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, the time of day at which trips are taken, parking locations, and the duration of activities for which travel is undertaken.

These multiple decisions are often envisioned as a sequence, typically starting with residential and job locations, then vehicle ownership, then other aspects. This sequence is in decreasing order of the time span over which the decision can be changed easily. However, it does not imply a sequential decision procedure whereby one decision is made without regard to its implications for later decisions. Rather, each decision is affected by others and so can be fully understood only as part of a simultaneous choice process. A given study may isolate just a few of these decisions for tractability; it is then all the more important to remember, in interpreting results, that other decisions are lurking in the background.

Furthermore, travel is a derived demand, usually undertaken not for its own sake but rather to facilitate a 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 also calls attention to an increasingly common form of travel: the linking together of several trip purposes into one integrated itinerary or *tour*, a process known as *trip chaining*.

The chapter begins (Section 2.1) by asking what can be learned from aggregate data about travel and how conventional economic demand theory is applied to such data. It then

moves on to disaggregate models (Section 2.2), also known as "behavioral" because they depict individual decision-making explicitly. Section 2.3 presents examples of models explaining some key travel choices: mode, time of day, and route. More specialized topics are then discussed (Sections 2.4-2.5). Finally, Section 2.6 analyzes two quantities of special interest to policy: travelers' willingness to pay for travel-time savings and for improved reliability. We mostly discuss passenger transportation, in part because more data are available than for freight; however studies of the demand for urban freight transportation tend to use similar methods.[1]

2.1 Aggregate Tabulations and Models

Much can be learned simply from cross-tabulating survey data. For the United States, there are two very useful sources. One, covering all trips, is the National Personal Transportation Survey (NPTS), collected at approximately six-year intervals; for the single year 2001 it was subsumed into a broader survey called the National Household Travel Survey (NHTS). The other useful source is the journey-to-work portion of the US Census, taken every ten years. This consists of responses to questions about work travel, asked on the Census "long form," which is administered to about 17% of all households. Pisarski (2006) provides a comprehensive analysis of the 2000 journey-to-work census and compares it to earlier censuses and to the 2001 NHTS.

The NHTS shows that work trips in the US continue to decline as a proportion of all trips, amounting to about 16% in 2001.[2] However, it would be a mistake to conclude that work trips are no longer important for urban transportation. Two-thirds of US work trips in 2000 occurred during the hours 6-9 a.m., and work trips account for about half of all person-miles of travel during those hours. Trips to work have been found to account for more than half of all trips in selected Belgian cities, and a very high fraction of peak-period trips on some of the most

[1] On freight demand, see D'Este (2000) and Small and Winston (1999). Collections of more specialized papers on travel demand are found in Hensher and Button (2000) and Mahmassani (2002).

[2] See Pisarski (2006), p. 3. He adjusts the 2001 survey tabulation upward to better match the way "trips" were defined in previous years; without that adjustment the figure is 15 percent. The reason for the adjustment is that the 2001 survey, unlike earlier surveys, counts each person, even a very small child, as a separate trip-maker. In the NHTS, "trips" includes such casual activities as walking the dog, which is probably why it shows a surprisingly high proportion of trips made by walking or bicycling: 9.5 percent according to Pucher and Renne (2003), compared to just 3.4 percent according to the journey-to-work Census (calculated from Pisarski 2006, Table ES-2, by dividing the percentages shown by 0.9674 to eliminate those working at home from their denominators).

notoriously congested Los Angeles freeways.[3] Thus work trips remain predominant contributors to congestion.

Private vehicles dominate all categories of trips in the US, and increasingly so. They account for 91% of all work trips in 2000, up from 89% in 1990. The share of work trips that use public transit is just 4.7%, down from 5.3% in 1990. However, transit share is much larger in large metropolitan areas, averaging 5.7% and reaching 11.5% in those areas with over five million residents. Carpooling accounts for 12.6% of all metropolitan work trips. Automobile ownership is high and continues to grow faster than the population. Yet many US households owned no car in 2001: roughly 8% of all households and 26% of those with incomes below $20,000.[4]

The geography of commuting continues to become more dispersed. Commuting within the main central cities of US metropolitan areas accounted for about one-fourth of all work trips by metropolitan residents, and those from suburbs to central cities another 17%; the rest are trips within or between suburbs (41%), reverse commuting from central city to suburbs (8%), and trips to workplaces outside the metropolitan areas (10%).[5] The increased dispersion of work trips shows up as longer distances traveled at somewhat higher speeds: average commuting distance has risen by about 14% over the decade (to 12.1 miles), while average commuting time has risen by only 9% (to 25.5 minutes).[6] Commuting times are slightly higher in metropolitan areas than in the US as a whole — 26.1 minutes in 2000 — and higher still in larger metropolitan areas, reaching 34 minutes in New York (where the percentage using public transit is especially high). Several metropolitan areas, all of them with high population growth, had a 20% increase in average commuting time over the decade 1990-2000; yet in only one of them, Atlanta, did the average commute exceed 30 minutes in 2000;[7] if we look just at commutes by private vehicles,

[3] These statistics are, respectively, from Pisarski (2006, p. 6), Van Dender (2001, p. 103), and Giuliano (1994, p. 261).

[4] The figures in these paragraphs are computed from Pisarski (2006, pp. xvi, 76-77); again we divide by (1–0.0326) to express them as proportions of work trips rather than of workers; and from Pucher and Renne (2003, Table 6).

[5] Calculated from Pisarski (2006), Fig. ES-2, p. xv.

[6] Pisarski (2006), Figures 307 (p. 51) and 3-68 (p. 101), the latter using the adjustment of the 1990 figure for measurement change as described by Pisarski.

[7] Pisarski (2006), Tables 3-40 (p. 102) and 3-41 (p. 106). Hu and Reuscher (2004, Table 26) document the much higher average travel time for commutes by public transit (47.9 minutes in 2001) compared to private vehicles (22.5 minutes).

their duration averaged 26 minutes in metropolitan areas over 3 million in population and considerably less in smaller areas (Hu and Reuscher 2004, Fig. 11).

Most of these trends apply to other parts of the world as well, although the actual numbers may be quite different. For example, rather different modal splits are observed for the Netherlands: 38% of all trips are by car as driver, 12% by car as passenger, 5% by public transport, 26% by bicycle, and 17% walking.[8] Of course, heavily urbanized areas have a much bigger proportion of trips by public transport — accounting for 23% of person-kilometers, compared to just 12% for the entire nation.[9] Total travel per person (in kilometers per day) is also about 7% less in the heavily urbanized areas, probably reflecting greater accessibility and possibly also the impact of congestion.

Work trips in the Netherlands account for 26% of person-kilometers, with car driver accounting for a higher proportion of them than is the case for all trips (66% *vs*. 54%), while car passenger accounts for a lower proportion of work trips than of all trips (8% *vs*. 21%). Public transport in the Netherlands captures a higher proportion of person-kilometers for work trips than for all trips (16% *vs*. 12%), but this may be because work trips are longer.

The stereotype of the US being more car-dependent than a European country like the Netherlands thus seems confirmed by these figures, and by other information about the European Union in Strelow (2006). It is also reflected in car ownership: 23% of the Dutch households have no car and 56% have one car; in heavily urbanized areas these shares are 40% and 50%, respectively. Local circumstances — including cultural, geographical, and historical differences, but also more "conventional" economic factors such as fuel prices, vehicle taxes, and parking policies — are among the factors commonly thought to explain such differences; more systematic analyses will be discussed below.

2.1.1 Aggregate Demand Models

For more formal analysis, 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

[8] Statistics Netherlands (2007). These figures refer to trips by people of age 12 or more. There were 3.1 trips per person per day on average with average trip length 11.2 km (6.9 miles).

[9] These figures refer to "very heavily urbanized areas," defined as those with more than 2500 addresses per km^2.

