2. TRAVEL DEMAND

In order to plan transportation facilities, it is necessary to forecast how much they will be used. In order to price them rationally and determine the best operating policies, it is necessary to know how users respond to prices and service characteristics. In order to evaluate whether a project is worthwhile at all, it is necessary to have a measure of the benefits it produces. All these requirements are in the province of travel demand analysis.

The demand for travel takes place in a multi-dimensional setting. The traditional sequential framework used by many metropolitan transportation planning agencies considers four choice dimensions: trip generation (the total number of trips originating from an area); trip distribution (the locations of the trips' destinations); modal choice (the means of travel, such as car, bus, train, bicycle, or walking); and trip assignment (the exact route used). More recently, researchers have paid greater attention to other dimensions of choice, such as residential and job location, household automobile ownership, and the time of day at which trips are taken.

Travel is a derived demand: it is usually undertaken not for its own consumption value, but rather to facilitate a complex and spatially varied set of activities such as work, recreation, shopping, and home life. This observation links the study of travel demand to studies of labor supply, firms' choices of technologies, and urban development. It furthermore calls attention to an increasingly common form of travel: the linking together of several trip purposes into one integrated travel itinerary or tour, a process known as trip chaining.

The chapter begins (Section 2.1) by describing what happens when a conventional aggregate approach to economic demand is applied to transportation. It then moves on to disaggregate models (Section 2.2), also known as "behavioral" because they depict the individual decision-making process. Section 2.3 presents examples of models explaining some key travel choices: mode, trip time, and express-lane option. More specialized topics are then discussed (Sections 2.4-2.5). Finally, Section 2.6 analyzes two behavioral results from travel-demand studies that have special importance for policy: travelers' willingness to pay for travel-time savings and improved reliability.

Other reviews of travel demand include Bates (1997) and Small and Winston (1999). Collections of more specialized papers include Garling, Laitila and Weston (1998), Hensher and Button (2000), and Mahmassani (2002). Our discussion, like most of the literature, is mainly...
mainly about passenger transportation; studies of the demand for urban freight transportation tends to use similar methods, although with many details specific to characteristics of the industry (D’Este, 2000).

2.1 Aggregate Models

The approach most similar to standard economic analysis of consumer demand is an aggregate one. The demand for some portion of the travel market 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.

Much can be learned simply from cross-tabulations of survey data. For example, Pucher and Renne (2003) tabulate results from the 2001 National Household Travel Survey (NHTS) in the United States.\(^1\) Among the many interesting findings are that public transit accounts for just 1.6 percent of all daily trips in 2001, considerably less than walking and bicycling (9.5 percent). For work trips transit has a higher share (3.7 percent). However, there is a 15-fold difference in transit share across the nine Census regions of the US, the highest being the region that includes New York City. Nearly 92 percent of US households own a car, varying across five income groups from 73.5 percent (lowest income) to 98.5 percent (highest income).

Hu and Young (1999, Fig. 13), using 1995 U.S. data from the National Personal Transportation Survey, present variations by trip purpose and time of day. Only during early morning hours, roughly 4-7 a.m., do work trips constitute a majority. Nevertheless, it would be a mistake to conclude that work trips are no longer important for urban congestion: they account for more than half of all trips in selected Belgian cities (Van Dender, 2001, p. 103), and a very high fraction of peak-period trips on some of the most notoriously congested Los Angeles freeways (Giuliano, 1994, p. 261).

We now turn to more formal statistical techniques for measuring travel demand.

---

\(^1\) This survey continues the well-known National Personal Transportation Survey (NPTS), conducted previously at approximately six-year intervals and covering daily travel, as well as the American Travel Survey covering long-distance travel.
2.1.1 Cross-Sections of Metropolitan Areas

Several studies have used aggregate data, such as are commonly reported by transit authorities or local governments, to study influences on travel behavior across cities. Gordon and Willson (1984) compile an international data set of 91 cities with light rail systems, and estimate simple regression equations explaining the ridership (per kilometer of line) on those systems. Using the semilogarithmic form with just four variables, they show that ridership is positively related to city population density and gross national product per capita; and that it is 52 percent lower in the U.S. and 55 percent higher in eastern Europe than elsewhere, other variables being held constant.\(^2\) Applying this model to three U.S. and two Canadian cities with light-rail systems then under construction, they predict far lower ridership than the official forecasts -- about half in most cases, but less than one-seventh in the case of Detroit. It has been shown subsequently by Pickrell (1992) that nearly every modern rail system built in the U.S. has in fact attained less than half the originally forecast patronage.

Pucher (1988) compares auto and transit use in twelve western European and North American nations. He then presents data on those nations' levels of transit and highway subsidies, transit service, gasoline and motor vehicle taxes, licensing and parking policies, and land-use policies; these data suggest that the main explanations for differences in travel modes are automobile taxation and land-use policies. But he also finds that auto ownership in western Europe is growing much faster than in the U.S. and that ownership rates per capita appear to be converging to a common value.

Black (1990) uses regression analysis on data from 120 U.S. metropolitan areas to see what factors influence the fraction of metropolitan workers who walk to work. This fraction varies from 1.9 to 15.7 percent, averaging 5.4 percent. It is higher in cities with military bases or universities, in small cities, and where incomes are low. Black's paper is a useful reminder that walking can be an important journey-to-work mode, its prevalence being sensitive to land uses and demographics. More recent data suggest that the importance of walking to work has

---

\(^2\) Their model 3, p. 137.
declined, but still its share was 3.4 percent for work trips in 2001.\(^3\) Plaut (2004) reports that in Tel Aviv, 9.4 percent of workers walk to work, a fraction that varies by gender, age, and household status.\(^4\)

2.1.2 Cross-Sections within a Metropolitan Area

Statistical analysis can also be used to analyze trip-making in different parts of one metropolitan area. This approach, known as direct demand modeling, was introduced by Domencich and Kraft (1970) to explain the number of round trips between zone pairs, by purpose and mode, in the Boston area. An example is the analysis by Kain and Liu (2002, Table 5-10) of mode share to work in Santiago, Chile. The share is measured for each of 34 districts ("communas") and its logarithm is regressed on variables such as travel time, transit availability, and household income. The most powerful predictors for automobile share are vehicle ownership and household income, both of which increase the share. The share for the Metro rail system is strongly increased, quite naturally, by presence of a Metro station in the district, and is strongly decreased by high household income.

Dagenais and Gaudry (1986), analyzing Montreal data, find it important to include observations of zone pairs with zero reported trips, using the standard "tobit" model for limited dependent variables to account for the fact that the number of trips between two zones cannot be negative. This illustrates a pervasive feature of travel-demand analysis: many of the variables to be explained are limited in range so that ordinary regression analysis, which typically assumes the dependent variable to have a normal (bell-shaped) distribution, is inappropriate. For this reason, travel-demand researchers have contributed importantly to the development of techniques, discussed later in this chapter, that are appropriate for such data (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

---

\(^3\) Pucher (2003, Table 3). The decline in walking is seen in Pucher’s Table 4 for 1977-1995, which tallies all daily travel; the subsequent trend cannot be discerned from the NPTS and NHTS data due to changes in survey methodology in 2001.

\(^4\) Computed as the weighted average of “percent walking to work” for the columns pertaining to Tel Aviv in Plaut’s Table 7, p. 246.
zero and one. Instead of explaining \( x \) directly, we can explain the logistic transformation of \( x \) as follows:

\[
\log \left( \frac{x}{1-x} \right) = \beta' z + \varepsilon
\]  

(2.1)

where \( \beta \) is a vector of parameters, \( z \) is a vector of independent variables, and \( \varepsilon \) is an error term with infinite range. Equivalently,

\[
x = \frac{\exp(\beta' z + \varepsilon)}{1 + \exp(\beta' z + \varepsilon)}.
\]  

(2.2)

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 \( i \). If the characteristics in the \( z \) variables encompass all the systematic influences on mode shares, then a simple extension of equation (2.2) ensures that they sum to one:

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

(2.3)

where \( J \) is the number of modes.\(^5\) Anas (1981) and Mackett (1985b) use counts of interzonal flows to estimate models similar to this, in order to explain location and mode choice in Chicago and in Hertfordshire (England), respectively.

2.1.3 Studies Using Time Series or Panel Data

One can estimate demand equations from aggregate time-series data from a single area. For example, Greene (1992) uses US nationwide data on vehicle-miles traveled (VMT) to examine the effects of fuel prices. Several studies have examined transit ridership using data over time from a single metropolitan area or even a single transit corridor—for example Gaudry (1975) and Gómez-Ibáñez (1996) use this method to study Montreal and Boston, respectively. Time-series studies are quite sensitive to the handling of auto-correlation among the error terms, which refers to the tendency for unobserved influences on the measured dependent variable to persist over

---

\(^5\) Eqn (2.2) is a special case of (2.3) in which \( J=2 \) and we define \( x=x_1, z=z_1, z_2 \) and \( \varepsilon=\varepsilon_1, \varepsilon_2 \).
time. They may also postulate “inertia” by including among the explanatory variables one or more lagged values of the variable being explained. For example, Greene considers the possibility that once people have established the travel patterns that produce a particular level of VMT, they change them only gradually if conditions such as fuel prices suddenly change. The coefficients on the lagged dependent variables then enable one to measure the difference between short- and long-run responses, but this measured difference is especially sensitive to the treatment of autocorrelation.

It is common to combine cross-sectional and time-series variation using panel data, also called longitudinal data. Such data often combine observations from many separate locations and two or more time periods. Kitamura (2000) provides a general review. 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. Surprisingly, he find that demographic characteristics have little independent effect; rather, he suggests that much of the observed correlation between demographic factors and rail ridership arises from reverse causation. For example, a neighborhood with good rail connections to the central business district (CBD) will attract residents who work in the CBD. A corollary of this finding is that the long-run effects of changes in fares or service levels, allowing for induced changes in residential location, are considerably greater than the short-run effects. Another study using panel data is that by Petitte (2001), who estimates fare elasticities from station-level data on Metrorail ridership in Washington, D.C.

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 an unnecessary loss of efficiency and an over-statement of the precision of the estimates; for nonlinear models, it may also bias the estimates. To account for the panel structure, at least three approaches are available. One is to “first difference” all variables, so that the variable explained describes changes in some quantity rather than the quantity itself; this reduces the size of the data set by $N$ if $N$ is the number of locations. A second is to estimate a “fixed effects” model, in which a separate constant is estimated for every location; this retains all observations but adds

---

6 A good example is the study of transit use in US cities by Baum-Snow and Kahn (2000).
N-1 new coefficients to be estimated (assuming one constant would be estimated in any case). A third is a “random effects” model, in which a separate random-error term is specified that varies only by location (not time), usually with an assumed normal distribution; this specification adds only one parameter to be estimated (the standard deviation of the new error term) and so is especially useful where only a few time periods are available. Statistical tests are available to determine whether the more restrictive random-effects model is justified. Voith (1997) uses both the first-difference and fixed-effects approaches.

2.1.4 Summary of Key Results

Several literature surveys have compiled estimates of summary measures such as own- and cross-elasticities of demand for auto or public transit with respect to cost and service quality. Service quality is typically proxied by annual vehicle-miles or vehicle-hours of service.

As a rough rule of thumb, a 10 percent increase in transit fare reduces transit demand by four percent: that is, transit's own-price elasticity is approximately -0.4 on average (Pratt et al. 2000: 12-9). This elasticity is higher for trips to a central business district, trips on buses (compared to urban rail), and off-peak trips (Lago et al. 1981). Goodwin (1992) and Pratt et al. (2000, ch. 12) provide thorough reviews, which also include some studies using the disaggregate techniques described later in this chapter. Transit elasticities with respect to service quality tend to be higher, especially where service quality is poor (Chan and Ou 1978).

Demand for automobile work trips is similarly more sensitive to service quality than to cost. For example, time- and cost-elasticities are measured at -0.8 and -0.5, respectively, for Boston; and at -0.4 and -0.1 for Louisville, Kentucky (Chan and Ou 1978: 43). Overall demand for travel in personal vehicles has more often been measured as a function of fuel price and/or fuel cost per mile; typical results show elasticities between 0.1 and 0.3, with short-run elasticities typically smaller (in absolute values) than long-run elasticities. The demand for fuel itself is apparently two to three

7 Luk and Hepburn (1993); Greening, Greene, and Difiglio (2000, section 3.1.4); Graham and Glaister (2002, Table 2). Johansson and Schipper (1997) estimate a useful breakdown of changes of per capita fuel consumption into changes in vehicle stock, average fuel economy, and average usage per vehicle.
times as large, indicating that changes in fuel price affect the composition of the motor-vehicle fleet more than its usage.

