\ (0.138) & 0.056) & (0.048) & (0.046) & & \end{array}$$

Here $D^{\text{tag}}\equiv D^2+D^3$ is a composite alternative-specific dummy variable for those choices involving a transponder, or "toll tag"; its negative coefficient presumably reflects the hassle and cost of obtaining one. Getting a transponder is apparently more attractive to people with high annual incomes (*Inc*, in $1000s per year) and less attractive to those speaking a foreign language (dummy variable *ForLang*). The statistical insignificance of the coefficient of D^3, an alternative-specific dummy for using the express lanes, suggests that the most important explanatory factors are included explicitly in the model.

The coefficients on per-person cost *c*, median travel time *T*, and unreliability *R* can be used to compute dollar values of time and reliability. Here we focus on two aspects of the resulting valuations. First, reliability is highly valued, achieving coefficients of similar magnitudes as travel time. Second, men seem to care less about reliability than women; their value is only 53 percent as high as women's according to the estimates of the coefficient of unreliability (-0.159 for women, -0.159+0.074 = -0.085 for men). (A qualification to this is that the difference, i.e. the coefficient of $Male\cdot R$, is not quite statistically significant at a 10-percent significance level.) Several studies of this particular toll facility have found women noticeably

[4] This is a partial listing of the coefficients in Lam and Small (2001), Table 11, Model 4b, with coefficients of *T* and *R* divided by 1.37 to adjust travel-time measurements to the time of the survey, as described on their p. 234 and Table 11, note *a*. Standard errors are in parentheses.

more likely to use the express lanes than men, and this formulation provides tentative evidence that the reason is a greater aversion to the unreliability of the free lanes.

9.4 Advanced Discrete-Choice Modeling

9.4.1 Generalized Extreme Value Models

Often it is implausible that the additive random utility components ε_j be independent of each other, especially if important variables are omitted from the model's specification. This will make either logit or iid probit predict poorly.

A simple example is mode choice among automobile, bus transit, and rail transit. The two public-transit modes are likely to have many unmeasured attributes in common. Suppose a traveler initially has available only auto (j=1) and bus (j=2), with equal systematic utilities V_j so that the choice probabilities are each one-half. Now suppose we want to predict the effects of adding a type of rail service (j=3) whose measurable characteristics are identical to those for bus service. The iid models would predict that all three modes would then have choice probabilities of one-third, whereas in reality the probability of choosing auto would most likely remain near one-half while the two transit modes divide the rest of the probability equally between them. The argument is even stronger if we imagine instead that the newly added mode is simply a bus of a different color: this is the famous "red bus, blue bus" example.

The probit model generalizes naturally, as already noted, by allowing the distribution function in equation (9.5) to be multivariate normal with an arbitrary variance-covariance matrix. It must be remembered that not all the elements of this matrix can be distinguished (*identified*, in econometric terminology) because, as already noted, it is only the (J-1) utility differences that affect behavior.[5]

The logit model generalizes in a comparable manner, as shown by McFadden (1978, 1981). The distribution function is postulated to be *Generalized Extreme Value* (GEV), given by

$$F(\varepsilon_1,...,\varepsilon_J) = \exp\left[-G(e^{-\varepsilon_1},...,e^{-\varepsilon_J})\right]$$

[5] The variance-covariance matrix of these utility differences has $(J\text{-}1)^2$ elements and is symmetric. Hence there are only $J(J\text{-}1)/2$ identifiable elements of the original variance-covariance matrix, less one for utility-scale normalization (Bunch, 1991).

where G is a function satisfying certain technical conditions. Logit is the special case $G(y_1,...,y_J) = y_1+...+y_J$.

The best known GEV model, other than logit itself, is *nested logit*, also called *structured logit* or *tree logit*. In this model, certain groups of alternatives are postulated to have correlated random terms. This is accomplished by grouping the corresponding alternatives in G in a manner we can illustrate using the auto-bus-rail example, with auto the first alternative:

$$G(y_1, y_2, y_3) = y_1 + \left(y_2^{1/\rho} + y_3^{1/\rho}\right)^{\rho} . \tag{9.22}$$

Here ρ is a parameter between 0 and 1 that indicates the degree of dissimilarity between bus and rail; more precisely, $\sqrt{1-\rho^2}$ is the correlation between ε_1 and ε_2 (Daganzo and Kusnic, 1993). The choice probability for this example may be written:

$$P_i = P(B_{r(i)}) \cdot P(i \mid B_r) \tag{9.23}$$

$$P(B_r) = \frac{\exp(\rho \cdot I_r)}{\sum_{s=1}^{2} \exp(\rho \cdot I_s)} \tag{9.24}$$

$$P(i \mid B_r) = \frac{\exp(V_i / \rho)}{\sum_{j \in B_r} \exp(V_j / \rho)} \tag{9.25}$$

where $B_1=\{1\}$ and $B_2=\{2,3\}$ are a partition of the choice set into groups; $r(i)$ indexes the group containing alternative i; and I_r denotes the *inclusive value* of set B_r, defined as the logarithm of the denominator of (9.25):

$$I_r = \log \sum_{j \in B_r} \exp(V_j / \rho) . \tag{9.26}$$

When ρ=1 in this model, ε_2 and ε_3 are independent and we have the logit model. As $\rho \downarrow 0$, ε_2 and ε_3 become perfectly correlated and we have an extreme form of the "red bus, blue bus" example, in which auto is pitted against the better (as measured by V_i) of the two transit alternatives; in this case $\rho I_1 \rightarrow V_1$ and $\rho I_2 \rightarrow \max\{V_2, V_3\}$.