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. Because behavior cannot be predicted precisely, an "error term" is added to represent behavior that, to the researcher at least, appears random. Thus a demand function might be represented as:

$$x = f(Z) + \varepsilon \tag{2.1}$$

where x is the quantity demanded, Z is a vector of values of all the relevant characteristics of the good and its potential consumers, $f(\cdot)$ is some mathematical function, and ε is the random error term. Statistical data on x and Z can be used to estimate the function f and the probability distribution of ε.

It is possible to estimate f using *non-parametric methods* that impose no prior assumptions about its shape. More commonly, f is *specified* to be a particular functional form with parameters whose values are to be determined from statistical data. For example, f might be specified as a general quadratic function:

$$f(Z) = \beta_0 + \sum_k \beta_{1k} Z_k + \sum_k \beta_{2k} Z_k^2 \sum_k \sum_{l \neq k} \beta_{3kl} Z_k Z_l \tag{2.2}$$

where Z_k is one of the characteristics included in vector Z, Z_l is a factor that possibly affects the extent the effect of Z_k upon f, and the β parameters are to be estimated empirically. This is an example of an equation that is *linear in parameters*. To see why, define variables $z_0 \equiv 1$, $z_{1k} = Z_k$, $z_{2k} = Z_k^2$, and $z_{3kl} = Z_k Z_l$, for all values of k and l. Combining all these variables into a single vector z and all the corresponding parameters into a single vector β, we can write (2.1) as:

$$x = \beta' z + \varepsilon \tag{2.3}$$

where the prime on β indicates transposition (changing it from a column vector to a row vector); thus $\beta' z$ is the inner product between β and z.[10]

Equation (2.3) is known as a *regression* of x on z. When it is linear in β, as in this example, one can very easily estimate the unknown parameters. Indeed, we chose this example to illustrate that even a quite complex relationship between x and Z can often be represented as a

[10] That is, $\beta' z \equiv \beta_0 z_0 + \Sigma_k \beta_{1k} z_{1k} + \Sigma_k \beta_{2k} z_{2k} + \Sigma_{kl} \beta_{3kl} z_{3kl}$.

linear regression.[11] The most common way of estimating the unknown parameters in (2.3) is *ordinary least squares*, in which the value of β is found that minimizes the sum of squared residuals for a set of observations labeled n=1,…,N:

$$\hat{\beta} = \arg\min_{\beta} \sum_{n=1}^{N} \left(x_n - \beta' z_n\right)^2$$

where now we have indexed each observed data point (x_n,z_n) by its observation label n. Ordinary least squares has particularly nice properties when the random error ε is assumed to have a normal (bell-shaped) distribution. However it is quite possible to estimate regression models non-linear functional forms or with other error distributions.[12]

The non-random part of the demand equation (2.1) is often based on an explicit theory of consumer choice. Such a theory is not necessary in order to specify and estimate a demand function, but it may help by suggesting likely functional forms for f and it is useful for interpreting the results. The most common such theory postulates that a consumer or group of consumers maximizes a *utility function*, $u(x,X)$; this function expresses preferences over the quantities of the good x under consideration and of other goods represented by the vector X. The consumer is limited by a budget constraint, expressed in terms of the price p of x and a price vector P consisting of prices of all the other goods X. Mathematically, then, consumption is determined as a solution to the following constrained maximization problem:

$$\underset{x,X}{Max}\, u(x,X) \quad \text{subject to: } px + P'X = y \tag{2.4}$$

where y is income and again the prime on P transposes it so that $P'X$ is an inner product, expressing the cost of consuming all the goods in vector X. Denoting the solution to (2.4) by vector (x^*,X^*), we note that it depends on prices and income:

---

[11] If the variables Z_k in (2.2) are replaced by their natural logarithms, and the mean values of these logarithms subtracted, then (2.2) becomes the *translog* functional form:

$$f(Z) = \beta_0 + \sum_k \beta_{1k} \widetilde{Z}_k + \sum_k \beta_{2k} \widetilde{Z}_k^{\,2} \sum_k \sum_{l \neq k} \beta_{3kl} \widetilde{Z}_k \widetilde{Z}_l$$

where $\widetilde{Z}_k \equiv \left(\log Z_k - \overline{\log Z_k}\right)$ and the bar indicates a sample average. This function is widely used in cost analysis and regarded as a particularly good approximation to an arbitrary unknown function: see Spady and Friedlaender (1978), Brauetigam (1999), and Chapter 3 of this book.

[12] For basic econometric texts, see Pindyck and Rubinfeld (1998) or Johnston and DiNardo (1997). For an advanced text, see Greene (2003).

$$x^* = x^*(p, P, y)$$
$$X^* = X^*(p, P, y). \tag{2.5}$$

Thus if we knew, or were willing to postulate, a form for the function u, and if we could solve the maximization problem, we would know the form of the demand function (2.1) except for its random term.[13]

The demand functions (2.5), when derived in this way, can be used to define a very useful quantity. We simply substitute them into the utility function to see how much utility can be achieved with a given set of prices and income. The result is known as the *indirect utility function*, often written as V:

$$V(p, P, y) = u\big(x^*(p, P, y), X^*(p, P, y)\big).$$

The indirect utility function has a property known as *Roy's identity*:

$$x^* = -\frac{\partial V / \partial p}{\partial V / \partial y} \tag{2.6}$$

where it is understood that all quantities in (2.6) depend on p, P, and y. We will make use of the indirect utility function and Roy's identity when we discuss disaggregate demand analysis in Section 2.2. For simplicity, our notation for demand functions will omit the asterisks in (2.5) and (2.6).

The demand functions (2.5) were defined as resulting from an individual consumer's optimization. An aggregate demand function can simply be derived from individuals' demands by summing quantities demanded over consumers. Under some quite restrictive conditions, the resulting aggregate demand function may look as if it could have resulted from optimization by a single "representative" consumer, which can be convenient in analyses. Specifically, a necessary and sufficient condition for this to be true is that individuals' indirect utility functions take the "Gorman form", meaning that for an individual i it can be written as $V_i(p,P,y) = a_i(p,P) + b(p,P)\cdot y_i$, where the a_i-term can vary across consumers but the b-term cannot (Varian 1992, p. 154). This condition is satisfied for several utility functions commonly used for theoretical analysis, but not for those underlying most empirical work.

[13] Utility theory is treated in most texts on microeconomic theory. Varian (1992) provides a concise entry-level graduate treatment.

2.1.2 Cross-Sectional Studies of Metropolitan Areas

Many studies have used aggregate data, such as are commonly reported by transit authorities or local governments, to study influences on travel behavior across cities. We illustrate here with three.

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. They show that ridership is positively related to city population density and gross national product per capita. Applying this model to five North American 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.

Winston and Shirley (1998) examine the shares of US work trips across metropolitan areas made by any of five different modes (two private, two public, plus taxi) within any of ten different time-of-day intervals. Their data are from the 1990 journey-to-work census. Among the many interesting findings are these: travel time has the largest effect on choice for trips of moderate distance, being less important for both short and long trips; route coverage and service frequency is important for public transit use, especially for commuters making long trips; and workers in the finance, real estate, and insurance industries tend to travel earlier in the day if they live further west, suggesting that they need to overlap with the operating hours of New York or other East Coast banking and financial markets.

Black (1990) uses regression analysis on data from 120 US 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%, averaging 5.4%. It is higher in cities with military bases or universities, in small cities, and where incomes are low. Black's study is a useful reminder that walking can be an important journey-to-work mode, its prevalence being sensitive to land uses and demographics. Plaut (2004) reports that in Tel Aviv, 9.4% of workers walk to work, a fraction that varies by gender, age, and household status.[14]

[14] Computed as the weighted average of "percent walking to work" for the columns pertaining to Tel Aviv in Plaut's Table 7, p. 246.

2.1.3 Cross-Sectional Studies 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"); 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. (Unfortunately, this does not control for the possibility that vehicle ownership is influenced by automobile share itself — an "endogeneity problem" that we discuss later.) 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 on trip-making, 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.[15] This illustrates a pervasive feature of travel-demand analysis: many of the variables to be explained are limited in range, making ordinary regression analysis 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). Here we describe one such technique that is applicable to aggregate data.