2.1.5 Travel Budgets
Some researchers have suggested that better predictions of travel behavior can be obtained by making use of certain regularities in the per-capita or per-household expenditures of time and money on travel. For example, Tanner (1961) notes that distances traveled by households varied only slightly across large and small urban areas and rural areas in Britain. Schafer (2000) finds similar stability across thirty data sets in more than ten nations. In a series of unpublished papers and reports, Yacov Zehavi has made these relationships the centerpiece of a model of travel demand known as Unified Mechanism of Travel (UMOT). No satisfactory theoretical foundation has ever been provided for the model, despite a heroic attempt by Golob, Beckmann, and Zehavi (1981).

The travel-budget approach is thoroughly reviewed by Gunn (1981) and Schafer (2000), who verify many intriguing regularities. Nevertheless, the evidence suggests that more conventional models can explain them and, furthermore, that the regularities are approximations and that violations of them occur as predicted by economic theory. For example, Kockelman (2001) statistically rejects the hypothesis of fixed travel-time budgets using data from the San Francisco Bay Area. She finds that total time spent traveling declines as nearby activities are made more accessible and as distant activities are made less accessible, the latter suggesting that substitution (of nearby for distant destinations) more than offsets the direct effects of increased travel time to distant locations.

2.1.6 Transportation and Land Use
Land-use patterns are among the most important factors influencing travel decisions. This fact has led to contentious policy debates as people seek to use land-use controls to solve transportation-related problems. Giuliano (2004) provides a thorough review.

Understanding the effects of land use on travel demand requires careful disentangling of related factors. First, one must control for variables, like income and fuel price, that independently affect travel and are also correlated with land-use characteristics. Some influential studies claiming far-reaching effects of urban density on travel, such Newman and Kenworthy
(1989), have been severely criticized for neglecting this fundamental requirement of statistical analysis.

Second, should one also control for variables measuring the provision of transportation infrastructure and services? Such provision is partly a response to travel choices and to that extent, including such variables as controls could lead to an overstatement of their effect on travel and an understatement of the effect of land-use patterns (since some of their effects would be attributed instead to the transportation service variables). One way to handle this is to control for specific policies that are deemed to be exogenous, such as federal cost-sharing formulas, rather than the actual amount of infrastructure or transit service. Another is to use such policies, or other variables correlated with infrastructure and services but not directly affecting travel itself, as instrumental variables.

Third, land-use patterns themselves are not entirely exogenous, but rather respond strongly to transportation systems, especially fixed infrastructure. As a result, transportation policies may cause unexpected feedback. For example, expanding highways to relieve congestion may attract development that undermines the intended effect; this is just one of many causes of “induced demand,” discussed in Chapter 5. Expanding mass transit can even exacerbate highway congestion because the induced development, even if relatively transit-oriented, still generates many automobile trips. For example, the Bay Area Rapid Transit system in San Francisco is credited with causing Walnut Creek, an outlying station, to develop into a major center of office employment – but despite its good transit access, 95 percent of commuting trips to this center are by automobile (Cervero and Wu 1996, Table 5).

Fourth, land-use policy acts only indirectly on urban structure and thus it is misleading to treat land-use patterns as though they could be modified by fiat. Even in countries with very strong land-use authority, such as the Netherlands, land-use policies do not always bring about the changes that were intended (Schwanen, Dijst, and Dieleman, 2004). In others, such as the United States, the changes in land use that can feasibly be accomplished through policy are quite limited. As Downs (2004, ch. 12) points out, the urbanized area of Portland, Oregon, remains relatively low density despite three decades of stringent policies aimed at increasing urban density. One reason is documented by Jun (2004): Portland’s policies have apparently diverted urban development to more outlying jurisdictions outside its control.
We turn now to empirical findings on how land use affects travel demand. At the level of an entire metropolitan area, modest effects have been documented. Within a cross-section of 49 US metropolitan areas, Keyes (1982) shows that per capita gasoline consumption rises with total urban population and with the fraction of jobs located in the central business district, and falls with the fraction of people living in high-density census tracts (defined as more than 10,000 people per square mile). Gordon, Kumar, and Richardson (1989) examine average commuting time among individual respondents to the 1980 National Personal Transportation Study. Commuting time, unlike gasoline consumption, goes up with residential density; presumably this reflects longer or more congested commutes. But commuting time goes down with the proportion of the metropolitan population that lives outside the central city. This last finding suggests that polycentric or dispersed land-use patterns enable people to bypass central congestion, which in turn may explain a paradox in commuting trends: even while congestion on particular roads has gotten worse over time, average commuting times have mostly not become longer (Gordon and Richardson, 1994).

Bento et al. (2004) define a number of land-use and transportation variables related to “urban sprawl,” and ask how they influence travel choices across 114 US metropolitan areas. While no one factor explains very much of the variation, all of them taken together make a significant difference. To illustrate, the authors predict annual vehicle use for a national sample of households if they all lived in metropolitan areas with specified characteristics related to land use and transportation supply. If the characteristics are those of Atlanta, their model predicts 16,900 vehicle-miles per household; if the characteristics are changed to those of Boston, the same households would travel 25 percent less. While many characteristics contribute to this difference, the most important is population centrality, which indicates that a higher proportion of Boston’s population than Atlanta’s lives within the central portions of the total urbanized land area. The next two most important characteristics are Boston’s higher supply of rail transit and its less circular shape, both of which favor greater use of public transit. Boston also has a higher urban population density and a more even balance between jobs and households at the zip-code level, both of which contribute more modestly to its lower vehicle use.

The results just cited control for transportation infrastructure and services. If one were to view such variables instead as consequences of travel decisions, and so omit them from the model, the effects of land use are even larger – by about two-thirds in Keyes’ study.
European contexts may yield different results. Schwanen et al. (2004) examine the effects of urban size and form on travel for both commuting and shopping trips in The Netherlands. For any given mode (auto or transit), they find that the largest cities have the shortest and fastest commutes, medium-sized cities and outlying “growth centers” have the longest, while suburbs are in between. They also report exceptionally high use of walking and bicycling, amounting to 40 percent of work trips and 67 percent of shopping trips in the three largest cities, with lower rates in other areas. Specific policies in The Netherlands may explain some of these results, but also the local culture and habits, and possibly external economies of scale in cycling, will matter (and may in fact partly justify these policies).

Turning to the neighborhood level, the evidence on how land use affects travel is quite mixed. Crane (2000) provides a useful review. It is clear that high-density neighborhoods near transit stops support higher use of public transit. It is less clear how much of this is due simply to sorting of households: those who want to use transit choose transit-oriented neighborhoods. This is not a concern if one simply wants to predict transit patronage in a proposed transit-oriented development, but if one wants to know the aggregate effects of building many such developments one needs to control statistically for sorting. When this is done, the effects of land use on the travel behavior are found to be much more modest, and to depend on a number of collateral factors such as how centralized the entire urban area is. For example, Cervero and Gorham (1994) find that transit-oriented development encourages more walking and bicycling in local areas within the San Francisco region, but not in those within the Los Angeles region.

One of the factors limiting the influence of land use is that people travel far greater distances than are required by land-use patterns alone (Giuliano and Small, 1993). Even when jobs-housing balance is achieved within a community, people do not predominately choose nearby jobs (Giuliano, 1991; Cervero, 1996). Thus, for example, new exurban communities intended to be relatively self-contained have generally discovered that most residents work elsewhere, even in the Netherlands (Schwanen et al., 2004).

Recent work has suggested that most of the effect of land use on travel can be understood by examining how trip costs are affected and how people respond to changes in trip costs (Boarnet and Crane, 2000). This observations opens the door to research explaining more specifically why land use has impacts in some settings and not others.
2.2 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 decision-making units), this approach explains behavior directly at the level of a person, household, or firm. Disaggregate models are more efficient in their use of survey data when it is available, and are based on a more satisfactory microeconomic theory of demand, a feature that is particularly useful when applying welfare economics. Most such models analyze choices among discrete rather than continuous alternatives and so are called *discrete-choice models*.

Discrete-choice modeling of travel demand has mostly taken advantage of data from large and expensive transportation surveys. Deaton (1985) shows that it can also be used with household-expenditure surveys, which are often conducted for other purposes and are frequently available in developing nations.

2.2.1 Basic Discrete-Choice Models

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

\[
U_{jn} = V(z_{jn}, s_n, \beta) + \epsilon_{jn}
\]

(2.4)

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 decision maker, \( s_n \) is a vector of characteristics of the decision maker (effectively allowing different utility structures for different identifiable groups of decision makers), \( \beta \) is a vector of unknown parameters, and \( \epsilon_{jn} \) is an unobservable component of utility which captures idiosyncratic preferences. \( U_{jn} \) and \( V \) are known as *conditional indirect utility* functions, since they are conditional on choice \( j \) and, just like the indirect utility function of standard consumer theory, they implicitly incorporate a budget constraint.

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 \( \epsilon_{jn} \). Once a

---

8 Reviews with a transportation focus include Ben-Akiva and Bierlaire (1999), Koppelman and Sethi (2000), and the books by Ben-Akiva and Lerman (1985) and Train (2003).
functional form for \( V \) is specified, the model becomes complete by specifying a joint cumulative distribution function (cdf) for the random terms, \( F(\varepsilon_{1n}, \ldots, \varepsilon_{Jn}) \). Denoting \( V(z_{jn}, s_{in}, \beta) \) by \( V_{jn} \), the choice probability for alternative \( i \) is then

\[
P_{in} = \Pr[U_{in} > U_{jn} \text{ for all } j \neq i]
= \Pr[\varepsilon_{jn} - \varepsilon_{in} < V_{in} - V_{jn} \text{ for all } j \neq i]
= \int_{-\infty}^{\infty} F_i(V_{in} - V_{jn} + \varepsilon_{jn}, \ldots, V_{in} - V_{jn} + \varepsilon_{in}) d\varepsilon_{in}
\]

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 \( \varepsilon_{in} \) conditional on the inequalities in (2.5).)

Suppose the cdf \( F(\cdot) \) is multivariate normal. Then (2.5) is the multimonial probit model with general covariance structure. However, neither \( F \) nor \( F_i \) can be expressed in closed form; instead, equation (2.5) 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 \((J=2)\), (2.5) becomes

\[
P_{1n} = \Phi\left(\frac{V_{1n} - V_{2n}}{\sigma}\right)
\]

where \( \Phi \) is the cumulative standard normal distribution function (a one-dimensional integral) and \( \sigma \) is the standard deviation of \( \varepsilon_{1n}, \varepsilon_{2n} \). In equation (2.6), \( \sigma \) cannot be distinguished empirically from the scale of utility, which is arbitrary; for example, doubling \( \sigma \) has the same effect as doubling both \( V_1 \) and \( V_2 \). Hence it is conventional to normalize by setting \( \sigma=1 \).

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

\[
Pr[\varepsilon_{jn} < x] = \exp(-e^{-\mu x})
\]

for all real numbers \( x \), where \( \mu \) is a scale parameter. Here the convention is to normalize by setting \( \mu=1 \). With this normalization, McFadden (1973) shows that the resulting probabilities calculated from (2.5) have the logit form:
\[ P_{in} = \frac{\exp(V_{in})}{\sum_{j=1}^{J} \exp(V_{jn})}. \]  

(2.8)

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 \(-\infty\) 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 (2.8), i.e. the form with \( J=2 \), is:

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

If graphed as a function of \( (V_{1n}-V_{2n}) \), this equation looks quite similar to (2.6).

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). 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 later.

As for the functional form of \( V \), by far the most common is linear in unknown parameters \( \beta \). More general forms such as Box-Cox and Box-Tukey transformations are studied by Gaudry and Wills (1978). Note that even with \( V \) forced to be linear in parameters, it can easily be made comparable normalization is accomplished by dividing the logit coefficients by \( \pi/\sqrt{3} \) in order to give the utilities the same standard deviations in the two models. In both models, the choice probabilities depend on \( (\beta/\sigma) \), where \( \sigma^2 \) is the variance of each of the random terms \( e_{in} \). In the case of probit, the variance of \( e_{1n} e_{2n} \) is set to one by the conventional normalization; hence \( \sigma^2_{\text{PROBIT}} = 1/2 \). In the case of logit, the normalization \( \mu=1 \) in equation (2.7) implies that \( e_{in} \) has standard deviation \( \sigma^2_{\text{LOGIT}} = \pi/\sqrt{6} \) (Hastings and Peacock, 1975, p. 60). Hence to make logit and iid probit comparable, the logit coefficients must be divided by \( \sqrt{\pi/3} = 1.814 \).
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:

$$V_{in} = \beta_1 \cdot \left( c_{in} / w_n \right) + \beta_2 T_{in} + \beta_3 T_{in}^2.$$  \hspace{1cm} (2.9)

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 2.4), the linear-in-parameters specification is simply written as

$$V_{in} = \beta z_{in}$$  \hspace{1cm} (2.10)

where $\beta$ is the transpose of column vector $\beta$.