The model just described can be generalized to any partition $\{B_r, r=1,\ldots,R\}$ of alternatives, and each group B_r can have its own parameter ρ_r in equations (9.22)-(9.26), leading to the form:

$$G(y_1,\ldots,y_J)=\sum_r\left(\sum_{j\in B_r}y_j^{1/\rho_r}\right)^{\rho_r}. \tag{9.27}$$

This is the general two-level nested logit model. It has choice probabilities (9.23)-(9.26) except that the index s in the denominator of (9.24) now runs from 1 to R. The welfare measure for the two-level nested logit model is:

$$W=\frac{1}{\lambda}\log\sum_r\exp(\rho_r\cdot I_r) \tag{9.28}$$

where again λ is the marginal utility of income.

In nested logit, $\{B_r\}$ is an exhaustive partition of the choice set into mutually exclusive subsets. Therefore equation (9.25) is a true conditional probability, and the model can be estimated sequentially: first estimate the parameters (β/ρ) from (9.25), use them to form the inclusive values (9.26), then estimate ρ from (9.24). Each estimation step uses an ordinary logit log-likelihood function, so it can be carried out with a logit algorithm. However, this sequential method is not statistically efficient and is rarely used today. Several studies show that maximum-likelihood estimation gives more accurate results (Brownstone and Small, 1989).

A different direction for generalizing the logit model is to maintain independence between error terms while allowing each error term to have a unique variance. This is the heteroscedastic extreme value model of Bhat (1995); it is a random-utility model but not in the GEV class, and its probabilities cannot be written in closed form so require numerical integration. Other extensions of the logit model are described by Koppelman and Sethi (2000).

9.4.2 Combined Discrete and Continuous Choice

In many situations, the choice among discrete alternatives is made simultaneously with some related continuous quantity. For example, a household's choice of type of automobile to own is closely intertwined with its choice of how much to drive. Estimating equations to explain usage, conditional on ownership, creates a *sample selection bias* (Heckman, 1979): for example, people who drive a lot are likely to select themselves into the category of owners of nice cars, so we could inadvertently overstate the independent effect of nice cars on driving. A variety of methods are available to remove this bias, as described in Train (1986, chap. 5) and Washington *et al.* (2003, ch. 12).

More elaborate systems of equations can be handled with the tools of *structural equations modeling*. These methods are quite flexible and allow one to try out different patterns of mutual causality, testing for the presence of particular causal links. They are often used when large data sets are available describing mutually related choices. Golob (2003) provides a review.

9.4.3 Disaggregate Panel Data

Just as with aggregate data, data from individual respondents can be collected repeatedly over time. A good example is the Dutch Mobility Panel, in which travel-diary information was obtained from the same individuals (with some attrition and replacement) at ten different times over the years 1984-1989. The resulting data have been widely used to analyze time lags and other dynamic aspects of travel behavior (Van Wissen and Meurs, 1989).

The methods described earlier for aggregate panel data are applicable to disaggregate data as well. In addition, attrition becomes a statistical issue: over time, some respondents will be lost from the sample and the reasons need not be independent of the behavior being investigated. The solution is to create an explicit model of what causes an individual to leave the sample, and to estimate it simultaneously with the choice process being considered. Pendyala and Kitamura (1997) and Brownstone and Chu (1997) analyze the issues involved.

9.4.4 Random Parameters and Mixed Logit

In the random utility model of (9.4)-(9.5), randomness in individual behavior is limited to an additive error term in the utility function. Other parameters, and functions of them, are deterministic: that is, the only variation in them is due to observed variables. Thus for example, the value of time defined by (9.13) varies with observed travel time and wage rate but otherwise is the same for everyone.

Experience has shown, however, that parameters of critical interest to transportation policy vary among individuals for reasons that we do not observe. Such reasons could be missing socioeconomic characteristics, personality, special features of the travel environment, and data errors. These, of course, are the same reasons for the inclusion of the additive error term in utility function (9.4). So the question is, why not also include randomness in the other parameters?