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

$$\log\left(\frac{x}{1-x}\right) = \beta' z + \varepsilon \tag{2.7}$$

[15] The tobit model, also known as a censored model, postulates a "latent" (unobserved) variable x^* that is explained by an ordinary regression equation like $x^*=\beta' z+\varepsilon$, and an observed variable $x=\max\{x^*,0\}$. The idea is that we observe x^* except when it is below a logical cutoff for observability (zero in this example), in which case we observe only that it fell below that cutoff. See any econometrics text, for example Johnston and DiNardo (1997, Ch. 13).

where β is a vector of parameters, z is a vector of independent variables, and ε is an error term with infinite range.[16] Equivalently,

$$x = \frac{\exp(\beta' z + \varepsilon)}{1 + \exp(\beta' z + \varepsilon)}. \tag{2.8}$$

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 (2.8) ensures that they sum to one:

$$x_i = \frac{\exp(\beta' z_i + \varepsilon_i)}{\sum_{j=1}^{J} \exp(\beta' z_j + \varepsilon_j)} \tag{2.9}$$

where J is the number of modes.[17] Anas (1981) and Mackett (1985) 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. The study by Winston and Shirley (1998), mentioned in the previous subsection, also uses this formula.

2.1.4 Studies Using Time Series 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 time. They may also postulate "inertia" by including among the explanatory variables one

[16] This assumes that x does not take values 0 or 1 in the observed data. If it does, an alternative would be a single- or double-censored tobit model. The double-censored tobit model is like that in the previous footnote except now $x=0$ if $x^*<0$, $x=x^*$ if $0\le x^*\le 1$, and $x=1$ if $x^*>1$.

[17] 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 $\varepsilon=\varepsilon_1-\varepsilon_2$.

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. This measured difference is especially sensitive to the treatment of autocorrelation.

It is common to combine cross-sectional and time-series variation, for example using observations from many separate locations and two or more time periods. 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 finds 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 cross-sectional time series is that by Petitte (2001), who estimates fare elasticities from station-level data on Metrorail ridership in Washington, D.C.

Studies using cross-sectional time series 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 error 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.[18] 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 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

[18] A good example is the study of transit use in US cities by Baum-Snow and Kahn (2000).

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.

2.1.5 Summary of Key Results of Aggregate Studies

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, p. 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, Mayworm and McEnroe 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). Unfortunately, the time period over which these elasticities apply varies from study to study and is often unstated; usually the elasticities are measured using data covering a period of a few months to about two years.

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, p. 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 elasticity for fuel itself is apparently two to three times as large, indicating that changes in fuel price affect the composition of the motor-vehicle fleet more than its usage.[19]

Several studies have found intriguing regularities in the per-capita or per-household expenditures of time and money on travel. In a series of papers and reports, Yacov Zahavi has

made these regularities the centerpiece of a model of travel demand known as Unified Mechanism of Travel. However, most scholars have concluded that the regularities can be explained by more conventional models and, furthermore, that the regularities are only approximations and that violations of them occur as predicted by economic theory (Schafer 2000). For example, Kockelman (2001) rejects the hypothesis of fixed travel-time budgets using data from the San Francisco Bay Area. She finds rather 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

Transportation is defined with respect to particular locations. Naturally, the way the land at those locations is used — the types and densities of buildings occupying them and the activities that occur there — are among the most important factors influencing travel decisions. For example, researchers have found that public transit ridership in a city is more heavily influenced by the number of jobs in the city's downtown area than by almost any other factor (Barnes 2005). Furthermore, transportation facilities have significant effects on land use. Thus, the two-way inter-relationship is important analytically and, because transportation and land use have significant spillover effects on quality of life for nearby residents, it fuels contentious policy debates.[20] Here, we consider just one of the two-way directions of influence: from land use to travel.

One approach to understanding this influence is to study the relationships among aggregate land-use and travel measures, typically at the level of a metropolitan area. One needs to keep track of which is causing which, and to carefully disentangle related factors. For example, one obviously must control for variables like income and fuel price, which are correlated with land-use characteristics and which independently affect travel. Some influential studies claiming far-reaching effects of urban density on travel, such Newman and Kenworthy

[19] 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.

[20] Giuliano (2004) provides a good review.

(1989), have been severely criticized for neglecting this fundamental requirement of statistical analysis.

But several other pitfalls afflict such aggregate studies. Most importantly, because the interaction is two-way, land-use patterns are not exogenous; instead, they respond strongly to transportation systems. As a result, trying to explain travel decisions with land-use patterns risks confusing cause with effect. Statistical techniques exist to handle this type of "endogeneity problem," notably the use of "instrumental variables," in this case variables related to land use but believed to be exogenous (see any of the econometrics texts listed in Section 2.1.1). An example might be the age of the city, or a measure of its density several decades earlier. Roughly speaking, the variable feared to be endogenous (land use) is replaced, in the equation explaining travel, by a predicted version formed by regressing the land-use variable on all exogenous variables in the system, including the instrumental variables. We discuss instrumental variables further in Section 2.4.5.

Even when the two directions of causality between transportation and land use are correctly measured, they can produce surprising policy paradoxes. For example, expanding highways to relieve congestion may attract development that undermines the intended effect; this is just one of many examples 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).

A similar endogeneity problem afflicts measures of transportation infrastructure or service provision, such as the extent of roadways and rail transit lines or the amount of bus service provided. Presumably policy makers provide infrastructure and services at least partly in response to actual or expected travel. Therefore any factors, unaccounted for by the model, that increase a particular type of travel may also tend to increase the associated infrastructure and service provision. Ignoring this reverse causality could lead to an overstatement of the effect of infrastructure or service provision on travel and, because infrastructure is closely related to urban

density, an understatement of the effect of land-use patterns (since some of their effects would be attributed instead to the transportation service variables).

Another problem is that land-use patterns cannot be modified by fiat. Even in countries with 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 US, the changes in land use that can feasibly be accomplished through policy are quite limited. The urbanized area of Portland, Oregon, remains relatively low density despite three decades of stringent policies aimed at increasing urban density; one reason is that Portland's policies have apparently diverted urban development to more outlying jurisdictions outside its control (Downs 2004, Jun 2004). Thus, even when it can be established that a certain type of land-use pattern has favorable effects on travel, there may be no way to bring about the desired change.

We turn now to empirical findings on how land use affects travel demand at an aggregate level. 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. These effects are estimated to be even larger if transportation infrastructure and services are omitted on the grounds that they are endogenous. 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: average commuting times have risen much more slowly than the amount of congestion on any given road.

Bento *et al.* (2004) ask how various land-use and transportation variables, intended to measure "urban sprawl," 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. 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 to its lower vehicle use. This limiting effect of density upon vehicle-miles traveled is consistent with the Dutch figures quoted in the introduction to Section 2.1.

Other contexts, or more precise breakdowns, may yield different results. Schwanen, Dijst and Dieleman (2004) examine the effects of urban size and form on travel for 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.

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 isolated 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 (1995) 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. Recent work has suggested that the neighborhood-level effects of land use on travel can be understood by examining how trip costs are affected (Boarnet and Crane 2000); this observation opens the door to research explaining more specifically why land use has impacts in some settings and not others.