### 2.2.2 Estimation

For a given model, data on actual choices, along with traits $z_{jn}$, can be used to estimate the unknown parameter vector $\beta$ in (2.10) and to carry out statistical tests of the specification (i.e., tests of whether the assumed functional form of $V$ and the assumed error distribution are valid). Parameters are usually estimated by maximizing the log-likelihood function:

$$L(\beta) = \sum_{n=1}^{N} \sum_{i=1}^{J} d_{in} \log P_{in}(\beta)$$  \hspace{1cm} (2.11)

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

A correction to (2.11) is available for choice-based samples, i.e., those in which the sampling frequencies depend on the choices made. 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). This correction does not, however, make efficient use of the information on aggregate mode shares that it requires. Imbens and Lancaster (1994) show quite generally how to incorporate aggregate information to greatly improve the efficiency of disaggregate econometric models, while Berry, Levinsohn, and Pakes (1995) show that with assumptions about firm behavior it is even possible even to estimate the parameters of a micro-level model using only aggregate data.
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 (2.8). With \( V \) linear in \( \beta \), the logit log-likelihood function is globally concave in \( \beta \), so finding a local maximum assures finding the global maximum. Fast computer routines to do this are widely available. In contrast, computing the log-likelihood function for multinomial probit with \( J \) alternatives entails computing for each member of the sample the \((J-1)\)-dimensional integral implicit in equation (2.5). This has generally proven difficult for \( J \) larger than 3 or 4, despite the development of computational-intensive simulation methods (Train 2003).

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 (2.4) 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. The problem is that income does too good a job as an explanatory variable: within this data set, the model exuberantly declares high income to make the alternative of owning a car infinitely desirable relative to not owning one. We know of course that this is not true and that a larger sample would contain counter-examples—even in the US, 1.5 percent of the highest-income households owned no car in 2001 (Pucher 2003). Given the sample we have, 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; despite certain statistical disadvantages, this model is able to measure the coefficient in question (Caudill, 1988) because there is a limit to how much income can affect probability.

2.2.3 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. The most common example is that travel times depend on congestion, which depends on amount of travel, which depends on travel times. Thus the actual use of a travel demand model may require a process of equilibration in which a solution is sought to a set of simultaneous relationships. An
elegant formulation of supply-demand equilibration on a congested network is provided in the remarkable study by Beckmann, McGuire, and Winsten (1956). Boyce, Mahmassani, and Nagurney (2005) provide a readable review of its history and subsequent impact.

With aggregate data, the endogeneity of travel characteristics is an important issue for obtaining valid statistical estimates of demand parameters. Fortunately, endogeneity can usually be ignored when using disaggregate data because, from the point of view of individual decision-making, the travel environment does not depend appreciably on that one individual’s decisions. Nevertheless, measuring the values of attributes $z_{it}$, 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 be quite different from the measures employed in policy analysis or forecasting. People know even less about alternatives they do not choose. Hence even if reported values accurately measure the perceptions that determine choice, the resulting models cannot be used for prediction unless one can predict how a given change will alter those perceptions. Worse still, the reports 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 study by MVA Consultancy et al. (1987, pp. 159-163) finds such bias to be severe in a study of the Tyne River crossing in England. In this case the explanatory variables of the model are endogenous to the choice, which makes the estimated model appear to fit very well (a typical finding for studies using reported values) but which renders it useless for prediction.

Objective estimates of travel attributes, on the other hand, may be very expensive and not necessarily accurate. Even something as simple as the travel time for driving on a particular highway segment at a particular time of day is quite difficult to ascertain. Measuring the day-to-day variability of that travel time is even more difficult. Three recent studies in California have accomplished this, one by applying sophisticated algorithms to data from loop detectors placed
in the highway and two by using the floating-car method, in which a vehicle with a stopwatch is driven so as to blend in with the traffic stream.\textsuperscript{10}

Ideally, one might formulate a model in which perceived attributes and actual choice are jointly determined, each influencing the other and both influenced by objective attributes and personal characteristics. This type of model most faithfully replicates the actual decision process. However, it is doubtful that the results would be worth the extra complexity unless there is inherent interest in perception formation for marketing purposes. For purposes of transportation planning, we care mainly about the relationship between objective values and actual choices. A model limited to this relationship may be interpreted as the reduced form of a more complex model including perceptions, so it is theoretically valid even though perception formation is only implicit. Hence the most fruitful expenditure of research effort is usually on finding ways to measure objective values as accurately as possible.

In a large sample, a cheaper way to compute objective values may be to assign values for a given alternative according to averages reported by people in the sample in similar circumstances who use that alternative. While subject to some inaccuracy, this at least eliminates endogeneity bias by using an identical procedure to assign values to chosen and unchosen alternatives.

The data described thus far measures \textit{revealed preference} (RP) information, that reflected in actual choices. There is growing interest in using \textit{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 by applying an appropriate experimental design (see for example Louviere, Hensher, and Swait, 2000). If administered in interviews using a portable computer, the questions posed can be adapted to information about the respondent collected in an earlier portion of the survey – as for example in the study of freight mode choice in India by Shinghal and Fowkes (2002). SP surveys also can elicit information about potential travel options not now available. It is still an open question, however, how accurately they describe what people really do.

It is possible to combine data from both revealed and stated preferences in a single estimation procedure in order to take advantage of the strengths of each (Ben-Akiva and

\textsuperscript{10} Brownstone \textit{et al.} (2003); Lam and Small (2001); Small, Winston, and Yan (2005).
Morikawa, 1990; 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, or more generally just to account for differences in surveys, it is recommended to estimate certain parameters separately in the two portions of the data: the scale factors \( \mu \) 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. For example, in the logit model of (2.8) and (2.12), one might constrain all parameters to be the same for RP and SP observations except for the scale, alternative-specific constants, and the first variable \( z_{in} \). Letting \( \beta'_{2i} z_{2in} \) represent the rest of \( \beta z_{in} \), adding superscripts for the parameters assumed distinct in the two data subsamples, and normalizing the RP scale parameter to one, the log-likelihood function (2.11) becomes the following:

\[
L(\alpha, \beta, \mu^{SP}) = \sum_{n \in RP} \sum_{i=1}^{J} d_{in} \left\{ \alpha_i^{RP} + \beta_i^{RP} z_{1in} + \beta_i^{SP} z_{2in} - \log \sum_{j=1}^{J} \exp(\alpha_j^{RP} + \beta_j^{RP} z_{1jn} + \beta_j^{SP} z_{2jn}) \right\} \\
+ \sum_{n \in SP} \sum_{i=1}^{J} d_{in} \left\{ \mu^{SP} \cdot (\alpha_i^{SP} + \beta_i^{SP} z_{1in} + \beta_i^{SP} z_{2in}) - \log \sum_{j=1}^{J} \exp(\mu^{SP} \cdot (\alpha_j^{SP} + \beta_j^{SP} z_{1jn} + \beta_j^{SP} z_{2jn})) \right\}
\]

where \((\alpha, \beta, \mu^{SP})\) denotes the entire set of parameters shown on the right-hand side (excluding \( \alpha_i^{RP} \) and \( \alpha_i^{RP} \), which can be normalized to zero). As before, the prime after a column vector indicates its transpose, so that it becomes a row vector. This expression is not as complicated as it looks: the first term in curly brackets is just the logarithm of the logit probability (2.8) for RP observations, while the second is the same thing for SP observations except utility \( V_{in} \) is multiplied by scale factor \( \mu^{SP} \).

2.2.4 Interpreting Coefficient Estimates

It is useful for interpreting empirical results to note that a change in \( \beta z_{in} \) in (2.10) by an amount of \( \pm 1 \) 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, multiplied by its relevant coefficient – if the result is on the order of 1.0 or larger, such variations have large effects on the relative odds. In fact some authors prefer to provide this information by listing, in addition to or instead of the coefficient estimates, the marginal effect of a specified change in the
independent variable on the probabilities; a disadvantage of this measure is that it depends on the values of the variables and so makes comparisons across models more difficult.

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

\[ V_{in} = \alpha_i + \beta' z_{in}. \tag{2.12} \]

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.

The constant \( \alpha_i \) may be interpreted as the average utility of the unobserved characteristics of the \( i \)-th alternative, relative to the base alternative. In a sense, specifying these constants is admitting the inadequacy of variables \( z_{in} \) to explain choice; hence the constants' estimated values are especially likely to reflect circumstances of a particular sample rather than universal behavior. The use of alternative-specific constants also 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. Quandt and Baumol (1966) coined the term "abstract mode" to indicate the desire to describe a travel mode entirely by its objective characteristics, rather than relying on alternative-specific constants. In practice, however, this goal is rarely achieved.

Equation (2.12) is really a special case of (2.10) in which one or more of the variables \( Z \) are alternative-specific dummy variables, \( D_k \), defined by \( D_{jn} = 1 \) if \( j = k \) and 0 otherwise (for each \( j = 1, \ldots, J \)). (Such a variable does not depend on \( n \).) In this notation, parameter \( \alpha_i \) in (2.12) is viewed as the coefficient of variable \( D^i \) included among the \( z \) variables in (2.10). 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 \( \beta \), thus allowing (2.10) still to represent the linear-in-parameters specification. An SP experiment designed such that alternative-specific coefficients can be estimated for all attributes is sometimes called a “labeled” experiment.

The most economically meaningful quantities obtained from estimating a discrete-choice model are often ratios of coefficients, which represent marginal rates of substitution. 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 is the marginal rate of substitution between time and money in the conditional indirect utility function, often called the value of travel-time savings, or value of time for short. It represents the monetary value that the traveler places on time savings, and is very important in evaluating the benefits of transportation improvements whose primary effects are to improve people's mobility. The value of time in the model (2.9) is

\[
(v_T)_n = -\frac{dc_{in}}{dT_{in}} = \frac{\partial V_{in}}{\partial T_{in}} = \left( \frac{\beta_2 + 2\beta_3 T_{in}}{\beta_1} \right) \cdot w_n ,
\]

(2.13)

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

As a more complex example, suppose we extend equation (2.9) by adding alternative-specific dummies, both separately (with coefficients \( \alpha_i \)) and interacted with travel time (with coefficients \( \gamma_i \)):

\[
V_{in} = \alpha_i + \beta_1 \cdot \left( c_{in} / w_n \right) + \beta_2 T_{in} + \beta_3 T_{in}^2 + \gamma_i T_{in}
\]

(2.14)

where one of the \( \alpha_i \) and one of the \( \gamma_i \) are normalized to zero. This yields the following value of time applicable when individual \( n \) chooses alternative \( i \):

\[
(v_T)_n = \left( \frac{\beta_2 + 2\beta_3 T_{in} + \gamma_i}{\beta_1} \right) \cdot w_n ,
\]

(2.15)

Now the value of time varies across modes even with identical travel times, due to the presence of \( \gamma_i \). There is a danger, however, in interpreting such a model. What appears to be variation in value of time across modes may just reflect selection bias: people who, for reasons we cannot observe, have high values of time will tend to self-select onto the faster modes (MVA Consultancy et al., 1987, pp. 90-92). This possibility can be modeled explicitly using a random-coefficient model, described later in this chapter.