The only reason is tractability, and that has largely been overcome by advances in computing power. Consider first how one could allow a single parameter in the logit model to vary randomly across individuals. Suppose we specify a distribution, such as normal with unknown mean and variance, for the parameter in question. The overall probability is then determined by embedding the integral in (9.5) within another integral over the density function of that distribution. This simple idea has been generalized to allow for general forms of randomness in many parameters, including alternative-specific constants, leading to a many-dimensional integral. Nevertheless the model is tractable because the outer integration (over the distribution defining random parameters) can be performed using simulation methods based on random draws, while the inner integration (that over the remaining additive errors ε_{jn}) is unnecessary because, conditional on the values of random parameters, it yields the logit formula (9.8). The model is called *mixed logit* because the combined error term has a distribution that is a mixture of the extreme value distribution with the distribution of the random parameters.

Writing this out explicitly, the choice probability conditional on random parameters is

$$P_{in|\beta} = \frac{\exp(\beta' z_{in})}{\sum_j \exp(\beta' z_{jn})}. \tag{9.29}$$

Let $f(\beta|\Theta)$ denote the density function defining the distribution of random parameters, which depends on some unknown "meta-parameters" Θ (such as means and variances of β). The unconditional choice probability is then simply the multi-dimensional integral:

$$P_{in} = \int P_{in|\beta} \cdot f(\beta \mid \Theta) d\beta . \tag{9.30}$$

Integration by simulation consists of taking R random draws β^r, r=1,…,R, from distribution $f(\beta|\Theta)$, calculating $P_{in|\beta}$ each time, and averaging over the resulting values:

$$P_{in}^{sim} = (1/R)\sum_{r=1}^{R} P_{in|\beta}^{r} .$$

Doing so requires, of course, assuming some trial value of Θ, just as calculating the usual logit probability requires assuming some trial value of β. Under reasonable conditions, maximizing the likelihood function defined by this simulated probability yields statistically consistent estimates of the meta-parameters Θ. Details are provided by Train (2003).

Brownstone and Train (1999) demonstrate how one can shape the model to capture anticipated patterns by specifying which parameters are random and what form their distribution takes – in particular, whether some of them are correlated with each other.[6] In their application, consumers state their willingness to purchase various makes and models of cars, each specified to be powered by one of four fuel types: gasoline (G), natural gas (N), methanol (M), or electricity (E). Respondents were asked to choose among hypothetical vehicles with specified characteristics. A partial listing of estimation results is as follows:

$$\begin{aligned} V = \; & -0.264\cdot[p/\ln(inc)] + 0.517\cdot range + (1.43+7.45\phi_1)\cdot size + (1.70+5.99\phi_2)\cdot luggage \\ & + 2.46\phi_3\cdot nonE + 1.07\phi_4\cdot nonN + (other\ terms) \end{aligned}$$

where *p* (vehicle price) and *inc* (income) are in thousands of dollars; the *range* between refueling (or recharging) is in hundreds of miles; *luggage* is luggage space relative to a comparably sized gasoline vehicle; *nonE* is a dummy variable for cars running on a fuel that must be purchased outside the home (in contrast to electric cars); *nonN* is a dummy for cars running on a fuel stored at atmospheric pressure (in contrast to natural gas); and ϕ_1-ϕ_4 are independent random variables with the standard normal distribution. All parameters shown above are estimated with enough precision to easily pass tests of statistical significance.

This model provides for observed heterogeneity in the effect of *price* on utility, since it varies with *income*. It provides for random coefficients on *size* and *luggage*, and for random constants as defined by *nonE* and *nonN*. This can be understood by examining the results term by term.

The terms in parentheses involving ϕ_1 and ϕ_2 represent the random coefficients. The coefficient of *size* is random with mean 1.43 and standard deviation 7.45. Similarly, the coefficient of *luggage* has mean 1.70 and standard deviation 5.99. These estimates indicate a wide variation in people's evaluation of these characteristics. For example, it implies that many people (namely, those for whom ϕ_2<-1.70/5.99) actually prefer less luggage space; presumably they do so because a smaller luggage compartment allows more interior room for the same size of vehicle. Similarly,

[6] The following simplified explanation is adapted from Small and Winston (1999).

preference for vehicle size ranges from negative (perhaps due to easier parking for small cars) to substantially positive.

The terms involving ϕ_3 and ϕ_4 represent random alternative-specific constants with a particular correlation pattern, predicated on the assumption that groups of alternatives share common features for which people have idiosyncratic preferences – very similar to the rationale for nested logit. Each of the dummy variables *nonE* and *nonN* is simply a sum of alternative-specific constants for those car models falling into a particular group. The two groups overlap: any gasoline-powered or methanol-powered car falls into both. If the coefficients of ϕ_3 and ϕ_4 had turned out to be negligible, then these terms would play no role and we would have the usual logit probability conditional on the values of ϕ_1 and ϕ_2. But the coefficients are not negligible, so each produces a correlation among utilities for those alternative in the corresponding group. For example, all cars that are not electric share a random utility component $2.46\phi_3$, which has standard deviation 2.46 (since ϕ_3 has standard deviation one by definition). Thus the combined additive random term in utility (including the random constants), $\varepsilon_{in}+2.46\phi_{3n}\cdot nonE_i+1.07\phi_{4n}\cdot nonN_i$, exhibits correlation across those alternatives *i* representing cars that are not electric. A similar argument applies to ϕ_4, which produces correlation across those alternatives representing cars that are not natural gas. Those alternatives falling into both *nonE* and *nonN* are even more highly correlated with each other. Note that because the distributions of ϕ_3 and ϕ_4 are centered at zero, this combined random term does not imply any overall average preference for or against various types of vehicles; such absolute preferences are in fact included in *other terms*.