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. Even when jobs-housing balance is achieved within a community, people do not predominately choose nearby jobs. Thus, for example, new exurban communities intended to be relatively self-contained have generally discovered that most residents work elsewhere.[21]

A quite different approach to the linkage between land use and transportation is to build detailed computer models that simulate both land-use decisions, such as whether to develop a property and at what density, and travel decisions. Typically this is done at the level of small- to moderate-sized zones within an urban area. Because the potential number of relevant decisions is enormous, it has proven difficult to simultaneously meet the goals of theoretical rigor (generally based on micro-economic theory), adequate spatial detail, computational tractability, data availability, and comprehensibility. Nevertheless such models have achieved some success. Anas (1982) describes a model, known as the Chicago Area Transportation/ Land Use Analysis System (CATLAS), calibrated for Chicago and used for a number of interesting policy evaluations. One of them was to predict impacts of a new transit line from downtown Chicago to Midway Airport. Subsequent studies by other researchers have verified two of its predictions: substantial impacts on housing prices for properties within half a mile of the line, and virtually no price impact on housing more than 1.5 miles distant. Another notable result is that much of the impact appeared during the several years after the line was approved but not yet constructed.[22]

More recently, Anas and Liu (2006) describe a model that attempts to more fully represent economic decision-making with computational tractability. The land-use portion of the model keeps track of product prices, wages, housing rents and prices, and stocks of buildings, allowing for dynamic decision-making accounting for the durability of decisions about buildings. The transportation portion of the model contains the usual components of a complete travel model as described earlier in Section 2.1, including equilibration of route choices on networks representing available transit lines and roads. Safirova *et al.* (2006) combine the same underlying land-use model with a different transportation model, known as START, and use it for policy analysis for the Washington, D.C., area. Another significant recent model system is UrbanSim,

[21] See Giuliano and Small (1993); Giuliano (1991); Cervero (1996); and Schwanen, Dijst and Dieleman . (2004).

[22] See McDonald and Osuji (1995) and McMillen and McDonald (2004).

which adopts an open computer architecture to facilitate adaptation by other researchers for specific purposes. It is described by Waddell (2002), who also provides a useful review of older models and of modeling strategies specifically designed for the needs of transportation agencies.

2.2 Disaggregate Models: Methods

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 such data are 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*.[23]

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,...,J chooses the one that maximizes utility as given by

$$U_{jn} = V\left(z_{jn}, s_n; \beta\right) + \varepsilon_{jn} \tag{2.10}$$

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), β is a vector of unknown parameters, and ε_{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, may depend on income and prices and thus implicitly incorporate a budget constraint.

[23] Reviews with a transportation focus include Train (2003), Ben-Akiva and Bierlaire (2003), Koppelman and Sethi (2000), and Ben-Akiva and Lerman (1985).

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

$$\begin{aligned} P_{in} &= \Pr[U_{in} > U_{jn} \qquad \text{for all } j \neq i] \\ &= \Pr[\varepsilon_{jn} - \varepsilon_{in} < V_{in} - V_{jn} \qquad \text{for all } j \neq i] \\ &= \int_{-\infty}^{\infty} F_i(V_{in} - V_{1n} + \varepsilon_{in}, \ldots, V_{in} - V_{Jn} + \varepsilon_{in}) d\varepsilon_{in} \end{aligned} \tag{2.11}$$

where F_i is the partial derivative of F with respect to its i-th argument. F_i is thus the probability density function of ε_{in}, conditional on the inequalities in (2.1).

Suppose the cdf $F(\cdot)$ is multivariate normal. Then (2.11) is the *multinomial probit* model with general covariance structure. However, neither F nor F_i can be expressed in closed form; instead, equation (2.11) 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), (2.11) becomes

$$P_{1n} = \Phi\left(\frac{V_{1n} - V_{2n}}{\sigma}\right) \tag{2.12}$$

where Φ is the cumulative standard normal distribution function (a one-dimensional integral) and σ is the standard deviation of $\varepsilon_{1n}-\varepsilon_{2n}$. In equation (2.12), σ cannot be distinguished empirically from the scale of utility, which is arbitrary; for example, doubling σ has the same effect as halving both V_1 and V_2. Hence it is conventional to normalize by setting σ=1.

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

$$\Pr[\varepsilon_{jn} < x] = \exp(-e^{-\mu x}) \tag{2.13}$$

for all real numbers x, where μ is a scale parameter. Here the convention is to normalize by setting μ=1. With this normalization, McFadden (1973) shows that the resulting probabilities calculated from (2.11) have the logit form:

$$P_{in} = \frac{\exp(V_{in})}{\sum_{j=1}^{J} \exp(V_{jn})}. \tag{2.14}$$

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

The binary form of (2.14), *i.e.* the form with J=2, is:

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

If graphed as a function of (V_{1n}–V_{2n}), this equation looks quite similar to (2.12).

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).[24] Furthermore, both probit and logit may be generalized by

[24] 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 (β/σ_ε), where σ_ε^2 is the variance of each of the random terms ε_{in}. In the case of probit, the variance of ε_{1n}-ε_{2n}, which is $2\sigma_\varepsilon^2$, is set to one by the conventional normalization; hence $\sigma_\varepsilon^{PROBIT} = 1/\sqrt{2}$. In the case of logit, the normalization μ=1 in equation (2.13) implies that ε_{in} has standard deviation $\sigma_\varepsilon^{LOGIT} = \pi/\sqrt{6}$ (Hastings and Peacock 1975, p. 60). Hence to make logit and iid probit comparable, the logit coefficients must be divided by $\sigma_\varepsilon^{LOGIT} / \sigma_\varepsilon^{PROBIT} = \pi/\sqrt{3} = 1.814$.

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 β. More general forms such as Box-Cox and Box-Tukey transformations are studied by Gaudry and Wills (1978). Just as with regression analysis, V can be linear in *parameters* while still nonlinear in *variables*, just by specifying new variables equal to nonlinear functions of the original ones. For example, the utility (2.10) on mode i of a traveler n with wage w_n facing travel costs c_{in} and times T_{in} could be specified as:

$$V_{in}(c_{in}, T_{in}, w_n; \beta) = \beta_1 \cdot (c_{in} / w_n) + \beta_2 T_{in} + \beta_3 T_{in}^2 . \tag{2.15}$$

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,[25] then the linear-in-parameters specification is simply written as

$$V_{in} = \beta' z_{in} \tag{2.16}$$

where β' is the transpose of column vector β.

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 β in (2.16) 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:

$$L(\beta) = \sum_{n=1}^{N} \sum_{i=1}^{J} d_{in} \log P_{in}(\beta) \tag{2.17}$$

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. Not only does maximizing this function give us the *maximum-likelihood estimates* of parameters, often written as $\hat{\beta}$; the derivatives of L also provide information about the statistical uncertainty in $\hat{\beta}$, usually summarized as its *variance-covariance matrix*, denoted $Var(\hat{\beta})$. The diagonal elements

[25] In this example there are three such combinations, so z_{in} is a vector with three components: c_{in}/w_n, T_{in}, and T_{in}^2.

of this matrix give the variances (*i.e.* the squares of the standard deviations) of the individual parameters, while the off-diagonal elements give the covariances between pairs of parameters. This information is crucial to knowing how firm we can be in making quantitative statements about the parameters based on the particular data set used.

A correction to (2.17) is available for choice-based samples, *i.e.*, those in which the sampling frequencies depend on the choices made. Choice-based samples often are available for practical reasons, such as the convenience of conducting a mode-choice survey at train stations and bus stops. 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 how to incorporate aggregate information to improve efficiency.

Recent work has shown that demand functions built up from discrete-choice models at the individual consumer level can be estimated using data solely on aggregate market shares. Bresnahan, Stern and Trajtenberg (1997) provide a particularly clear exposition. Basically, they use an extension of the idea, described earlier in connection with regression analysis, of minimizing the sum of squared residuals. Recall that a residual is the discrepancy between an observed quantity (such as amount of a good consumed) and that predicted by the model at any particular set of parameters; it is those parameters that are adjusted to minimize the sum of squared residuals. In the treatment by Bresnahan, Stern and Trajtenberg , the sum of squared residuals is generalized to a quadratic form in the vector of residuals, a common procedure in econometric models. The innovation, derived from Berry (1994), is in constructing the residuals themselves, one for each alternative j; they are formed as the differences between the indirect utility V_j (as computed from the discrete-choice model at a trial set of parameter values) and the indirect utility δ_j that is implied by its observed market share. (The authors assume that the discrete-choice model contains no variables describing characteristics of consumers, which enables us to omit subscript n from indirect utility.) The values of δ_j used in this procedure are determined (again for a given set of trial parameters) by solving the equation for probabilities,

e.g. (2.14), so as to give each indirect utility V_j as a function of the probabilities $P_i(i=1,\ldots,J)$; the solution is interpreted as giving δ_j in terms of the observed market shares.[26]

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 (2.14). With V linear in β, the logit log-likelihood function is globally concave in β, so finding a local maximum assures finding the global maximum. Fast computer routines to do this are widely available. 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 (2.11). 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 (2.10) 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% of the highest-income households owned no car in 2001 (Pucher and Renne 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 strongly income can affect probability.