Confidence bounds for a ratio of coefficients, or for more complex functions of coefficients, can be estimated by standard approximations for transformations of normal variates. Specifically, if vector \( \beta \) is asymptotically normally distributed with mean \( b \) and variance-covariance matrix \( \Sigma \), then a function \( f(\beta-b) \) is asymptotically normally distributed with mean zero
and variance-covariance matrix $(\nabla f) \Sigma (\nabla f)'$, where $\nabla f$ is the vector of partial derivatives of $f$.\footnote{Chow (1983: 182-3). The result requires asymptotic convergence of $\beta$ at a rate proportional to the square root of the sample size. (That is, as the sample size $N$ increases, the difference between the estimated and true values of $\beta$ tends to diminish proportionally to $1/\sqrt{N}$.) In the simple case where $v = \beta_1/\beta_2$, it implies that the standard deviation $\sigma_v$ of $v$ obeys the intuitive formula: $(\sigma_v^2) \equiv (\sigma_1^2/\beta_2^2) + (\sigma_2^2/\beta_1^2) - 2\sigma_1\sigma_2(\beta_1/\beta_2)$, where $\sigma_1$ and $\sigma_2$ are the standard deviations of $\beta_1$ and $\beta_2$, and where $\sigma_1\sigma_2$ is their covariance. Such an approximation requires that $\sigma_1/\beta_1$ and $\sigma_2/\beta_2$ be small, which in turn helps ensure that the variance of $\beta_1/\beta_2$ exists (it would not exist, for example, if the mean of $\beta_2$ were zero).} A more accurate estimate may be obtained by taking repeated random draws $\beta_r$ from the distribution of $\beta$, which is estimated along with $\beta$ itself, and then examining the resulting values $f(\beta_r)$. As an example, the 5th and 95th percentile values of those values define a 90 percent confidence interval for $\beta$. See Train (2003, ch. 9) for how to take such random draws.

### 2.2.5 Randomness, Scale of Utility, and Measures of Benefit

The variance of the random utility term in equation (2.4) 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 (2.8), suppose systematic utility is linear in parameter vector $\beta$ as in (2.10). If all the elements of $\beta$ are small in magnitude, the corresponding variables have little effect on probabilities so choices are dominated by randomness. If the elements of $\beta$ 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, or entropy, in aggregate behavior. Indeed, it can be measured by the entropy-like quantity $-\Sigma_j \Sigma_n P_{jn} \log P_{jn}$, which is larger when the choice probability is divided evenly among the alternatives than when one alternative is very likely and others very unlikely. Anas (1983) derives the disaggregate logit model by maximizing an aggregate objective function that includes such an entropy term, subject
to constraints that guarantee consistency with observed aggregate shares and average values of characteristics. Similarly, Anderson et al. (1988) show that the aggregate logit model can be derived by maximizing a utility function for a representative traveler that includes an entropy term, subject to a consistency constraint on aggregate choice shares. Thus entropy is a link between aggregate and disaggregate models: at the aggregate level we can say the system tends to favor entropy or that a representative consumer craves variety, whereas at the disaggregate level we represent the same phenomenon as randomness in utility.

It is sometimes useful to have a measure of the overall desirability of the choice set being offered to a decision maker. 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, i.e. determined solely by the ranking of $V_i$, the decision maker 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 $V_i$ 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:

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

where $\gamma=0.5772$ is Euler’s constant (it accounts for the nonzero mean of the error terms $\varepsilon_j$ in the standard normalization). Here we have retained the parameter $\mu$ from (2.7), rather than normalizing it, to make clear how randomness affects expected utility. When the amount of randomness is small (large $\mu$), 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 $\mu^{-1} \log[\exp(V_{j*}/\rho)] = V_{j*}$, the utility of the dominating alternative. When randomness dominates (small $\mu$), 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 \( \lambda_n \), the marginal utility of income; thus a useful welfare measure for such a set of individuals, with normalization \( \mu=1 \), is:

\[
W = \frac{1}{\lambda_n} \log \sum_{j=1}^{J} \exp(V_{jn}),
\]

(2.17)
a formula also derived by Williams (1977). (The constant \( \gamma \) drops out of welfare comparisons so is omitted.) Because portions of the utility \( V_i \) that are common to all alternatives cannot be estimated from the choice model, \( \lambda_n \) cannot be estimated directly.\(^\text{12}\) However, typically it can be determined from Roy's Identity:

\[
\lambda_n = -\frac{1}{x_{in}} \frac{\partial V_{in}}{\partial c_{in}}
\]

(2.18)
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). Expression (2.18) is valid provided that its right-hand-side is independent of \( i \); when it is not, tractable approximations are available (Chattopadhyay, 2001).

\[2.2.6\text{ Aggregation and Forecasting}\]

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 decision makers is drawn, each assumed to represent a subpopulation with identical observable

12 If income \( y \) is included as an explanatory variable, it might be tempting to simply compute \( \partial V_j / \partial y \) as a measure of \( \lambda \). This is completely wrong because the indirect utility is strongly influenced by income, independently of which alternative is chosen, whereas \( V_j \) captures only the relative effects of income on utility of the various alternatives.
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.

One can simulate the effects of a policy by determining how it changes the values of independent variables for each sample member, and recomputing the predicted probabilities accordingly. Doing so requires that these variables be explicitly included in the model. For example, to simulate the effect of better schedule coordination at transfer points on a transit system, the model must include a variable for waiting time at the transfer points. Such a specification is called policy-sensitive, and its absence in earlier aggregate models was one of the main objections to the traditional travel-demand modeling framework. The ability to examine complex policies by computing their effects on an enumeration sample is one of the major advantages of disaggregate models.

Aggregate forecasts may display a sensitivity to policy variables that is quite different from a naive calculation based on a representative individual. For example, suppose the choice between travel by automobile (alternative 1) and bus (alternative 2) is determined by a logit model with utilities given by equation (2.9) with $\beta_2 = 0$. Then the probability of choosing bus travel is:

$$P_{2n} = \frac{1}{1 + \exp\left[\left(\frac{\beta_1}{w_n}\right)\cdot(c_{1n} - c_{2n}) + \beta_2 \cdot (T_{1n} - T_{2n})\right]}.$$  \hspace{1cm} (2.19)

Suppose everyone's bus fare is $c_2$ and everyone's wage is $w$. Then

$$\frac{\partial P_{2n}}{\partial c_2} = \left(\frac{\beta_1}{w}\right) \cdot P_{2n} \cdot (1 - P_{2n}).$$  \hspace{1cm} (2.20)

Now suppose half the population has conditions favorable to bus travel, such that $P_{2n} = 0.9$; whereas the other half has $P_{2n} = 0.1$. Aggregate bus share is then 0.5. The rate of change of aggregate bus share with respect to bus fare is, from (2.20), $\left(\frac{\beta_1}{w}\right) \cdot \left[\frac{1}{2}(0.9)(0.1) + \frac{1}{2}(0.1)(0.9)\right] = 0.09 \cdot \left(\frac{\beta_1}{w}\right)$. But if we were to calculate it from (2.20) as though there were a single representative traveler with $P_2 = 0.5$, we would get $\left(\frac{\beta_1}{w}\right)(0.5)(0.5) = 0.25 \cdot \left(\frac{\beta_1}{w}\right)$. This would
overestimate the true sensitivity by 178 percent. Again, the existence of variety reduces the actual sensitivity to changes in independent variables, in this case because there are only a few travelers (those with extreme values of $\varepsilon_1 - \varepsilon_2$) who have a close enough decision to be affected.

McFadden and Reid (1976) derive an especially illuminating result illustrating this phenomenon in the case of a binary probit model where the independent variables are normally distributed in the population. They show that if a single individual's choice probability (2.6) is written in the form

$$P_i = \Phi(\beta' z_i),$$

then the expected aggregate share $\bar{P}_i$ for the subpopulation represented by this individual is given by

$$\bar{P}_i = \Phi\left(\frac{\beta' \bar{z}}{\sqrt{1 + \sigma^2}}\right)$$

where $\bar{z}$ and $\sigma^2$ are the average of $z$ and the variance of $\beta' z$, respectively, within this subpopulation. Once again, the existence of population variance reduces policy sensitivity and causes the naive calculation using an average traveler (equivalent to setting $\sigma=0$) to overestimate that sensitivity.

Equation (2.21) illustrates a danger in using aggregate models for policy forecasts. If an aggregate probit model fitting $\bar{P}_i$ to $\bar{z}$ were estimated, its coefficients would correspond to $\beta / \sqrt{1 + \sigma^2}$. If a policy being investigated changed $\sigma$, these coefficients would no longer accurately represent behavior under the new policy.

2.2.7 Specification Searches

Like most applied statistical work, travel demand analysis requires balancing completeness against tractability. The model that includes every relevant influence on behavior may require too much data to estimate with adequate precision, or it may be too complex to serve as a practical guide to policy analysis. A related problem, also common to most empirical work, is that the statistical properties of the model, such as standard errors of estimated coefficients, are valid only when the model’s basic assumptions are known in advance to be correct. But in practice the researcher normally chooses a model’s specification (i.e. its functional form and set
of included variables) using guidance from the same data as those from which its parameters are then estimated.

A good way to handle both problems is to base empirical models on an explicit behavioral theory. Rather than try out dozens of specifications to see what fits, one gives preference to relationships that are predicted by a plausible theory. For example, a specification like (2.9) would be chosen if there is good theoretical reason to think the value of time is proportional to the wage rate—a question explored later in this chapter.

Bayesian methods offer a more formal approach to using prior information or judgments when specifying empirical models. Instead of all-or-nothing decisions about model structure, they allow one to explicitly describe prior uncertainty and to calculate the manner in which prior beliefs need to be modified in light of the data. Such methods have recently been developed for parameter estimation in discrete choice models (Train 2003, ch. 12). Bayesian methods for model selection, in which the data (along with prior beliefs) determine explicit probabilities for competing model structures, are also available and could be usefully applied to transportation problems. See Berger and Pericchi (2001) for a good introduction.

2.2.8 Transferability
One of the goals of disaggregate travel-demand modeling is to describe behavioral tendencies that are reasonably general. This would enable a model estimated in one time and place to be used for another. The progress toward this goal has been disappointing, but some limited success has been achieved by making certain adjustments. Notably, the alternative-specific constants and the scale of the utility function are often found to be different in a new location, presumably because they reflect our degree of ignorance which may vary from one setting to another. Such adjustments can be made relatively inexpensively by using limited data collection in a new location or, in the case of alternative-specific constants, just by adjusting them to match known aggregate shares (Koppelman and Wilmot, 1982; Koppelman and Rose, 1985).

2.2.9 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” $y_n^* = \beta' z_n + \epsilon_n$, with choice $j$ occurring if this latent variable falls in a particular interval $[\mu_{j-1}, \mu_j]$ of the real line, where $\mu_0 = -\infty$ and $\mu_J = \infty$. The interval boundaries $\mu_1, \ldots, \mu_{J-1}$ are estimated along with $\beta$, except that one of them can be normalized arbitrarily if $\beta' z_n$ contains a constant term. The probability of choice $j$ is then

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

(2.22)

where $F(\cdot)$ is the cumulative distribution function assumed for $\epsilon_n$. In the ordered probit model $F(\cdot)$ is standard normal, while in the ordered logit model it is logistic, i.e. $F(x) = [1 + \exp(-x)]^{-1}$. Thus probabilities depend entirely on a single index, $\beta' z_n$, calculated for individual $n$. When this index is strongly positive, all the terms $F(\mu_j - \beta' z_n)$ are small except for the last, $F(\mu_J - \beta' z_n) = F(\infty) = 1$, so the most likely choice will be alternative $J$. When the index is strongly negative, the most likely choice will be alternative 1. At intermediate values it becomes more likely that alternatives between 1 and $J$ will be chosen. 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, ch. 10).

Sometimes information is available not only on the most preferred alternative, but on the individual’s ranking of other alternatives. In this case, we effectively observe “choices” among numerous situations, including some where the most preferred alternative is hypothetically absent. Efficient use can be made of such data through the rank-ordered logit model analyzed by Beggs, Cardell, and Hausman (1981) and Hausman and Ruud (1987). In the case where a complete ranking of $J$ alternatives is obtained, the probability formula for rank-ordered logit is a product of $J$ logit probability formulas, one for each ranked alternative, giving the probability of

---

13 Rank-ordered logit is sometimes called “expanded logit” or “exploded logit.” Beggs et al. call it “ordered logit,” but that name is now usually reserved for an ordered response model as described here.
choosing that alternative from the set of itself and all lower-ranked alternatives. One may want to ignore the stated ordering among some low-ranked alternatives, or alternatively to estimate a separate scale factor for those choices, to allow for the possibility that a respondent pays less attention when answering questions about alternatives of little interest.

2.3 Examples of Disaggregate Models

Discrete-choice models have been estimated for nearly every conceivable travel decision, forming a body of research that cannot possibly be reviewed here.\(^\text{14}\) In some cases, these models have been linked into large simultaneous systems requiring extensive computer simulation. An example is the system of models developed to analyze a proposal for congestion pricing in London (Bates et al. 1996).

In this section we present three very modest disaggregate models, each chosen for its compact representation of a behavioral factor that is central to urban transportation policy as analyzed in later chapters.