The lesson from this example is that mixed logit can be used not only to specify unobserved randomness in the coefficients of certain variables, but also to mimic the kinds of correlation patterns among the random constants for which the GEV model was developed. Indeed, McFadden and Train (2000) show that it can closely approximate virtually any choice model based on random utility.

9.5 Value of Time and Reliability

Among the most important quantities inferred from travel demand studies are the monetary values that people place on saving various forms of travel time or improving the predictability of travel time. The first, loosely known as the *value of time* (VOT), is a key parameter in cost-benefit analyses that measure the benefits brought about by transportation policies or projects. The second, the *value of reliability* (VOR), also appears important, but accurate measurement is a science in its infancy. The benefits or losses due to changes in time and reliability are normally captured as part of consumer surplus, for example that given by (9.17), so long as they are part of the demand model. However, it is often enlightening to separate them explicitly.

9.5.1 Value of Time

The most natural definition of value of time is in terms of compensating variation. The value of saving a given amount and type of travel time by a particular person is the amount that person could pay, after receiving the saving, and be just as well off as before. This amount, divided by the time saving, is that person's average value of time saved for that particular change. Aggregating over a class of people yields the *average value of time* for those people in that situation. The limit of this average value, as the time saving shrinks to zero, is called the *marginal value of time*, or just "value of time;" by definition, it is independent of the amount of time saving. It was defined empirically in equation (9.13).

Value of time may depend on many aspects of the trip-maker and of the trip itself. To name just a few, it depends on trip purpose (e.g. work or recreation), demographic and socio-economic characteristics, time of day, physical or psychological amenities available during travel, and the total duration of the trip. There are two main approaches to specifying a travel-demand model so as to measure such variations. One is known as *market segmentation*: the sample is divided according to criteria such as income and type of household, and a separate model is estimated for each segment. This has the advantage of imposing no potentially erroneous constraints, but the disadvantage of requiring many parameters to be estimated, with no guarantee that these estimates will follow a reasonable pattern. The second approach uses theoretical reasoning to postulate a functional form for utility that determines how VOT varies. This approach often builds on a framework due to Becker (1965), in which utility is maximized subject to a time constraint. Becker's theory has been elaborated in many directions, most of which predict some relationship between value of time and the wage rate. For example, the theory of Oort (1969) predicts that the value of time will exceed the wage rate if time spent at work is enjoyed relative to that spent traveling, and fall short of it if the opposite is true. Thus the value of time, even for non-work trips, depends on conditions of the job.

These theories can provide guidance about how to specify the systematic utilities V_k in a discrete choice model. Suppose, for example, one believes that work is disliked (relative to travel), that its relative marginal disutility is a fixed fraction of the wage rate. Then the value of time is a fraction of the wage rate, as for example with specification (9.9) with β_3=0. Alternatively, one might think that work enjoyment varies nonlinearly with the observed wage rate: perhaps negatively due to wage differentials that compensate for working conditions, or perhaps positively due to employers' responses to an income-elastic demand for job amenities. Then the value of time is a nonlinear function of the wage rate, which could suggest using (9.9) with a non-zero term β_3.

9.5.2 Value of Reliability

It is well known that uncertainty in travel time, which may result from congestion or poor adherence to transit schedules, is a major perceived cost of travel. A parallel with other types of products is fairly obvious: uncertainty in how well a product will perform the desired function will reduce its value to the user .

How can reliability be captured in a theoretical model of travel? Adapting Noland and Small (1995), we can begin with a model of trip-scheduling choice, in which trip cost depends on the degree of adherence to a desired time of arrival at work. Define *schedule delay*, S_D, as the difference (in minutes, rounded to nearest five minutes) between the arrival time represented by a given alternative and the official work start time t^*. Define "Schedule Delay Late" as SDL=Max$\{S_D,0\}$ and "Schedule Delay Early" as SDE=Max$\{-S_D,0\}$. Define a "late dummy," DL, equal to one for the on-time and all later alternatives and equal to 0 for the early alternatives. Define T as the travel time (in minutes) encountered at each alternative. Suppose, then, that trip cost is a linear function of these variables:

$$C(t_d,T_r) = \alpha\cdot T + \beta\cdot SDE + \gamma\cdot SDL + \theta\cdot DL \qquad (9.31)$$