2.2.3 Interpreting Coefficient Estimates

[26] The fact that (2.14) can be inverted is due to a property of the logit model: if a full set of alternative-specific constants is included, the predicted choice shares for each of the alternatives can be made exactly equal to the observed shares in the sample by setting parameter β to its maximum-likelihood estimate.

It is useful for interpreting empirical results to note that a change in $\beta' z_{in}$ in (2.16) by an amount of ±1 increases or decreases the relative odds of alternative *i*, compared to each other alternative, by a factor exp(1)=2.72. Thus a quick gauge of the behavioral significance of any particular variable can be obtained by considering the size of typical variations in that variable, 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; this marginal effect however depends on the values of the variables.

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

$$V_{in} = \alpha_i + \beta' z_{in} . \tag{2.18}$$

Since only utility differences matter, at least one of the alternative-specific constants α_i must be normalized, usually to zero: that alternative then serves as a "base alternative" for comparisons.

The constant α_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 (2.18) is really a special case of (2.16) in which one or more of the variables *Z* are *alternative-specific dummy variables*, D^k, defined by D^k_{jn}=1 if *j*=*k* and 0 otherwise (for each *j*=1,…,*J*). (Such a variable does not depend on *n*.) In this notation, parameter α_i in (2.18) is viewed as the coefficient of variable D^i included among the *z* variables in (2.16). Such dummy variables can also be interacted with (*i.e.*, multiplied by) any other variable, making it possible for the latter variable to affect utility in a different way for each alternative. All such variables

and interactions may be included in z, and their coefficients in β, thus allowing (2.16) still to represent the linear-in-parameters specification.

The most economically meaningful quantities obtained from estimating a discrete-choice model are often ratios of coefficients, which represent marginal rates of substitution — that is, the rates at which two variables can be traded against each other without changing utility. 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 (2.15) is

$$(v_T)_{in} \equiv -\left(\frac{dc_{in}}{dT_{in}}\right)_{V_{in}} \equiv \frac{\partial V_{in}/\partial T_{in}}{\partial V_{in}/\partial c_{in}} = \left(\frac{\beta_2 + 2\beta_3 T_{in}}{\beta_1}\right)\cdot w_n, \tag{2.19}$$

which varies across individuals since it depends on w_n and T_{in}. We emphasize that this is a marginal concept: it is a measure, per unit of time, that applies to small time savings. For a larger time saving, say from T_{in}^0 to T_{in}^1, the total value to travelers would be the integral of (2.19), which can be computed analytically:

$$\int_{T_{in}^0}^{T_{in}^1} (v_T)_{in}\, dT = \left(\frac{\beta_2 + 2\beta_3 \overline{T}}{\beta_1}\right)\cdot w_n \cdot \Delta T$$

where $\overline{T} = (T_{in}^0 + T_{in}^1)/2$ and $\Delta T = (T_{in}^1 - T_{in}^0)/2$. Thus, in this example, the value of a finite time saving is computed as the time saving multiplied by the marginal value of time, with the latter evaluated at the average trip duration before and after the change.

As a more complex example, suppose we extend equation (2.15) by adding alternative-specific dummies, both separately (with coefficients α_i) and interacted with travel time (with coefficients γ_i):

$$V_{in} = \alpha_i + \beta_1 \cdot (c_{in}/w_n) + \beta_2 T_{in} + \beta_3 T_{in}^2 + \gamma_i T_{in} \tag{2.20}$$

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

$$(v_T)_{in} = \left(\frac{\beta_2 + 2\beta_3 T_{in} + \gamma_i}{\beta_1} \right) \cdot w_n . \tag{2.21}$$

Now the value of time varies across modes even with identical travel times, due to the presence of γ_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 β is asymptotically normally distributed with mean b and variance-covariance matrix Σ, then a function $f(\beta–b)$ is asymptotically normally distributed with mean zero and variance-covariance matrix $(\nabla f)\Sigma(\nabla f)'$, where ∇f is the vector of partial derivatives of f.[27] A more accurate estimate may be obtained by taking repeated random draws from the probability distribution of β (that distribution being estimated along with β itself), and then examining the values of f corresponding to these draws. As an example, the 5th and 95th percentile values of those values of f define a 90-percent confidence interval for f.[28]

2.2.4 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

[27] Chow (1983, pp. 182-3). The result requires asymptotic convergence of β 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 β tends to diminish proportionally to $1/\sqrt{N}$.) In the simple case where $v_T=\beta_2/\beta_1$, it implies that the standard deviation σ_v of v_T obeys the intuitive formula: $(\sigma_v/v_T)^2 \cong (\sigma_1/\beta_1)^2 + (\sigma_2/\beta_2)^2 - 2\sigma_{12}/(\beta_1\beta_2)$, where σ_1 and σ_2 are the standard deviations of β_1 and β_2 and where σ_{12} is their covariance. Such an approximation requires that σ_1/β_1 and σ_2/β_2 be small, which in turn helps ensure that the variance of β_1/β_2 exists (it would not exist, for example, if the mean of β_2 were zero).

travel, which depends on travel times. Thus the application 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. In most cases, endogeneity can 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. (See Section 2.4.5 for an important exception.) Nevertheless, measuring the values of attributes z_{in}, which typically vary by alternative, is more difficult than it may first appear. How does one know the attributes 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 measured values for explanatory variables 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.

[28] This method is described by Armstrong, Garrido and Ortúzar (2001). See Train (2003, Ch. 9) for how to take random draws from distributions.

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.[29]

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 type of data described thus far measures *revealed preference* (RP) information, that reflected in actual choices. There is growing interest in using *stated preference* (SP) data, based on responses to hypothetical situations (Hensher 1994). SP data permit more control over the ranges of and correlations among the independent variables 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

[29] Brownstone *et al*. (2003); Lam and Small (2001); Small, Winston and Yan (2005).

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 Morikawa 1990; Louviere and Hensher 2001). As 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 μ for the two parts of the sample (with one but not both normalized), any alternative-specific constants (see next subsection), and any critical behavioral coefficients that may differ. For example, in the logit model of (2.14) and (2.18), 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_{1in}. Letting $\beta_2' z_{2in}$ 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 (2.17) becomes the following:

$$L(\alpha,\beta,\mu^{SP}) = \sum_{n\in RP}\sum_{i=1}^{J} d_{in}\left\{\alpha_i^{RP} + \beta_1^{RP} z_{1in} + \beta_2' z_{2in} - \log\sum_{j=1}^{J}\exp(\alpha_j^{RP} + \beta_1^{RP} z_{1jn} + \beta_2' z_{2jn})\right\}$$

$$+ \sum_{n\in SP}\sum_{i=1}^{J} d_{in}\left\{\mu^{SP}\cdot(\alpha_i^{SP} + \beta_1^{SP} z_{1in} + \beta_2' z_{2in}) - \log\sum_{j=1}^{J}\exp\left[\mu^{SP}\cdot(\alpha_j^{SP} + \beta_1^{SP} z_{1jn} + \beta_2' z_{2jn})\right]\right\}$$

where (α,β,μ^{SP}) denotes the entire set of parameters shown on the right-hand side (excluding α_1^{RP} and α_1^{SP}, which can be normalized to zero). This expression is not as complicated as it looks: the first term in curly brackets is just the logarithm of the logit probability (2.14) for RP observations, while the second is the same thing for SP observations except utility V_{in} is multiplied by scale factor μ^{SP}.

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.