2.3.1 Mode Choice

Kenneth Train (1978, 1980) and colleagues have developed a series of models explaining automobile ownership and commuting mode, estimated from survey data collected before and after the opening of the Bay Area Rapid Transit (BART) system in the San Francisco area. Here we present one of the simplest, explaining only mode choice: the "naive model" reported by McFadden et al. (1977, pp. 121-123). It assumes 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 BART system.

Mode choice is explained by just three independent variables plus 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 \(T_{in}^\alpha\), the out-of-vehicle travel time including walking, waiting, and

\(^{14}\) For additional examples, see McCarthy (2001, ch. 3-4) and Small and Winston (1999).
transferring. Cost $c_{in}$ includes parking, tolls, gasoline, and maintenance (Train, 1980, p. 362).

The estimated utility function is:

$$V = -0.0412 \cdot \frac{c}{w} - 0.0201 \cdot T - 0.0531 \cdot T'^{0} - 0.89 \cdot D^{1} - 1.78 \cdot D^{3} - 2.15 \cdot D^{4}$$ (2.23)

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.

This utility function is a simplification of (2.14) (with $\beta_{3} = \gamma_{i} = 0$), except that travel time is broken into two components, $T$ and $T^{0}$. Adapting (2.15), we see that the "value of time" for each of these two components is proportional to the post-tax wage rate, the proportionality constant being the ratio of the corresponding time-coefficient to the coefficient of $c/w$. Hence the 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 (2.23), 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.

2.3.2 Trip-Scheduling Choice

One of the key decisions affecting congestion is the timing or scheduling of work trips. There is now a substantial body of empirical work on this subject, reviewed by Mahmassani (2000).

Although the scheduling decision is inherently continuous, most authors model it as a discrete choice among time intervals. There are two reasons for this: survey responses are
rounded off to a few even numbers, and disaggregate models can easily portray the complex manner in which travel time varies across possible schedules. Small (1982) estimates the choice among twelve possible five-minute intervals for work arrival time, using a set of auto commuters from the San Francisco Bay Area who have an official work-start time. The data set includes characteristics of the workers and a network-based engineering calculation of the travel time that each would encounter at each arrival time. Commuters are assumed to have full information; reliability of arrival is not considered except that the specification allows the commuter to avoid (through an estimated utility penalty) arriving just in time for work.

The utility specification postulates a linear penalty for arriving early, on the assumption that time spent before work is relatively unproductive; and a much larger linear penalty for arriving late, on the assumption that employer sanctions take hold with gradually increasing severity. 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. Define "Schedule Delay Late," \( SDL \), as \( \max\{S_D,0\} \) and "Schedule Delay Early," \( SDE \), as \( \max\{-S_D,0\} \). 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.

The utility function estimated by Small (1982, Table 2, Model 1), with estimated standard errors in parentheses, is:

\[
V = -0.106 \cdot T - 0.065 \cdot SDE - 0.254 \cdot SDL - 0.58 \cdot DL. 
\]

\[(2.24)\]

This excludes two variables used to represent a tendency of respondents to round off answers to the nearest 10 or 15 minutes. More complex models are also estimated, in which the various penalties are nonlinear or depend upon such factors as the worker's family status or occupation, whether solo driver or carpooler, and how much flexibility for late arrivals the employer is said to allow.

Figure 2.1 shows utility function (2.24), divided by the coefficient of travel time. The marginal rates of substitution indicate that the commuter is willing to suffer an extra 0.61 minutes of congestion to reduce the amount of early arrival by one minute;\(^ {15}\) and 2.40 minutes of congestion to reduce late arrival by one minute, plus an extra 5.47 minutes congestion to avoid

\(^{15}\) Calculated as 0.065/0.106=0.61.
any of the just-on-time or late alternatives. These turn out to be key parameters in models, to be
presented in the next chapter, which describe equilibrium when congestion occurs in the form of
queueing behind a bottleneck. They also can be used to formulate models of traveler response to
network unreliability, as described by Bates et al. (2001).

These alternatives have a natural ordering; so why is the ordered response model not
used? There are two reasons. First, as already noted, the ordered response model cannot take
advantage of information that varies by alternative, such as travel time. Second, even accounting
just for socioeconomic variables, there is no plausible combination of them that would exert a
monotonic influence on the time of day; rather, there are likely to be some variables that favor
peak times, others that favor times either earlier or later than that, others still that would affect
the strength of preference for low travel times, and so forth.

2.3.3 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, documented in the previous subsection, that people are more averse to unexpected delays
than to unexpected early arrivals.

The model explains a pair of related decisions: (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 decisions 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:16

\[ V = -0.862 \cdot D^{ag} + 0.0239 \cdot Inc \cdot D^{ag} - 0.766 \cdot ForLang \cdot D^{ag} - 0.789 \cdot D^3 \]

\[ (0.411) \quad (0.0058) \quad (0.412) \quad (0.853) \]

\[ -0.357 \cdot c - 0.109 \cdot T - 0.159 \cdot R + 0.074 \cdot Male \cdot R + \text{(other terms)}. \]  

(2.25)

(0.138) 0.056) (0.048) (0.046)

Here \( D^{ag} = D^2 + D^3 \) 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 point estimates, although the difference (i.e. the coefficient of \( Male \cdot R \)) is not quite statistically significant even at a 10-percent significance level. Several studies of this particular toll facility have found women noticeably 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.

2.4 Advanced Discrete-Choice Modeling

2.4.1 Generalized Extreme Value Models

16 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.

17 The coefficient for women is -0.159, and that for men is -0.159+0.074 = -0.085.
Often it is implausible that the additive random utility components $\epsilon_j$ be independent, 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 (2.5) 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.\(^{18}\)

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

$$F(\epsilon_1, ..., \epsilon_J) = \exp\left[-G(e^{-\epsilon_1}, ..., e^{-\epsilon_J})\right]$$

where $G$ is a function satisfying certain technical conditions. With this distribution, the choice probabilities are of the form

$$P_i = \frac{e^{V_i} \cdot G_i(e^{V_1}, ..., e^{V_J})}{G(e^{V_1}, ..., e^{V_J})}$$

where $G_i$ is the $i$-th partial derivative of $G$. The expected maximum utility is

---

\(^{18}\) The variance-covariance matrix of these utility differences has $(J-1)^2$ elements and is symmetric. Hence there are only $J(J-1)/2$ identifiable elements of the original variance-covariance matrix, less one for utility-scale normalization (Chu, 1981, pp. 58-59; Bunch, 1991).
\[ E \max_j (V_j + \epsilon_j) = \log G(e^{V_1}, \ldots, e^{V_j}) + \gamma \] (2.27)

where again \( \gamma \) is Euler’s constant.\(^{19}\) Logit is the special case \( G(y_1, \ldots, y_J) = y_1 + \cdots + y_J \).

The best known GEV model, other than logit itself, is nested logit, also called structured logit or tree logit and first developed by Ben-Akiva (1974). McFadden (1981) discusses its theoretical roots and computational characteristics. 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:

\[ G(y_1, y_2, y_3) = y_1 + \left( y_2^{1/\rho} + y_3^{1/\rho} \right)^\rho. \] (2.28)

In this equation, \( \rho \) is a parameter between 0 and 1 that indicates the degree of dissimilarity between bus and rail; more precisely, \( 1-\rho^2 \) is the correlation between \( \epsilon_1 \) and \( \epsilon_2 \) (Daganzo and Kusnic, 1993). The choice probability for this example, computed from (2.26), may be written:

\[ P_i = P(B_{r(i)}) \cdot P(i \mid B_r) \] (2.29)

\[ P(B_r) = \frac{\exp(\rho \cdot I_r)}{\sum_{s=1}^2 \exp(\rho \cdot I_s)} \] (2.30)

\[ P(i \mid B_r) = \frac{\exp(V_i / \rho)}{\sum_{j \in B_r} \exp(V_j / \rho)} \] (2.31)

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 (2.31):

\[ I_r = \log \sum_{j \in B_r} \exp(V_j / \rho). \] (2.32)

\(^{19}\) This is demonstrated by Lindberg, Eriksson, and Mattsson (1995, p. 134). For a simple and elegant proof, see Choi and Moon (1997, p. 131).
When $\rho=1$ in this model, $\varepsilon_2$ and $\varepsilon_3$ are independent and we have the logit model. As $\rho \downarrow 0$, $\varepsilon_2$ and $\varepsilon_3$ 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 $\rho I_1 \to V_1$ and $\rho I_2 \to \max \{V_2, V_3\}$.

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 $\rho_r$ in equations (2.28)-(2.32), leading to the form:

$$G(y_1, \ldots, y_J) = \sum_r \left( \sum_{j \in B_r} y_j^{1/\rho_r} \right)^{\rho_r}. \tag{2.33}$$

This is the general two-level nested logit model. It has choice probabilities (2.29)-(2.32) except that the index $s$ in the denominator of (2.30) now runs from 1 to $R$. Like logit, it can be also derived from an entropy formulation (Brice, 1989). The groups $B_r$ can themselves be grouped, and those groupings further grouped, and so on, giving rise to even more general “tree structures” of three or more levels.

As in the logit model, the inclusive value is a summary measure of the overall desirability (expected maximum utility) of the relevant group of alternatives. This fact gives the “upper-level” probability (2.30) a natural interpretation as a choice among groups, taking the logit form with $I_r$ playing the role of the independent variable and $\rho_r$ its coefficient. As an example of an application of this interpretation, the inclusive value of a set of locations can serve as a measure of accessibility in models of destination choice (Ben-Akiva and Lerman, 1979). The welfare measure for the two-level nested logit model is, from (2.27), (2.32), and (2.33):

$$W = \frac{1}{\lambda} \log \sum_r \exp(\rho_r \cdot I_r) \tag{2.34}$$

where again $\lambda$ 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 (2.31) is a true conditional probability, and the model can be estimated sequentially: first estimate the parameters ($\beta', \rho$) from (2.31), use them to form the inclusive values (2.32), then estimate $\rho$ from (2.30). 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, nor does it produce consistent estimates of the standard
errors of the coefficients (Amemiya, 1978). Several studies show that maximum-likelihood estimation, although computationally more difficult, gives more accurate results (Hensher, 1986; Daly, 1987; Brownstone and Small, 1989).\(^{20}\)

There is some experience with other GEV models. Most of them generalize (2.33) by not requiring the subsets \(B_r\) to be mutually exclusive. Small (1987) defines subsets each encompassing two or more alternatives that lie close to each other on some ordering. Chu (1981) and Koppelman and Wen (2000) study models in which the subsets \(B_r\) include all possible pairs of alternatives:

\[
G(y_1, \ldots, y_J) = \sum_{j=1}^{J-1} \sum_{k=j+1}^{J} \left( y_j^{1/\rho_j} + y_k^{1/\rho_k} \right)^{\rho_j}.
\]

Such a model has the same number of estimable parameters in its variance-covariance matrix as multinomial probit (again, the arbitrary scale requires one more normalization), and so might have comparable generality. In practice this model seems to be easier to estimate than multinomial probit.

Nevertheless, estimation of GEV models is often difficult because of the highly nonlinear manner in which \(\rho\) enters the equation for choice probabilities. When the true model is GEV but differs only moderately from logit, a reasonable approximation can be estimated using two steps of a standard logit estimation routine, a procedure that appears to be considerably more stable than maximum likelihood estimation of the exact GEV model (Small, 1994).

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).

### 2.4.2 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

\(^{20}\) If maximizing the log-likelihood function is numerically difficult, one can start with the sequential estimator and carry out just one step of a Newton-Raphson algorithm toward maximization; this yields a statistically efficient estimate and seems to work well in practice (Brownstone and Small, 1989).
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), Mannering and Hensher (1987), and Washington et al. (2003, ch. 12).

The essence of the problem can be illustrated within an example of binary choice: that of owning a new or used automobile, denoted \( j = 1 \) or \( 2 \). Each type of car has a fixed measurable quality level \( Q_j \) that we can assume is higher for new cars, i.e. \( Q_1 > Q_2 \). For example, \( Q \) could be the number of safety features offered from a particular list, or simply an alternative-specific dummy variable equal to 1 for a new car. Let us suppose that the decision of how much to drive depends on car quality and income \( Y \), as follows:

\[
x = \beta_0 + \beta_Q Q + \beta_Y Y + u
\]  

(2.36)

where \( Y \) is income and \( u \) is a random error term. For simplicity we have omitted the subscript \( n \) denoting the individual in the sample. Car quality can be written in terms of the choice variable \( d_{1n} \) defined earlier, as follows (again omitting subscript \( n \)):

\[
Q = d_1 Q_1 + (1 - d_1) Q_2.
\]  

(2.37)

Substituting (2.37) into (2.36) makes explicit the dependence of the usage decision \( x \) on the ownership decision \( d_1 \).