where $\alpha \equiv v_T/60$ is the per-minute value of travel time, β and γ are per-minute costs of early and late arrival, and θ is a fixed cost of arriving late. The functional notation $C(t_d,T_r)$ is to remind us that each of the components of trip cost depends on the departure time, t_d, and a random (unpredictable) component of travel time, $T_r \geq 0$. Our objective is to measure the increase in expected cost C due to the dispersion in T_r, given that t_d is subject to choice by the traveler. Letting C^* denote this expected cost after the user chooses t_d optimally, we have

$$C^* = \underset{t_d}{Min}\, E\left[C(t_d, t_r)\right] = \underset{t_d}{Min}\left[\alpha \cdot E(T) + \beta \cdot E(SDE) + \gamma \cdot E(SDL) + \theta \cdot P_L\right] \qquad (9.32)$$

where E denotes an expected value taken over the distribution of T_r, and where $P_L \equiv E(DL)$ is the probability of being late. This equation can form the basis for specifying the reliability term in a model like (9.21).

To focus just on reliability, let's ignore congestion for now by assuming that E(T) is independent of departure time. Remarkably, the optimal value of t_d then does not depend on the distribution of T_r, provided that its probability density is everywhere finite. To find this optimal departure time, let $f(T_r)$ be the distribution function, and let T_f be value of travel time when T_r=0. The next to last term in the square brackets of (9.32) can then be written as

$$\begin{aligned}\gamma \cdot E(SDL) &= \gamma \cdot E(t_d + T_r - \tilde{t} \mid T_r > \tilde{t} - t_d) \\ &= \gamma \cdot \int_{\tilde{t}-t_d}^{\infty} \left(t_d + T_r - \tilde{t}\right) \cdot f(T_r) dT_r\end{aligned}$$

where $\tilde{t} \equiv t^* - T_f$ is the time the traveler would depart if T_r were equal to zero with certainty. Differentiating yields:

$$\frac{d}{dt_d} \gamma \cdot E(SDL) = 0 + \gamma \cdot \int_{\tilde{t}-t_d}^{\infty} \left[\frac{d}{dt_d}\left(t_d + T_r - \tilde{t}\right) \cdot f(T_r)\right] dT_r = \gamma P_L^*$$

where P_L^* is the optimal value of the probability of being late.[7] Similarly, differentiating the term involving β in (9.32) yields $-\beta \cdot \left(1 - P_L^*\right)$. Finally, differentiating the last term yields $-\theta f^0$ where $f^0 \equiv f(\tilde{t} - t_d^*)$ is the probability density at the point where the traveler is neither early nor late. Combining all three terms and setting them equal to zero gives the first-order condition for optimal departure time:

$$P_L^* = \frac{\beta + \theta f^0}{\beta + \gamma} \,. \qquad (9.33)$$

[7] The term "0" in this equation arises from differentiating the lower limit of integration: $-\left[d(\tilde{t} - t_d)/dt_d\right] \cdot \left[(t_d + T_r - \tilde{t}) \cdot f(T_r)\right]_{T_r = \tilde{t} - t_d}$ =1·0=0.

In general this does not yield a closed-form solution for t_d^* because f^0 depends on t_d^*. However, in the special case θ=0, it yields $P_L^* = \beta/(\beta+\gamma)$, a very intuitive rule for setting departure time that is noted by Bates *et al.* (2001, p. 202). The rule balances the aversions to early and late arrival.

The cost function itself has been derived in closed form for two cases: a uniform distribution and an exponential distribution for T_r. In the case of a uniform distribution with range *b*, (9.33) again simplifies to a closed form:

$$P_L^* = \frac{\beta + (\theta/b)}{\beta+\gamma}.$$

The value of C^* in this case is given by Noland and Small (1995) and Bates *et al.* (2001). In the special case θ=0, it is equal to the cost of expected travel time, $\alpha \cdot E(T)$, plus the following cost of unreliability:

$$v_R = \left(\frac{\beta\gamma}{\beta+\gamma}\right)\cdot\frac{b}{2}. \qquad (9.34)$$

The quantity in parentheses is a composite measure of the unit costs of scheduling mismatch, which plays a central role in the cost functions considered in the next chapter. Thus (9.34) indicates that reliability cost derives from the combination of costly scheduling mismatches and dispersion in travel time.

More generally, we can see from (9.32) that whatever the form of the distribution of uncertain travel time, expected trip cost will increase with dispersion in that distribution. Furthermore, if γ>β and/or if θ is large, both of which are confirmed by the empirical findings of Small (1982), expected cost will be especially sensitive to the possibility of values of T_r high enough to make the traveler late even though t_d is chosen optimally. Therefore the cost of unreliability depends especially on the upper tail of the distribution of uncertain travel times. This property was used in creating the reliability variable in the study by Lam and Small (2001) described earlier.

In a similar manner, the reliability of a products design may need to be measured primarily by one part of the distribution of random events associated with the product's functioning. If a boat rudder bends under certain wave conditions, this may reduce its efficiency, with some minor loss of value; whereas if it bends so far as to break, the loss is much greater.

9.5.3 Empirical Results

Research has generated an enormous literature on empirical estimates of value of time, and a much smaller one on value of reliability. Here we rely mainly on reviews of this literature by others.