2.2.5 Randomness, Scale of Utility, and Measures of Benefit

The variance of the random utility term in equation (2.10) 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 (2.14), suppose systematic utility is linear in parameter vector β as in (2.16). If all the elements of β are small in magnitude, the corresponding variables have little effect on probabilities so choices are dominated by randomness. If the elements of β are large, most of the variation in choice behavior is explained by variation in 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 $-\Sigma_n\Sigma_j P_{jn}\log P_{jn}$, which is larger when the choice probability is divided evenly among the alternatives than when one alternative is very likely and others very unlikely. Anderson, de Palma and Thisse. (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_{in} across alternatives 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_{in} will nevertheless be chosen; so it is desirable for such an

alternative to be offered and to be made as attractive as possible. A natural measure of the desirability of choice set J is the expected maximum utility of that set, which for the logit model has the convenient form:

$$E \max_{j} (V_j + \varepsilon_j) = \mu^{-1} \log \sum_{j=1}^{J} \exp(\mu V_j) + \gamma \tag{2.22}$$

where γ=0.5772 is Euler's constant (it accounts for the nonzero mean of the error terms ε_j in the standard normalization). Here we have retained the parameter μ from (2.13), rather than normalizing it, to make clear how randomness affects expected utility. When the amount of randomness is small (large μ), the summation on the right-hand side is dominated by its largest term (let's denote its index by j*); expected utility is then approximately $\mu^{-1}\cdot\log[\exp(\mu V_{j*})] = V_{j*}$, the utility of the dominating alternative. When randomness dominates (small μ), all terms contribute more or less equally (let's denote their average utility value by V); then expected utility is approximately $\mu^{-1}\cdot\log[J\cdot\exp(\mu V)] = V+\mu^{-1}\cdot\log(J)$, which is the average utility plus a term reflecting the desirability of having many choices.

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

$$W = \frac{1}{\lambda_n} \log \sum_{j=1}^{J} \exp(V_{jn}), \tag{2.23}$$

a formula also derived by Williams (1977). (The constant γ 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, λ_n cannot be estimated directly.[30] However, typically it can be determined from Roy's Identity:

$$\lambda_n = -\frac{1}{x_{in}} \cdot \frac{\partial V_{in}}{\partial c_{in}} \tag{2.24}$$

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

There is an important condition for the validity of (2.23), which is that any *income effects* in the response of transportation demand be small. Changes in transportation costs or service quality exert most of their influence directly: travelers respond by shifting toward or away from the particular choices that have become more or less desirable. However, there is also a somewhat indirect effect caused by travelers being made more or less well off overall in terms of their standard of living. For example, suppose the price of gasoline rises by 50 percent, and gasoline accounts for 2 percent of total expenditures by a particular group of people. The price increase makes activities that use gasoline somewhat less attractive than before (how much so is something we examine in Section 3.4.6), and so tilts consumers toward choosing less of them. (They might travel less, or they might buy smaller cars with less energy-consuming options.) This is sometimes expressed as a *compensated price elasticity*: the responsiveness to price they would exhibit if their incomes were supplemented so as to leave them just as well off as before. But presuming they are not so compensated, the price increase makes them worse off overall by reducing their discretionary income for other consumption. As a first-order approximation, in the above example people would behave as though they had 1 percent less discretionary income (50 percent of 2 percent). If gasoline is a *normal* good — one whose consumption increases with income — then this indirect effect also reduces gasoline consumption, by an amount proportional

[30] 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 λ. 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.

to the product of the income elasticity of gasoline and the fraction of income spent on gasoline.[31] This indirect effect is called an *income effect*. The combined direct and indirect effects may be expressed as the total price elasticity, sometimes called the *uncompensated price elasticity* to distinguish it from the compensated elasticity. For normal goods the two effects reinforce each other, so the uncompensated price elasticity is larger in magnitude than the compensated one.[32]

Equation (2.23) is valid when λ_n , the marginal income of utility, is the same before and after the change under consideration. This condition is as assured as a close approximation if any income effects of the change are small — as is likely for transportation analysis because the fraction of income spent on any one transportation activity is usually quite small.

2.2.6 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, Ch. 6) discuss several methods.

The most straightforward and common is *sample enumeration*. A sample of consumers is drawn, each assumed to represent a subpopulation with identical observable characteristics. (The estimation sample itself may satisfy this criterion and hence be usable as an enumeration sample.) Each individual's choice probabilities, computed using the estimated parameters, predict the shares of that subpopulation choosing the various alternatives. These predictions can then simply be added, weighting each sample member according to the corresponding subpopulation size. Standard deviations of forecast values can be estimated by Monte Carlo simulation methods.

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

---

[31] The income elasticity of a good is the ratio of the percentage change in its consumption to the percentage change in income, holding prices constant.

[32] The relationship between the uncompensated and price elasticities, ε^u and ε^c, is given by the Slutsky equation:

$$\varepsilon^u = \varepsilon^c - s \cdot \eta_y$$

where s is the share of income spend on the good in question and η_y is its income elasticity. Recall that both ε^u and ε^c are negative; hence the negative sign assures that the income effect (the last term) reinforces the compensated price elasticity. See Varian (1992) or any microeconomics textbook for a derivation.

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 (2.15) with β_3=0. Then the probability of choosing bus travel is:

$$P_{2n} = \frac{1}{1+\exp\left[\left(\beta_1 / w_n\right)\cdot\left(c_{1n}-c_{2n}\right)+\beta_2\cdot\left(T_{1n}-T_{2n}\right)\right]}\,. \tag{2.25}$$

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

$$\frac{\partial P_{2n}}{\partial c_2} = \left(\beta_1 / w\right)\cdot P_{2n}\cdot\left(1-P_{2n}\right). \tag{2.26}$$

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. Applying (2.26) to each half of the population, we see that the rate of change of aggregate bus share with respect to bus fare is $(\beta_1/w)\cdot[½(0.9)(0.1)+½(0.1)(0.9)] = 0.09\cdot(\beta_1/w)$. But if we were to apply (2.26) as though there were a single representative traveler with P_2=0.5, we would get $(\beta_1/w)(0.5)(0.5) = 0.25(\beta_1/w)$, more than twice the true value. 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 $\varepsilon_{1n}-\varepsilon_{2n}$) who have a close enough decision to be affected.

McFadden and Reid (1976) derive a more formal 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 (2.12) is written in the form $P_1 = \Phi\left(\beta' z\right)$, then the expected aggregate share is

$$\overline{P}_1 = \Phi\left(\frac{\beta'\bar{z}}{\sqrt{1+\sigma^2}}\right) \tag{2.27}$$

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

Equation (2.27) illustrates a danger in using aggregate models for policy forecasts. If an aggregate probit model fitting $\overline{P}_1$ to $\bar{z}$ were estimated, its coefficients would correspond to $\beta/\sqrt{(1+\sigma^2)}$. If a policy being investigated changed σ, these coefficients would no longer accurately represent behavior under the new policy.

2.2.7 Specification

Like most applied statistical work, travel demand analysis requires balancing completeness against tractability. A 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 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 (2.15) 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. We discuss some other specification issues in connection with an example in Section 2.3.3.

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) and for selection among competing model specifications (Berger and Pericchi 2001).