Suppose also that the ownership decision depends on some set of observable variables \( X \), which could include \( Y \) and \( (Q_1 - Q_2) \):

\[
d_1 = 1 \text{ if } U_1 > U_2, \quad 0 \text{ otherwise};
\]

\[
U_1 - U_2 = \beta'_X X + \varepsilon.
\]  

(2.38)

This equation defines a binary probit model if \( \varepsilon \) is assumed normal, binary logit if \( \varepsilon \) is assumed logistic.

Selection bias is present if \( u \) and \( \varepsilon \) are correlated, which is likely because unobservable factors may affect both usage and the relative desirability of a new car. (An example of such a factor is how much this individual likes listening to a high-quality car stereo.) If \( u \) is correlated with \( \varepsilon \), it is also correlated with the car-type indicator \( d_1 \) and therefore with car quality \( Q \) via
This biases the estimated coefficients in (2.36), especially $\beta_Q$, because $Q$ is endogenous there.

If we can find an exogenous proxy for $Q$, we can use it instead and solve the problem. This can be accomplished using the following two-step procedure proposed by Heckman (1979).

**Step 1** consists of estimating a reduced-form version of (2.38). In this stage, the endogenous variable $x$ is replaced by the exogenous variables that are postulated to determine it: namely, $Q_1$, $Q_2$, and $Y$. Thus the ownership decision is modeled as some function of $X$, $Q_1$, $Q_2$, and $Y$. Unfortunately theory does not provide definitive guidance on the functional form for these variables, but usually some experimentation will produce a satisfactory fit. In this step, all that matters is that we obtain a reasonably good predictor of the probability $\hat{P}_1$ that the person will choose a new car. For convenience, let $Z$ represent all the variables (and transformations of them, if any) used in this reduced form, and $\varepsilon_R$ be the error term, so that the utility difference is

$$U_1 - U_2 = \beta_Z Z + \varepsilon_R.$$

From the estimated coefficient vector $\hat{\beta}_Z$, we can compute a predicted probability $\hat{P}_1$ of choosing a new car, equal to $\Phi(\hat{\beta}_Z Z)$ if the model is probit or $\left[1 + \exp(-\hat{\beta}_Z Z)\right]^{-1}$ if the model is logit.

**Step 2** consists of estimating a variant of (2.36) that is purged of endogeneity. There are two alternative strategies for doing this:

**Step 2 Version (a): Replace $Q$ by an exogenous predictor $\hat{Q}$.**

We look for an unbiased estimate of $Q$ that does not use the observed ownership choice, $d_1$, as does (2.37). There are at least three possibilities, which are Methods II, I, and III of Train (1986, p. 90):

(i) Compute $\hat{Q}$ as $E(Q) \equiv \hat{P}_1 \cdot Q_1 + (1 - \hat{P}_1) \cdot Q_2$.

(ii) Compute $\hat{Q}$ as the predicted value from an auxiliary regression of observed $Q$ on all the exogenous variables of the system, namely $Z$. (Note this method does not actually require that Step 1 be carried out.)

(iii) Compute $\hat{Q}$ from an auxiliary regression as in (ii) with $E(Q)$, calculated as in (i), as an additional variable in the regression. This procedure is more statistically efficient than
either (i) or (ii) because it incorporates data on actual choices (via the process for computing \( \hat{P}_1 \)) as well as on variables Z.

Method (iii) is probably the best choice in most cases, although like (ii) it requires one to specify arbitrarily the exact functional form of the auxiliary regression.

**Step 2 Version (b): Add a “correction term” to the error term in (2.36) to make it independent of u.**

One way to look at selection bias is that observed \( Q \) is conditional on the individual’s ownership decision. Therefore using \( Q \) as a variable in (2.36) would be appropriate if (2.36) could be transformed into an equation describing usage *conditional on* ownership. This can be done by making its error term conditional on ownership. If \( u \) is assumed to be normal, as is usual, the required transformation is accomplished by subtracting the conditional expectation of a normal variable, given its link to ownership via (2.38), from \( u \); the remaining error term is can be assumed independent of \( Q \) and so (2.36) is purged of selectivity bias. A recent example of use of this technique is West’s (2004) model of automobile type choice and amount of use.

That conditional expectation can be computed explicitly for binary probit and logit models.\(^21\) We write the new term to be added to (2.36) as \( \gamma C \) where \( \gamma \) is a parameter to be estimated and \( C \) is a “correction variable” computed from the results of Step 1. The estimated value of \( \gamma \) will give us information about the correlation between \( u \) and \( \varepsilon \), which we denote by \( \rho \).

It is this correlation that causes the problem, so we can test for selection bias by testing whether \( \gamma \) is different from zero.

Table 2.1 gives formulas for the correction variable \( C = d_1 C_1 + (1-d_1) C_2 \); it also shows how parameter \( \gamma \) is related to correlation \( \rho \). In this table, \( \Phi \) denotes the probability distribution function of the standard normal distribution, and \( \phi \) its derivative (i.e. the normal density function); \( \sigma_u \) is the standard deviation of \( u \); and \( \hat{P}_2 = 1 - \hat{P}_1 \). Sometimes data are lacking on people

---

\(^21\) Supposedly it can be done for multinomial logit model as well, but it is extremely complex. Dubin and McFadden (1984) specify a usage model conditional on a single choice, \( i \); the include \( J-1 \) correction terms, and so estimate \( J-1 \) correlations (between \( u \) and \( \varepsilon, j \neq i \)). However they do not discuss how to pool the data with observations of individuals who choose other alternatives.
making choice \( j = 2 \), in which case the correction factor is simply \( C_1 \) and the usage equation is estimated on just the subsample of new car owners.\(^{22}\)

<table>
<thead>
<tr>
<th>Table 2.1. Selectivity Correction Terms ( \gamma(D^1C_1 + D^2C_2) )</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Model</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Probit</td>
</tr>
<tr>
<td>Logit</td>
</tr>
</tbody>
</table>

What this procedure does is add correction \( \gamma C_1 \) for those individuals in the sample who chose a new car, and \( \gamma C_2 \) for the others. Note that in each row, \( C_1 \) is positive and \( C_2 \) is negative. Thus if \( \rho \) is positive, indicating that people choosing new cars are likely to drive more for unobserved reasons, the adjustment indicates that \( \epsilon \) has a positive expected value for those individuals who choose new cars, and a negative expected value for those who choose used cars. The extent of the adjustment is determined by estimated coefficient \( \gamma \), which is positively related to \( \rho \) as shown in the last column.

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 decisions. Golob (2003) provides a review.

### 2.4.3 Disaggregate Panel Data

\(^{22}\) Sign conventions vary in the literature. In the probit case, some references replace \( \hat{\beta}_Z Z \) by the equivalent quantity \( \Phi^{-1}(\hat{P}_1) \), where \( \Phi^{-1} \) denotes the inverse of the standard normal cumulative distribution function.
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.

2.4.4 Random Parameters and Mixed Logit

In the random utility model of (2.4)-(2.5), 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 (2.13) 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 (2.4). 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. Boyd and Mellman (1980) and Cardell and Dunbar (1980) showed how one could allow a parameter in the logit model to vary randomly across individuals. The idea is to specify a distribution, such as normal with unknown mean and variance, for the parameter in question; the overall probability is determined by embedding the integral in (2.5) within another integral over the density function of that distribution. Subsequently, this simple idea was generalized to allow for general forms of randomness in all the parameters – even alternative-specific constants, where further randomness might seem redundant yet it proves a simple way to
produce correlation patterns like those in GEV without the complexity of the GEV probability formulas. Such models are 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 $\varepsilon_{jn}$) is unnecessary because, conditional on the values of random parameters, it yields the logit formula (2.8). 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.

The mixed logit model is simple to write out. Using the logit formulation of (2.8) and (2.10), the choice probability conditional on random parameters is

$$P_{in|\beta} = \frac{\exp(\beta^T z_{in})}{\sum_j \exp(\beta^T z_{jn})}.$$  

(2.39)

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

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

(2.40)

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

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

Doing so requires, of course, assuming some trial value of $\Theta$, just as calculating the usual logit probability requires assuming some trial value of $\beta$. Under reasonable conditions, maximizing the likelihood function defined by this simulated probability yields statistically consistent estimates of the meta-parameters $\Theta$. 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.\footnote{The following simplified explanation is adapted from Small and Winston (1999).} 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:

\[ V = -0.264 \cdot \frac{p}{\ln(inc)} + 0.517 \cdot \text{range} + (1.43 + 7.45 \phi_1) \cdot \text{size} + (1.70 + 5.99 \phi_2) \cdot \text{luggage} + 2.46 \phi_3 \cdot \text{nonE} + 1.07 \phi_4 \cdot \text{nonN} + \text{(other terms)} \]

where \( p \) (vehicle price) and \( inc \) (income) are in thousands of dollars; the \text{range} between refueling (or recharging) is in hundreds of miles; \text{luggage} is luggage space relative to a comparably sized gasoline vehicle; \text{nonE} is a dummy variable for cars running on a fuel that must be purchased outside the home (in contrast to electric cars); \text{nonN} is a dummy for cars running on a fuel stored at atmospheric pressure (in contrast to natural gas); and \( \phi_1 - \phi_4 \) 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 \text{nonE} and \text{nonN}. This can be understood by examining the results term by term.

The terms in parentheses involving \( \phi_1 \) and \( \phi_2 \) 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 actually prefer less luggage space namely, those for whom \( \phi_2 < -1.70/5.99 \); presumably they do so because a smaller luggage compartment allows more interior room for the same size of vehicle. Similarly, preference for vehicle size ranges from negative (perhaps due to easier parking for small cars) to substantially positive.

The terms involving \( \phi_3 \) and \( \phi_4 \) 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 $\phi_3$ and $\phi_4$ 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 $\phi_1$ and $\phi_2$. 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; this is in addition to other random utility components including $\varepsilon_{i\theta}$ in (2.4), which as we have seen has standard deviation equal to $\pi/\sqrt{6}=1.28$ by normalization – a requirement for (2.39) to be valid. Thus the combined additive random term in utility, \(^\text{24}\)

\[ (\varepsilon_{i\theta}+2.46 \phi_3 \cdot \text{nonE}_i+1.07 \phi_4 \cdot \text{nonN}_i), \]

exhibits correlation across those alternatives $i$ representing cars that are not electric and, by similar argument involving $\phi_4$, 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 $\phi_3$ and $\phi_4$ 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. The model described above acts much like a GEV model with overlapping nests for alternatives in groups nonE and nonN, and with random parameters for size and luggage. It is probably easier to estimate than such a nested logit model, especially if one is already committed to random parameters. Even more complicated error structures can be accommodated within this framework, for example one designating repeated observations from a given individual (Small,

\[^{24}\text{As Brownstone and Train point out, the terms in } \phi_3 \text{ and } \phi_4 \text{ may also be viewed as part of an additive random utility term, but one that is not a constant, i.e. it depends on values of observed variables.}\]
Winston, and Yan, 2005) or spatial correlation related to geographical location (Bhat and Guo, 2004).

In principle, the mixing idea can be applied to any choice model, not just logit, in order to randomize its parameters. Indeed, it happens that the multinomial probit model was first developed with a random-parameters formulation (Hauman and Wise, 1978), a fact that has caused some confusion about the relationship between probit and logit. There may be cases where it is easier to estimate a random-parameters multinomial probit than a mixed logit model, but usually it is harder because one needs to simulate not only the explicit integral in (2.40) but also the integral that, for probit, is part of the definition the conditional choice probability \( P_{in|β} \).

2.5 Activity Patterns

A more fundamental approach to the demand for travel would be to explain the entire structure of decision-making about what activities to undertake in what locations. This idea has proven difficult to translate into workable models that use available data, but important progress has been made. For example, some surveys now elicit multi-day diaries describing all activities and travel undertaken during a period of time. Descriptive statistics on activity patterns have been compiled showing surprising similarities across nations including US, UK, Japan, Canada, and The Netherlands (Timmermans et al., 2002).

Ettema and Timermans (1997), Ben-Akiva and Bowman (1998), and Bhat and Koppelman (1999) provide reviews of activity-based models. There are two main classes. Econometric models extend the basic framework of this chapter to deal with additional choice dimensions such as trip frequency, destination, and type and duration of activities undertaken. Simulation models may also utilize a utility-maximization choice framework, but they emphasize more the enumeration of feasible activities based on various constraints such as that relating the starting and ending times of each trip to its duration.