Reviewing studies for the UK, Wardman (1998, Table 6) finds an average VOT of £3.58/hour in late 1994 prices, which is 52% of the corresponding wage rate.[8] Gunn (2001) find that Dutch values used by planners in the late 1980s track British results (by household income) quite well; however, he also finds a substantial unexplained downward shift in the profile for 1997, possibly resulting from better in-vehicle amenities. Transport Canada (1994) and US Department of Transportation (1997) recommend using a VOT for personal travel by automobile equal to 50 percent of the gross wage rate. A French review by the Commissariat Général du Plan (2001, p. 42) finds VOT to be 59 percent of the wage on average for urban trips. Finally, a Japanese review suggests using 2,333 yen/hour for weekday automobile travel in 1999, which was 84 percent of the wage rate.[9]

There is considerable evidence that value of time rises with income but less than proportionally. The easiest way to summarize this issue is in an elasticity of value of time with respect to income. Wardman (2004), using a formal meta-analysis, finds that elasticity to be 0.72 when income is measured as gross domestic product per capita. Wardman's (2001) meta-analysis focuses on how value of time depends on various trip attributes. There is a small positive relationship (elasticity 0.13) with trip distance, a 16 percent differential between commuting and leisure trips, and considerable differences across modes, with bus riders having a lower than average value and rail riders a higher than average value – possibly due to self-selection by speed. Most important, walking and waiting time are valued much higher than in-vehicle time – a universal finding conventionally summarized as 2 to 2-1/2 times as high, although Wardman finds them to be only 1.6 times as high.

One unsettled methodological issue is an apparent tendency for SP data to yield considerably smaller values of time than RP data. Brownstone and Small (2005) find that SP

[8] Mean gross hourly earnings for the UK were £6.79 and £7.07/hour in spring 1994 and 1995, respectively. Source: National Statistics Online (2004, Table 38).

[9] Japan Research Institute Study Group on Road Investment Evaluation (2000), Table 3-2-2, using car occupancy of 1.44 (p. 52). Average wage rate is calculated as cash earnings divided by hours worked, from Japan Ministry of Health, Labour and Welfare (1999).

results for VOT are one-third to one-half the corresponding RP results. One possible explanation for this difference is hinted at by the finding from other studies that people overestimate the actual time savings from the toll roads by roughly a factor of two; thus when answering SP survey questions, they may indicate a per-minute willingness to pay for *perceived* time savings that is lower than their willingness to pay for *actual* time savings. If one wants to use a VOT for purposes of policy analysis, one needs it to correspond to actual travel time since that is typically the variable considered in the analysis. Therefore if RP and SP values differ when both are accurately measured, it is the RP values that are relevant for most purposes.

From this evidence, it appears that the value of time for personal journeys is almost always between 20 and 90 percent of the gross wage rate, most often averaging close to 50 percent. Although it varies somewhat less than proportionally with income, it is close enough to proportional to make its expression as a fraction of the wage rate a good approximation and more useful than expression as an absolute amount. There is universal agreement that value of time is much higher for travel while on business, generally recommended to be set at 100 percent of total compensation including benefits. The value of walking and waiting time for transit trips is probably 1.6 to 2.0 times that of in-vehicle time, not counting some context-specific disutility of having to transfer from one vehicle to another.

There has been far less empirical research on value of reliability (VOR). Most of it has been based on SP data, for at least two reasons: if is difficult to measure unreliability in actual situations, and unreliability tends to be correlated with travel time itself. However, a few recent studies, including Lam and Small (2001), have had some success with RP data. Brownstone and Small (2005) review several such studies in which unreliability is defined as the difference between the 90th and 50th percentile of the travel-time distribution across days, or some similar measure; in those studies, VOR tends to be of about the same magnitude as VOT. One of those studies, using data from the high-occupancy toll (HOT) lane on State Route 91 in the Los Angeles region, finds that roughly two-thirds of the advantage of the HOT lane to the average traveler is due to its lower travel time and one-third is due to its higher reliability.[10] In prospective studies of a possible £4 cordon toll for Central London, May, Coombe and Gilliam (1996) estimate that reliability would account for 23 percent of the benefits to car users.

[10] An updated version of that study is Small, Winston, and Yan (2005).

9.6 Conclusions

The methods discussed here have spread far beyond transportation to applications in labor economics, industrial organization, and many other fields. The field of marketing has taken them up with special vigor, adapting and refining them to match the kinds of data often elicited in marketing surveys. Some of the refinements involve more sophisticated models, sometimes made feasible by large volumes of data. Others involve stated preference (SP) methodology, which is prevalent in marketing studies. Researchers have paid considerable attention to using information on the demand for product characteristics to forecast the reaction to new products.

In these and other ways, methods from travel demand analysis can bring information to bear on how consumers value the characteristics under consideration in design problems, and how the demand for products will depend on those design decision. There is ample room for specialists in design to both use and contribute to the tools described here.