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 of *transferability* 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 choice shares (Koppelman and Rose 1985). The adjustment procedure is sometimes called calibration, and requires an iterative algorithm for matching the choice shares predicted by the model to observed shares; the match is exact so the algorithm is numerical rather than statistical.[33]

2.2.8 Ordered and Rank-Ordered Models

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

In such a case, an *ordered response* model might be assumed. In this model, the choice of individual n is determined by the size of a "latent variable" $y_n^* = \beta' z_n + \varepsilon_n$, with choice j occurring if this latent variable falls in a particular interval $[\mu_{j-1},\mu_j]$ of the real line, where $\mu_0=-\infty$ and $\mu_J=\infty$. The interval boundaries $\mu_1,...,\mu_{J-1}$ are estimated along with β, except that one of them can be normalized arbitrarily if $\beta' z_n$ contains a constant term. The probability of choice j is then

$$P_{jn} = \Pr[\mu_{j-1} < \beta' z_n + \varepsilon_n < \mu_j] = F(\mu_j - \beta' z_n) - F(\mu_{j-1} - \beta' z_n) \qquad (2.28)$$

[33] A commonly used algorithm, sometimes known as a contraction procedure, adjusts the alternative-specific constants $\{\alpha_i\}$ at each iteration by an amount $\log[s_i] - \log[\hat{s}_i(\alpha_1,...,\alpha_J)]$, where s_i is the observed share and $\hat{s}_i$ is the share predicted by the model (Train 1986, p. 105).

where $F(\cdot)$ is the cumulative distribution function assumed for ε_n. In the *ordered probit* model $F(\cdot)$ is standard normal, while in the *ordered logit* model it is logistic, *i.e.* $F(x) = \left[1 + \exp(-x)\right]^{-1}$. 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.[34] 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 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 Disaggregate Models: Examples

[34] Rank-ordered logit, sometimes called "expanded logit" or "exploded logit," is analyzed by Beggs, Cardell and Hausman (1981) and Hausman and Ruud (1987) Beggs *et al.* call it "ordered logit," but that name is now usually reserved for an ordered response model as described here.

Discrete-choice models have been estimated for nearly every conceivable travel decision, forming a body of research that cannot possibly be reviewed here.[35] 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^o_{in}, the out-of-vehicle travel time including walking, waiting, and transferring. Cost c_{in} includes parking, tolls, gasoline, and maintenance (Train 1980, p. 362). The estimated utility function is:

$$\begin{array}{cccccc} V = -0.0412 \cdot c/w - 0.0201 \cdot T - 0.0531 \cdot T^o - 0.89 \cdot D^1 - 1.78 \cdot D^3 - 2.15 \cdot D^4 \\ \quad (0.0054) \quad (0.0072) \quad (0.0070) \quad (0.26) \quad (0.24) \quad (0.25) \end{array}, \qquad (2.29)$$

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.

[35] For additional examples, see McCarthy (2001, Ch. 3-4) and Small and Winston (1999).

This utility function is a simplification of (2.20) (with $\beta^3=\gamma^j=0$), except that travel time is broken into two components, T and T^o. Adapting (2.21), we see that the "value of time" for each of these two components is assumed to be 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); this is probably 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 fit of Train's model's could undoubtedly be 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 such a variable 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 (2.29), then, is as a "reduced-form" model that implicitly incorporates 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.

More complete models typically aim to directly measure some of the preferences indicated here by coefficients of alternative-specific dummy variables. Currie (2005) compares results of ten mode-choice studies that estimate a "transfer penalty," *i.e.*, that include a dummy variable indicating if a transfer was necessary (or how many were necessary) for the trip being explained. They measure the penalty in terms of the equivalent number of minutes of in-vehicle time, which means they are measuring the rate of substitution between additional travel time and the saving of one transfer. The average penalty ranges from 8 minutes for transfers between subway lines to 22 minutes for transfers between bus lines, with intermediate values for transfers where one or both modes is suburban rail or light rail. They also compare alternative-specific constants, finding that "bus rapid transit," a type of bus service designed to mimic rail in convenience, does indeed achieve mode-specific constants comparable to those of rail.

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 assumes the consumer needs to arrive *before* the work-start time in order to avoid a penalty.

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 $\text{Max}\{S_D,0\}$ and "Schedule Delay Early", *SDE*, as $\text{Max}\{-S_D,0\}$. 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:

$$\begin{matrix} V = -0.106 \cdot T - 0.065 \cdot SDE - 0.254 \cdot SDLE - 0.58 \cdot DL \\ \qquad\quad (0.038) \qquad (0.007) \qquad\quad (0.030) \qquad\quad (0.21) \end{matrix} . \tag{2.30}$$

Here we exclude two variables used by Small 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, occupation, car occupancy, and stated work-hour flexibility.

Figure 2.1 shows utility function (2.30), 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;[36] and 2.40 minutes of congestion to reduce late arrival by one minute, plus an extra 5.47 minutes of congestion to avoid 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 queuing behind a bottleneck. They also can be used to formulate models of traveler response to network unreliability, as we describe in Section 2.6.4.

The twelve alternatives in the choice model just described have a natural ordering in terms of chronological time; 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, there is no plausible combination of variables 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 earlier or later times, others still that that would affect the strength of preference for low travel times, and so forth. Such richness can be incorporated into the specification of a discrete-choice model based on random utility maximization, one of its great advantages.

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, State Route 91 (SR91), in southern California. 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 90^{th} and 50^{th} 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.

[36] Calculated as 0.065/0.106=0.61.

The model explains a pair of related decisions: (a) whether to acquire an electronic toll-collection transponder (required for any use of the express lanes), and (b) which lanes to take on the day in question. A natural way to view these decisions is as a hierarchical set, with transponder choice governed by the potential benefits of express-lane travel and other factors. 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. However, they obtain virtually identical results with a simpler "joint logit" model with three choice 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 model that Lam and Small estimate is:[37]

$$\begin{aligned} V = & \underset{(0.411)}{-0.862 \cdot D^{\text{tag}}} + \underset{(0.0058)}{0.0239 \cdot Inc \cdot D^{\text{tag}}} - \underset{(0.412)}{0.766 \cdot ForLang \cdot D^{\text{tag}}} - \underset{(0.853)}{0.789 \cdot D^3} \\ & - \underset{(0.138)}{0.357 \cdot c} - \underset{(0.056)}{0.109 \cdot T} - \underset{(0.048)}{0.159 \cdot R} + \underset{(0.046)}{0.074 \cdot Male \cdot R} + (\text{other terms}) \end{aligned} \quad . \qquad (2.31)$$

Here $D^{\text{tag}} \equiv D^2 + D^3$ is a composite of alternative-specific dummy variables 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 (recall that both variables are measured in units of time). Second, men seem to care less about reliability than women; their value is only 53 percent as high according

[37] 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.

to the point estimates,[38] although the difference (*i.e.* the coefficient of *Male·R*) is not quite statistically significant even at a 10-percent significance level.

This provides a good opportunity to consider how one chooses the list of variables to be included in a travel-demand model. In this case, the model is specified to yield a constant value of time and a constant value of reliability. Although theory suggests these values might vary by income or other factors, prior experimentation showed that including such variations (through interactions like those in Section 2.3.1) resulted in imprecise and ambiguous results. However, the authors did find that those same factors strongly affect the alternative-specific constants (as determined by interacting these variables with alternative-specific dummies, part of "other terms" in the results shown here). Furthermore, the hierarchical nature of the decision process, already mentioned, suggests that certain alternatives are likely to fall within groups subject to common influences. Specifically, one might assume that certain unobserved factors would influence the extent to which the requirement to get a "toll tag", specific to two of the alternatives, is regarded as onerous. For example, the financial set-up arrangements for a toll tag might be of lesser significance to people with high incomes and of greater significance for people who have difficulty communicating in English — precisely what is indicated by the signs of the second and third coefficients in equation (2.31). The additional alternative-specific constant D^3 is included on the grounds that yet other unmeasured factors could affect the actual decision to pay for the express lanes. The variable *Male·R* is included because several studies of this particular toll facility have found women noticeably more likely to use the express lanes than men. This could be accounted for with a simple interaction variable $Male \cdot D^3$, but experimentation shows that the formulation shown here fits better. It is more specific about why men use the express lanes less often: apparently, the reason is less aversion to the unreliability of the free lanes. To be specific, reliability enters women's utility with the negative coefficient -0.159, whereas it enters men's utility with the more weakly negative coefficient $(-0.159 + 0.074) = -0.085$. One could attempt to be even more specific: perhaps the gender difference is due to differing responsibility for children; the authors investigated this by

[38] The coefficient for women is –0.159, and that for men is –0.159+0.074 = –0.085.

including a variable equal to R multiplied by a dummy for women with children, but they were unsuccessful in pinpointing the effect in this manner.[39]

Equation (2.31) also makes it easy to define a quantity used often in demand and cost analysis: the *generalized price* of a particular type of travel.[40] The idea is that if a traveler considers cost and other aspects of travel, such as time and reliability, in fixed proportions, it may be analytically useful to combine them into a single index denominated in money. The conditional indirect utility function, V in this example, summarizes the relative weights the traveler puts on these various aspects of travel. Thus in (2.31), the generalized price p for females would be defined as

$$p = c + \frac{0.109}{0.357} T + \frac{0.159}{0.357} R \tag{2.32}$$

and the same for males except the coefficient of R would have numerator (0.159–0.074). It is not hard to see that the first ratio of coefficients is just the value of time defined earlier (perhaps in different units); and the second will be defined later as the value of reliability. The concept of generalized price is most straightforward if the price is the same for everyone; it is somewhat less straightforward in this example since there are two groups with two different generalized prices. It would be even more complicated in our mode-choice example of (2.29), where the value of time depends on the traveler's wage rate.