One significant advance has been to model an entire tour (a round trip visiting one or more destinations in sequence) as an object of choice. The problem is that this quickly leads to enormous numbers of possible alternatives, especially when one considers a daily schedule containing several possible tours. Bowman and Ben-Akiva (2001) improve tractability by breaking the overall decision about the daily schedule into parts, including a primary tour type,
secondary tour type(s), and destinations and modes of travel for each tour. Such a model lends itself to a structured choice model such as nested logit. Illustrating the difficulty of designing realistic models, the authors acknowledge that the results in their example are able to explain only a small part of variations in observed activity patterns.

Few if any formal models have been able to account for flexibility in both the times of day and the locations at which activities take place, both of which are fundamental to describing the trips connecting them. Furthermore, to fully understand the processes generating travel, one needs to model the substitution between in-home and out-of-home activities, which adds further to the sheer number of possibilities to consider.

As an example of what can be accomplished with such models, Shiftan and Suhrbier (2002) utilize one of the best data sets for activity analysis – a 1994 household survey in Portland, Oregon – to analyze several policies classified as “travel demand management.” One result is illustrative. A policy to encourage telecommuting is predicted to reduce long-distance work trips to downtown Portland. But it increases the number of short tours as people make special-purpose trips for activities that previously were handled as part of a tour from home to work and back. Other studies of telecommuting have found only a very small net reduction in travel (Choo, Mokhtarian, and Salomon, 2005). More generally, many types of telecommunication appear to be complements to, rather than substitutes for, travel (Plaut, 1997).

2.6 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 (2.17), so long as they are part of the demand model. However, it is often enlightening to separate them explicitly.

2.6.1 Theory of 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 despite confusion on this subject.25

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 second approach is pursued here.

A useful theoretical framework builds on that of Becker (1965), in which utility is maximized subject to a time constraint. Becker's theory has been elaborated in many directions; here, we present ideas developed mainly by Oort (1969) and DeSerpa (1971), adapting the exposition of MVA Consultancy et al. (1987).

Let utility $U$ depend on consumption of goods $G$, time $T_w$ spent at work, and times $T_k$ spent in various other activities $k$. We can normalize the price of consumption to one. Utility is

---

25 It is sometimes claimed that the average value of time savings diminishes rapidly as the time savings shrink to zero, which would imply a very low marginal rate. But these claims are based on ad hoc empirical specifications, and ignore the fact that travel patterns are in constant flux so the time saving from one particular policy cannot long be distinguished from other sources of differences in trip times. Studies based on consistent definitions have not found such dependencies (MVA Consultancy et al., 1987, pp. 65-68), and theory refutes the alleged rationale for them (Mackie, Jara-Diaz, and Fowkes, 2001).
maximized subject to several constraints. First, there is the usual budget constraint involving unearned income $Y$ and earned income $wT_w$, where $w$ is the wage rate. Second, a time constraint requires that time spent on all activities equal total time available, $\bar{T}$. Finally, the nature of certain activities (such as travel) imposes a minimum $\bar{T}_k$ on time $T_k$ spent in activity $k$. (We will consider as an extension the possibility that $T_w$ is also constrained.)

This problem can be solved by maximizing the following Lagrangian function with respect to $G$, $T_w$, and $\{T_k\}$:

$$L = U(G,T_w,\{T_k\}) + \lambda \left[ Y + wT_w - G \right] + \mu \left[ \bar{T} - T_w - \sum_k T_k \right] + \sum_k \phi_k \left[ T_k - \bar{T}_k \right],$$

(2.41)

where $\lambda$, $\mu$, and $\{\phi_k\}$ are Lagrangian multipliers that indicate how tightly each of the corresponding constraints limits utility. The first-order condition for maximizing (2.41) with respect to one activity time $T_k$ is

$$U_{T_k} - \mu + \phi_k = 0$$

(2.42)

while that with respect to $T_w$ is

$$U_{T_w} + \lambda w + \lambda T_w \cdot (dw/dT_w) - \mu = 0,$$

(2.43)

where subscripts on $U$ indicate partial derivatives. We have allowed for a nonlinear compensation schedule by letting $w$ depend on $T_w$.

We can denote the value of utility at the solution to this maximization problem by $V$, the indirect utility function; it depends on $Y$, $\bar{T}$, wage schedule $w(T_w)$, and minimum activities times $\{\bar{T}_k\}$. The rate at which utility increases as the $k$-th minimum-time constraint is relaxed is given by its Lagrange multiplier, $\phi_k$; the increase with respect to unearned income is $\lambda$. Hence the marginal value of time for the $k$-th time component is their ratio:

$$v^k = \frac{\partial Y}{\partial \bar{T}_k} = \frac{\phi_k}{\lambda}. \quad (2.44)$$

Those activities for which the minimum-time constraint is not binding, i.e. those for which $\phi_k=0$, are called by DeSerpa pure leisure activities. The others, which presumably include most travel, are intermediate activities.

Equations (2.42)-(2.44) imply:

$$v^k_T = \frac{\mu - U_{T_k}}{\lambda} = w + T_w \cdot \frac{dw}{dT_w} + \frac{U_{T_w}}{\lambda} - \frac{U_{T_k}}{\lambda}. \quad (2.45)$$
This equation decomposes the value of travel-time savings into the opportunity cost of time that could be used for work, \( \mu/\lambda \), less the value of the marginal utility of time spent in travel. The opportunity cost is both pecuniary (the first two terms after the last equality) and nonpecuniary (the third term, which could be positive or negative).

Most of the theoretical literature assumes that the wage is fixed, in which case equation (2.45) gives the result noted by Oort (1969): the value of time exceeds the wage rate if time spent at work is enjoyed relative to that spent traveling, and falls short of it if time at work is relatively disliked. This is a fundamental insight into how the value of time, even for non-work trips, depends on conditions of the job. It suggests a modeling strategy that interacts variables believed to be related to compensation and work enjoyment with those measuring time or cost. In addition, we might expect \( \nu^k_T \) to rise with total trip time because the total time constraint in (2.41) will bind more tightly, causing \( \mu \), the marginal utility of leisure, to rise.

### 2.6.2 Empirical Specifications

The most common situation for measuring values of time empirically is one where a discrete choice is being made, such as among modes or between routes. To clarify how the general theory just presented corresponds to empirical specifications, assume that there is only one pure leisure activity, \( k=0 \), and that the other activities are all mutually exclusive travel activities, each consisting of one trip. We can also add travel cost \( c_k \delta_k \) to the budget constraint, where \( \delta_k \) is one if activity \( k \) is chosen and zero otherwise. The indirect utility function has the same derivatives with respect to exogenous variables \( c_k \) and \( T_k \) as does the Lagrangian function:

\[
\frac{\partial V}{\partial c_k} = -\lambda \delta_k ; \quad \frac{\partial V}{\partial T_k} = -\phi_k \delta_k .
\]

Equivalently, the conditional indirect utility functions needed for a discrete-choice model satisfy:

\[
\frac{\partial V_k}{\partial c_k} = -\lambda ; \quad \frac{\partial V_k}{\partial T_k} = -\phi_k ,
\]

which imply that our definition of value of time in (2.44) is identical to that in (2.13):

\[
v^k_T \equiv \frac{\phi_k}{\lambda} = \frac{\partial V_k / \partial T_k}{\partial V_k / \partial c_k} .
\]
Note also that the first of equations (2.46) is identical to (2.18) since here there is assumed just one trip per time period. (It is easy to generalize this model to allow for an endogenously chosen number of trips per time period.)

Our theory provides some 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) and that its relative marginal disutility is a fixed fraction of the wage rate. Suppose further that the wage rate is fixed, so the second term in (2.45) disappears. Then (2.45) implies that the value of time is a fraction of the wage rate, as for example with specification (2.9) with $\beta_3=0$. 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 (2.45) implies that value of time is a nonlinear function of the wage rate, which could suggest using (2.9) with a non-zero term $\beta_3$ or with additional terms involving cost divided by some other power of the wage.

Train and McFadden (1978) demonstrate how specific forms of the utility function in (2.41) can lead to operational specifications for the conditional indirect utility function with a desired relationship between value of time and wage rate. For example, in (2.9), we could have achieved the same relationship by multiplying the time-related variables by wage rather than by dividing the cost by wage; but doing so would imply a different underlying function $U(\cdot)$.

### 2.6.3 Extensions

Several extensions to the theory just presented are interesting. First, consider work-hour constraints. People are not always free to change the amount of time they spend at work, perhaps because they are locked into a particular job with fixed hours or because there are few jobs offered with the work hours that they prefer. To some extent this is handled by allowing $w$ to depend on $T_w$. We could represent a stricter constraint by adding a term $\phi_w \cdot [T_w - T_w^*]$ to (2.41), where $\phi_w$ is another Lagrangian multiplier whose sign indicates whether this person would prefer fewer ($\phi_w > 0$) or more ($\phi_w < 0$) hours at the job. This modification adds a term $\phi_w / \lambda$ to the value of time as given by (2.45). Possibly a positive value is suggested by the finding of MVA Consultancy et al. (1987, pp. 149-150) that people who are required to work extra hours at short notice have 15-20 percent higher values of travel time than other workers.
Another extension is suggested by Jara-Díaz (2000, 2003). Suppose goods consumption requires using an amount of leisure time proportional to those goods. This constraint is represented by modifying the term $\phi_0 \cdot [T_0 - T_0]$ in the last summation in (2.41), where subscript 0 indicates the leisure activity; it now becomes $\phi_0 \cdot [T_0 - \ell \cdot G]$, where $\ell$ is the unit time requirement for consumption. The constraint is binding if $\phi_0 > 0$. (This modification involves a modification of the definition of leisure, which previously was defined, following DeSerpa, by the condition $\phi_0 = 0$.) While this modification does not alter the formulas derived for value of time, it does change the meaning of $\lambda$ (the marginal utility of income) in those formulas. Previously, $\lambda = U_G$, as is easily seen by writing the first-order condition for maximizing (2.41) with respect to $G$. But this modified leisure constraint introduces a new term in that first-order condition, resulting in $\lambda = U_G - \phi_0 \cdot \ell$. Since $\lambda$ appears in the denominator of expressions for value of time, this change would tend to raise the value of time if the constraint is binding. This may be viewed as yet another model producing the “harried leisure class” postulated by Linder (1970) as a byproduct of rising productivities.

A different extension is considered by De Borger and Van Dender (2003). Suppose the time required for travel depends on the amount of time worked, so that $\bar{T}_k = t_k \cdot T_w$ for fixed parameter $t_k$. This might happen for a number of reasons: secondary workers entering or leaving the work force, part-time workers changing the number of days worked, or part-time workers having to accept more distant jobs in order to increase time worked. In that case, (2.43) acquires the additional term $-\phi t_k$, and all terms on the right-hand side of (2.45) are divided by $(1 + t_k)$. Thus greater commuting time decreases the value of time – opposite to the effect noted earlier from a rising marginal utility of leisure – because it reduces the hourly wage rate net of commuting cost. De Borger and Van Dender suggest in numerical simulations that the effect can be quite large and causes unexpected results: for example, reducing congestion can cause the value of time to rise so much that total travel cost actually increases.

Other theoretical extensions show that value of time can depend on tax rates (Forsyth, 1980) and on scheduling considerations (Small, 1982).

2.6.4 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 (e.g., MVA Consultancy et al., 1987, pp. 61-62). This conclusion is supported by attitudinal surveys (Prashker, 1979), and perhaps by the frequent finding that time spent in congestion is more onerous than other in-vehicle time. How can this aversion to unreliability be captured in a theoretical model of travel?

One approach, adapting Noland and Small (1995), is to begin with the model of tripscheduling choice presented in equation (2.24). Dividing utility by minus the marginal utility of income, we can write this model in terms of trip cost, in a conventional notation that we will use extensively in the next chapter:

\[
C(t_d, T_r) = \alpha \cdot T + \beta \cdot SDE + \gamma \cdot SDL + \theta \cdot DL
\]

where \(\alpha \equiv v_T/60\) is the per-minute value of travel time, \(\beta\) and \(\gamma\) are per-minute costs of early and late arrival, and \(\theta\) 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 \(C^\ast\) denote this expected cost after the user chooses \(t_d\) optimally, we have

\[
C^\ast = \min_{t_d} E[C(t_d, t_r)] = \min_{t_d} \left[ \alpha \cdot E(T) + \beta \cdot E(SDE) + \gamma \cdot E(SDL) + \theta \cdot P_L \right]
\]

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 (2.25). It captures the effect of travel time uncertainty upon expected schedule delay costs, but may omit other reasons why uncertainty could cause disutility.