References

Bates, J., Polak, J., Jones, P., and Cook, A., 2001, "The Valuation of Reliability for Personal Travel." *Transportation Research E: Logistics and Transportation Review*, **37**, pp. 191-229.

Becker, G. S., 1965, "A Theory of the Allocation of Time." *Economic Journal*, **75**, pp. 493-517.

Ben-Akiva, M., and Lerman, S. R., 1985, *Discrete Choice Analysis: Theory and Application to Travel Demand*, Cambridge, Mass.: MIT Press.

Bhat, C., 1995, "A Heteroscedastic Extreme Value Model of Intercity Travel Mode Choice." *Transportation Research Part B*, **29**, pp. 471-483.

Brownstone, D., and Chu, X., 1997, "Multiply-Imputed Sampling Weights for Consistent Inference with Panel Attrition," in Golob, T. F., R. Kitamura, and L. Long (eds.), *Panels for Transportation Planning: Methods and Applications,* 259-273.

Brownstone, D., and Train, K., 1999, "Forecasting new product penetration with flexible substitution patterns." *Journal of Econometrics*, **89**, pp. 109-129.

Brownstone, D., and Small, K. A., 1989, "Efficient Estimation of Nested Logit Models." *Journal of Business and Economic Statistics*, **7**, pp. 67-74.

Brownstone, D., and Small, K. A., 2005, "Valuing Time and Reliability: Assessing the Evidence from Road Pricing Demonstrations," *Transportation Research Part A*, **39**, pp. 279-293.

Bunch, D. S., 1991, "Estimability in the Multinomial Probit Model." *Transportation Research Part B,* **25**, pp. 1-12.

Commissariat Général du Plan, 2001, *Transports: Choix des Investissements et coût des nuisances* (*Transportation: Choice of Investments and the Cost of Nuisances*), Paris, June.

Daganzo, C. F. and Kusnic, M., 1993, "Two Properties of the Nested Logit Model," *Transportation Science*, 27, pp. 395-400.

Gómez-Ibáñez, J. A., 1996, "Big-City Transit Ridership, Deficits and Politics: Avoiding Reality in Boston." *Journal of the American Planning Association*, **62**, pp. 30-50.

Greene, D. L., 1992, "Vehicle Use and Fuel Economy: How Big is the Rebound Effect?" *Energy Journal*, **13**, pp. 117-143.

Gunn, H., 2001, "Spatial and Temporal Transferability of Relationships between Travel Demand, Trip Cost and Travel Time," *Transportation Research Part E*, **37**, pp. 163-189.

Hastings, N.A.J., and Peacock, J.B., 1975, *Statistical Distributions: A Handbook for Students and Practitioners*. London: Butterworth.

Hausman, J. A., and Ruud, P. A., 1987, "Specifying and testing econometric models for rank-ordered data." *Journal of Econometrics*, **34**, pp. 83-104.

Heckman, J. J., 1979, "Sample Selection Bias as a Specification Error." *Econometrica*, **47**, pp. 153-162.

Hensher, D. A., 1994, "Stated preference analysis of travel choices: the state of practice." *Transportation*, **21**, pp. 107-133. .

Horowitz, J. L., 1980, "The Accuracy of the Multinomial Logit Model as an Approximation to the Multinomial Probit Model of Travel Demand." *Transportation Research Part B*, **14**, pp. 331-341.

Japan Ministry of Health, Labour and Welfare, 1999, *Final Report of Monthly Labour Survey: July 1999*. Available at: http://www.mhlw.go.jp/english/database/db-l/, accessed Dec. 30, 2004.

Japan Research Institute Study Group on Road Investment Evaluation, 2000, *Guidelines for the Evaluation of Road Investment Projects*, Japan Research Institute, Tokyo, May.

Kain, J. F., and Liu, Z., 2002, "Efficiency and Locational Consequences of Government Transport Policies and Spending in Chile," in Glaeser, E. L., and J. R. Meyer (eds.), *Chile: Political Economy of urban Development*, pp. 105-195.

Kitamura, R., 2000, "Longitudinal Methods," in Hensher, D., and K. Button (eds.), *Handbook of Transport Modelling*, pp. 113-129.

Koppelman, F. S., and Sethi, V., 2000, "Closed-Form Discrete-Choice Models," in Hensher, D., and K. Button (eds.), *Handbook of Transport Modelling*, pp. 211-227.

Lam, T. C., and Small, K. A., 2001, "The Value of Time and Reliability: Measurement from a Value Pricing Experiment." *Transportation Research Part E*, **37**, pp. 231-251.

Louviere, J. J., and Hensher, D. A., 2001, "Combining Sources of Preference Data," in: David A. Hensher (ed.), *Travel Behaviour Research: The Leading Edge*, Pergamon, Oxford, pp. 125-144.

Manski, C. F., and Lerman, S. R., 1977, "The Estimation of Choice Probabilities from Choice Based Samples." *Econometrica*, **45**, pp. 1977-1988.

May, A.D., Coombe, D., and Gilliam, C., 1996, "The London Congestion Charging Research Programme: 3: The Assessment Methods," *Traffic Engineering and Control*, 37, pp. 277-282.

McFadden, D., 1973, "Conditional Logit Analysis of Qualitative Choice Behavior," in P. Zarembka. (eds.), *Frontiers in Econometrics*. New York: Academic Press, **pp.** 105-142.

McFadden, D., Talvitie, A. P., and Associates, 1977, *Demand Model Estimation and Validation. Urban Travel Demand Forecasting Project*. Phase I Final Report Series, Vol. V, Berkeley: University of California Institute of Transportation Studies.Special Report UCB-ITS-SR-77-9.

McFadden, D., 1978, "Modelling the Choice of Residential Location," ". in Karlqvist, A., L. Lundqvist, F. Snickars, and J. W. Weibull (eds.). *Spatial Interaction Theory and Planning Models*, Amsterdam: North-Holland, pp. 75-96.

McFadden, D., 1981, "Econometric Models of Probabilistic Choice," in Manski, C. F., and D. McFadden (eds.), *Structural Analysis of Discrete Data with Econometric Applications*, Cambridge, Mass.: MIT Press, pp. 198-272.

McFadden, D., 2001, "Economic Choices," *American Economic Review*, **91**, pp. 351-378.

McFadden, D., and Train, K., 2000, "Mixed MNL Models for Discrete Response," *Journal of Applied Econometrics*, **15**, pp. 447-470.

National Statistics Online, 2004, *Labour Force Survey (LFS) Historical Quarterly Supplement*. Available at: http://www.statistics.gov.uk/STATBASE/Expodata/Spreadsheets/D7938.xls, accessed Dec. 18, 2004.

Noland, R. B., and Small, K. A., 1995, "Travel-Time Uncertainty, Departure Time Choice, and the Cost of Morning Commutes," *Transportation Research Record*, **1493**, pp. 150-158.

Oort, C. J., 1969, "The Evaluation of Travelling Time." *Journal of Transport Economics and Policy*, **3**, pp. 279-286.

Pendyala, R. M., and Kitamura, R., 1997, "Weighting Methods for Attrition in Choice-Based Panels," in Golob, T. F., R. Kitamura, and L. Long (eds.), *Panels for Transportation Planning: Methods and Applications*, pp. 233-257.

Small, K. A., 1982, "The Scheduling of Consumer Activities: Work Trips." *American Economic Review*, **72**, pp. 467-479.

Small, K. A., and Rosen, H. S., 1981, "Applied Welfare Economics with Discrete Choice Models." *Econometrica*, **49**, pp. 105-130.

Small, K. A., and Winston, C., 1999, "The Demand for Transportation: Models and Applications," in Gomez-Ibanez, J. A., W. Tye, and C. Winston (eds.), *Transportation Policy and Economics: A Handbook in Honor of John R. Meyer*, pp. 11-55.

Small, K A., Winston, C., and Yan, J., 2005, "Uncovering the Distribution of Motorists' Preferences for Travel Time and Reliability: Implications for Road Pricing," *Econometrica*, **73**, pp. 1367-1382.

Train, K., 1980, "A Structured Logit Model of Auto Ownership and Mode Choice." *Review of Economic Studies*, **47**, pp. 357-370.

Train, K., 1986, *Qualitative Choice Analysis: Theory, Econometrics, and an Application to Automobile Demand*, Cambridge, Mass.: MIT Press.

Train, K., 2003, *Discrete Choice Methods with Simulation*, Cambridge, UK: Cambridge University Press.

Transport Canada, 1994, *Guide to Benefit-Cost Analysis in Transport Canada*, September. Available at: http://www.tc.gc.ca/finance/BCA/en/TOC_e.htm, accessed Dec. 30, 2004.

USDOT, 1997, *The Value of Travel Time: Departmental Guidance for Conducting Economic Evaluations*, US Department of Transportation, Washington.

Van Wissen, L. J. G., and Meurs, H. J., 1989, "The Dutch Mobility Panel: Experiences and Evaluation." *Transportation*, **16**, pp. 99-119.

Voith, R., 1997, "Fares, Service Levels, and Demographics: What Determines Commuter Rail ridership in the Long run?" *Journal of Urban Economics*, **41**, pp. 176-197.

Wardman, M., 1998, "The Value of Travel Time: A Review of British Evidence," *Journal of Transport Economics and Policy*, **32**, pp. 285-316.

Wardman, M., 2001, "A Review of British Evidence on Time and Service Quality Valuations," *Transportation Research Part E*, **37**, pp. 107-128.

Wardman, M., 2004, "Public Transport Values of Time," *Transport Policy*, **11**, pp. 363-377.

Washington, S. P., Karlaftis, M. G., and Mannering, F. L., 2003, *Statistical and Econometric Methods for Transportation Data Analysis*, Chapman and Hall, Boca Raton, Florida.