To focus just on reliability, let’s ignore the dynamics of 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 this probability density function, \(T_r\) the travel time when \(T_r = 0\), and \(t^\ast\) the desired arrival time at the destination. The next to last term in the square brackets of (2.48) can then be written as

\[P_L(T_r) = P_L(T_r < t^\ast)
\]

26 See for example MVA Consultancy et al. (1987), p. 149; Small, Noland, Chu, and Lewis (1999); and Hensher (2001).
\[
\gamma \cdot E(SDL) = \gamma \cdot E(t_d + T_r - \tilde{t} \mid T_r > \tilde{t} - t_d ) = \gamma \cdot \int_{\tilde{t} - t_d}^{\infty} (t_d + T_r - \tilde{t}) \cdot f(T_r) dT_r
\]

where \( \tilde{t} \equiv t^* - T_r \) is the time the traveler would depart if \( T_r \) were equal to zero with certainty.

Differentiating yields:

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

where \( P^*_L \) is the optimal value of the probability of being late.\(^{27}\) Similarly, differentiating the term involving \( \beta \) in (2.48) yields \(-\beta \cdot (1 - P^*_L)\). Finally, differentiating the last term yields \(-\theta \).

\[
\text{where } f(\tilde{t} - t^*_d) \text{ is the probability density at the point where the traveler is neither early nor late. Combining all three terms and setting them equal to zero gives the first-order condition for optimal departure time:}
\]

\[
P^*_L = \frac{\beta + \theta f^0}{\beta + \gamma}.
\]

(2.49)

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 \( \theta = 0 \), 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 \( b \), (2.49) again simplifies to a closed form:

\[
P^*_L = \frac{\beta + (\theta / b)}{\beta + \gamma}.
\]

---

\(^{27}\) The term “0” in this equation arises from differentiating the lower limit of integration:

\[
- \left[ \frac{d}{dt_d} (\tilde{t} - t_d) / dt_d \right] \left[ (t_d + T_r - \tilde{t}) \cdot f(T_r) \right]_{\tilde{t} - t_d = 1} = 1 - 0 = 0.
\]
The value of $C^*$ in this case is given by Noland and Small (1995) and Bates et al. (2001). In the special case $\theta=0$, it is equal to the cost of expected travel time, $\alpha \cdot E(T)$, plus the following cost of unreliability:

$$v_R = \left( \frac{\beta \gamma}{\beta + \gamma} \right) \frac{b}{2}. \tag{2.50}$$

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 (2.50) indicates that reliability cost derives from the combination of costly scheduling mismatches and dispersion in travel time.

More generally, the last two terms in (2.48) are potentially important if $\gamma > \beta$ or if $\theta$ is large, conditions that are in fact true according to the empirical findings in (2.24). These terms are sensitive to the values of $E(SDL)$ and $P_L$, which depend especially on the shape of the distribution of $T_r$ in its upper ranges, since this determines the likelihood that $T_r$ takes a high enough value to make the traveler late. Thus we might expect the expected cost of unreliability to depend more on this part of the distribution (its “upper tail”) than on other parts.

Equation (2.50) applies equally to the expected cost of schedule mismatches on a transit trip, under the common assumption that people arrive at a transit stop at a steady rate, if $b$ is reinterpreted as the headway between transit vehicles. Although under that interpretation $v_R$ is proportional to expected waiting time, it is not a representation of waiting-time cost but rather must be added to it. In the case where the transit headway is itself uncertain, or where the vehicle might be too full to accommodate another passenger, the derivation of reliability cost for transit becomes much more complicated (Bates et al., 2001).

2.6.5 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.

---

28 This is pointed out by Wardman (2004, p. 364), who attributes the point to unpublished work by John Bates.
Waters (1996) reviews 56 value-of-time estimates from 14 different nations. Each is stated as a fraction of the gross wage rate. Focusing on those where the context is commuting by automobile, he finds an average ratio of VOT to wage rate of 48 percent, and a median ratio of 42 percent. He suggests that “a representative [VOT] for auto commuting would be in the 35 to 50 percent range, probably at the upper end of this range for North America.” Consistent with this last statement, both Transport Canada (1994, sect. 7.3.2) and US Department of Transportation (1997) currently recommend using a ratio of 50 percent for personal travel by automobile.

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. Mackie et al. (2003), reviewing a larger set of UK studies, recommend best hourly values for VOT of £3.96 for commuting and £3.54 for other trips at 1997 prices; their average is 51% of the relevant wage rate. Gunn (2001) find that Dutch values used in 1988, differentiated by level of household income, track well various British results for a similar time. However, Gunn reports that there was a substantial unexplained downward shift in the profile for 1997 – a phenomenon possibly resulting from better amenities in vehicle. Another Dutch study – using a novel methodology in which the “choice” is job termination rather than mode or route – finds a ratio of VOT to wage rate of one-third for shorter commutes (less than one hour round trip) and two-thirds for longer ones, for an average of “almost half” (Van Ommeren, Van den Berg, and Gorter, 2000). A French review by the Commissariat Général du Plan (2001, p. 42) finds VOT to be 77 and 42 percent of the wage for commuting and other urban trips, respectively, for an average of 59 percent. Finally, a Japanese review suggests using 2,333 yen/hour for weekday automobile travel in 1999, which was 84 percent of the wage rate.

There is considerable evidence that value of time rises with income but less than proportionally, which makes the expression of VOT as a fraction of the wage rate, as above,

---

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

30 Mean gross hourly earnings in 1997 were £7.42/hour, from same source as previous footnote.

31 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).
somewhat less attractive. An earlier set of studies in England found that although members of the
highest-income group had incomes more than three times those of the lowest group, their values
of time were only 30 to 40 percent higher.\footnote{MVA Consultancy et al., pp. 133-135, 150, 152.} Similarly, Mackie \textit{et al.} (2003) find that values of
time for the highest of three broad income groups are 1.5 to 2.4 times those for the lowest group.
The easiest way to summarize this issue is in an elasticity of value of time with respect to
income. Wardman (2001, p. 116), using a formal meta-analysis, finds that elasticity to be 0.51
when income is measured as gross domestic product per capita; with a larger sample he obtains
0.72 (Wardman, 2004, p. 373), and he is part of a group that recommends using an elasticity of
0.8 (Mackie \textit{et al.}, 2003). These elasticities could be subject to a downward bias if there is
indeed a downward trend, independent of income, as suggested by Gunn.

Wardman’s (2001) meta-analysis is especially useful for tracking the effects of various
trip attributes on value of time. For example, there is a 16 percent differential between value of
time for 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. Wardman actually
gets a considerably smaller differential, namely a ratio of 1.62, which is quite precisely
estimated; nevertheless Mackie \textit{et al.} (2003) recommend using a ratio of 2.0. There is
considerable dispersion in the reported estimates of these relative valuations, especially in the
relative value of waiting time (MVA Consultancy et al., p. 130). This may indicate that the
disutility of transfers (which entail waiting as well as other possible difficulties) is quite variable,
and suggests a payoff from research into the sources of this variation.

A number of studies have been carried out using Chilean data. Munizaga \textit{et al.} (2004),
using an innovative model that combines choices of activities and travel modes by residents of
Santiago, obtain average VOT equal to 46\% and 67\% of the wage rate for middle and upper
income groups, respectively.

SP data often yield considerably smaller values of time than RP data. For example,
Hensher (1997) and Calfee and Wintson (1998) obtain values using SP surveys of car commuters
of 19 percent and 20 percent, respectively, of the wage rate. Brownstone and Small (2005) take advantage of three data sets, all from “high occupancy toll lane” facilities in southern California, that obtained RP and SP data from comparable populations, in some cases from the same individuals. They find that SP results for VOT are one-third to one-half the corresponding RP results, the latter being 50-90 percent of the wage rate. One possible explanation for this difference is hinted at by the finding, from other studies of these same corridors, 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. (This is not to prejudge whether it may be desirable to use a constant absolute amount in cost-benefit analysis for political or distributional reasons, a subject we consider in chapter 5.) There is universal agreement that value of time is much higher for travel while on business, generally taken as 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.

Several studies have applied mixed logit to measure variation from unobserved sources in the disutility of time and reliability. Hensher (2001) allows for random coefficients of three types of travel time, using SP data on New Zealand commuters, resulting in standard deviations of

---

33 This statement is based on Calfee and Winston’s summary of the average over the entire sample (p. 91), and on Hensher’s Table 3.7 (p. 274), panel for “private commute,” using his preferred VOT of $4.35/hour.

34 See their Table 1, rows 4-5, 13-14
VOT equal to 41-58 percent of the corresponding mean VOT. The California studies reviewed by Brownstone and Small (2005) measure heterogeneity as the inter-quartile range (75th minus 25th percentile values) of the distribution of VOT or VOR with that measure, unobserved heterogeneity in VOT (i.e., that due just to random coefficients) is 55-125 percent of median VOT in two cases using RP data, and 144 percent in one case using SP data.

There has been far less empirical research on value of reliability. Almost all of it has been based on SP data, for at least two reasons: it is difficult to measure unreliability in actual situations, and unreliability tends to be correlated with travel time itself. However, a few recent studies have had some success with RP data. One key development is to measure unreliability as a property of the upper percentiles of the distribution of travel times, as suggested by the theory discussed earlier. It turns out that such a measure is less correlated with travel time than is a symmetric measure like standard deviation, because the upper-percentile travel times (i.e., travel times that occur only rarely) tend to arise from incidents such as accidents or stalled vehicles. The occurrence of such incidents is closely correlated to congestion, but the delays they cause are less so because the effects of the incident persist long after it occurs.

Bates et al. (2001) review several SP studies of car travel that define unreliability as the standard deviation of travel time. Those that they deem most free of methodological problems produce a value of reliability (VOR), expressed in units of money per unit increase in that standard deviation, on the order of 0.8 to 1.3 times the value of time (VOT). Brownstone and Small (2005) review 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 also, 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.

If reliability is not controlled for in studies of value of time, the estimated VOT may include some aversion to unreliability to the extent that time and unreliability are correlated. Nevertheless, the studies reviewed by Brownstone and Small (2005) obtain high VOT for

35 This statement is based on Hensher’s Table 3, Model 3a, the lower panel showing values of time in which the cost coefficient is that on a variable measuring the toll.

36 An updated version of that study is Small, Winston, and Yan (2005).
automobile users even when simultaneously measuring VOR. In the case of the value of waiting time for public transit, the bias is measured explicitly by (2.50) under the interpretation stated earlier, in which (2.50) is proportional to expected waiting time $b/2$.

Turning to freight transportation, it is clear that values of time and reliability are important, but empirical evidence is sparse and definitions inconsistent. Most studies use SP methodology and many involve mostly involving inter-city travel. De Jong (2000) provides a recent review of studies, which suggests that for countries like The Netherlands, where a high proportion of travel is urban, values of time are quite high, consistent with conventional theory that travel time for road-freight vehicles it is viewed similar to business time and includes some inventory value for equipment and payload. Kawamura (2000), Wigan et al. (2000), and Fowkes et al. (2004) provide some evidence on values of both travel time and reliability.

### 2.7 Conclusions

All tractable approaches to travel-demand analysis are based upon greatly simplified portrayals of travel behavior. This is necessary because the purposes of travel and the variety of choices available make travel choices so complex. As a result, distinct or even mutually contradictory analytical approaches may each provide useful information for particular circumstances, and the sophisticated planner will want to understand a wide variety of approaches.

Both aggregate and disaggregate models can be instructive, the choice between them depending in any given instance on availability of micro data and on how important it is to have an explicit representation of individual decision-making processes. Many of the problems plaguing the traditional planning process are not inherent in aggregate models, but rather in simplifications that obscure important feedback effects. Disaggregate models, even if they have not always improved forecasting accuracy, have performed well in many circumstances and have enabled researchers to undertake new and sophisticated types of policy analysis. They have also enriched our understanding of how variability affects travel behavior, and they have given new insight into aggregate measures of attractiveness, accessibility, and welfare.

The theory of time allocation is well developed and permits us to rigorously address conceptual issues concerning value of time and reliability. Despite uncertainty, a consensus has developed over many of the most important empirical magnitudes for values of time, permitting
them to be used confidently in benefit assessment. Another decade should bring similar consensus to value of reliability.

Figure 2.1 Disutility of Schedule Delay