The generalized price could also incorporate more than one component of travel time, as in (2.29), and it could incorporate scheduling costs, as in (2.30). Note that (2.30) itself is insufficient to define a generalized price because it does not contain a cost variable. However, it gives the rates of tradeoff between schedule delay and time, and (2.31) gives the tradeoff between time and cost, so by combining the two studies we could extend (2.32) as follows:

$$p = c + \frac{0.109}{0.357} \cdot \left[T + \frac{0.065}{0.106} SDE + \frac{0.254}{0.106} SDL + \frac{0.58}{0.106} DL \right] + \frac{0.159}{0.357} R \,. \tag{2.33}$$

In Chapter 3, we will see that a generalized price based on those parts of (2.33) involving c, T, SDE, and SDL, has been used extensively to analyze equilibria with travelers individually choosing their trip schedules in response to time-varying congestion.

---

[39] See Small, Winston and Yan (2005) for further discussion of this same issue in a different data set from SR91.

[40] Generalized price is also sometimes known as *generalized cost*; but we avoid this term because in our terminology, price includes taxes and tolls, if levied, while cost does not.

2.4 Advanced Discrete-Choice Modeling

2.4.1 Generalized Extreme Value Models

Often it is implausible that the additive random utility components ε_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 have many unmeasured attributes in common, such as occasional crowding. 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 rail service (j=3) with measurable characteristics identical to those for bus. The iid models would predict that all three modes would then have choice probabilities of one-third; 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 (2.11) 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.[41]

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

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

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

[41] The variance-covariance matrix of these utility differences has $(J–1)^2$ elements and is symmetric. Hence it has only $J(J–1)/2$ identifiable elements, less one for utility-scale normalization.

$$P_i = \frac{e^{V_i} \cdot G_i(e^{V_1},...,e^{V_J})}{G(e^{V_1},...,e^{V_J})} \tag{2.34}$$

where G_i is the i-th partial derivative of G. (The technical conditions assure that these probabilities add to one.) The expected maximum utility is

$$E \max_j (V_j + \varepsilon_j) = \log G(e^{V_1},...,e^{V_J}) + \gamma \tag{2.35}$$

where again γ is Euler's constant.[42] Logit is the special case $G(y_1,...,y_J) = y_1+...+y_J$.

The best known GEV model, other than logit itself, is *nested logit*, also called *structured logit* or *tree logit* 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:

$$G(y_1, y_2, y_3) = y_1 + \left(y_2^{1/\rho} + y_3^{1/\rho}\right)^{\rho}. \tag{2.36}$$

In this equation, ρ 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 ε_1 and ε_2 (Daganzo and Kusnic 1993). The choice probability for this example, computed from (2.34), may be written:

$$P_i = P(B_{r(i)}) \cdot P(i \mid B_r) \tag{2.37}$$

$$P(B_r) = \frac{\exp(\rho \cdot I_r)}{\sum_{s=1}^{2} \exp(\rho \cdot I_s)} \tag{2.38}$$

$$P(i \mid B_r) = \frac{\exp(V_i / \rho)}{\sum_{j \in B_r} \exp(V_j / \rho)} \tag{2.39}$$

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 (2.39):

[42] This is demonstrated by Lindberg, Eriksson and Mattsson (1995, p. 134). For a simpler proof, see Choi and Moon (1997, p. 131).

$$I_r = \log \sum_{j \in B_r} \exp(V_j / \rho) . \tag{2.40}$$

When ρ=1 in this model, ε_2 and ε_3 are independent and we have the logit model. As $\rho \downarrow 0$, ε_2 and ε_3 become perfectly correlated and we have an extreme form of the "red bus, blue bus" example, in which auto is pitted against the better (as measured by V_i) of the two transit alternatives; in this case $\rho I_1 = V_1$ 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 ρ_r in equations (2.36)-(2.40), leading to the form:

$$G(y_1,..., y_J) = \sum_r \left(\sum_{j \in B_r} y_j^{1/\rho_r} \right)^{\rho_r} . \tag{2.41}$$

This is the general two-level nested logit model. It has choice probabilities (2.37)-(2.40) except that the index s in the denominator of (2.38) 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 (2.38) a natural interpretation as a choice among groups, taking the logit form with I_r playing the role of the independent variable and ρ_r its coefficient. Thus, for example, in a destination-choice model the inclusive value of a set of shopping destinations (from a given origin) can serve as a measure of accessibility of that origin to shopping (Ben-Akiva and Lerman 1979). The welfare measure for the two-level nested logit model is, from (2.35), (2.40), and (2.41):

$$W = \frac{1}{\lambda} \log \sum_r \exp(\rho_r \cdot I_r) \tag{2.42}$$

where again λ is the marginal utility of income and the constant γ is omitted.

In nested logit, $\{B_r\}$ is an exhaustive partition of the choice set into mutually exclusive subsets. Therefore equation (2.39) is a true conditional probability, and the model can be estimated sequentially: first estimate the parameters (β/ρ) from (2.39), use them to form the

inclusive values (2.40), then estimate ρ from (2.38). 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. Several studies show that maximum-likelihood estimation, although computationally more difficult, gives more accurate results (Hensher 1986, Brownstone and Small 1989).[43]

Most other GEV models that have been studied generalize (2.41) 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. Vovsha (1997) defines a *cross-nested logit* model with nests that can overlap in a general way and with alternatives that can belong to a nest in a partial sense, with weights to be estimated. Chu (1981) and Koppelman and Wen (2000) define a model in which the subsets B_r include all possible pairs of alternatives:

$$G(y_1,...,y_J) = \sum_{j=1}^{J-1} \sum_{k=j+1}^{J} \left(y_j^{1/\rho_{jk}} + y_k^{1/\rho_{jk}} \right)^{\rho_{jk}} . \tag{2.43}$$

This model, known as *paired combinatorial logit*, has error terms whose variance-covariance matrix contains the same number of estimable parameters as does multinomial probit with a general covariance structure (again, the arbitrary scale requires one more normalization); thus the two models can be expected to have comparable degrees of generality, although in practice paired combinatorial logit seems to be easier to estimate.

Nevertheless, estimation of GEV models is often difficult because of the highly nonlinear manner in which ρ_r 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 (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

[43] 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).

GEV class, and its probabilities cannot be written in closed form so require numerical integration.[44]

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 is closely intertwined with its choice of how much to drive. One can formulate an equation to explain usage, conditional on ownership, but it is subject to *sample selection bias* (Heckman 1979). To illustrate, suppose that people who drive a lot tend to select themselves into the category of owners of nice cars; the conditional model would overstate the independent effect of nice cars on driving by ignoring this reverse causality (a form of endogeneity). A variety of methods are available to remove this bias.[45]

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 Q, income Y, and other explanatory variables X as follows:

$$x = \beta_X X + \beta_Q Q + \beta_Y Y + u \tag{2.44}$$

where 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):

$$Q = d_1 Q_1 + (1 - d_1) Q_2 . \tag{2.45}$$

Substituting (2.45) into (2.44) makes explicit the dependence of the usage decision (x) on the ownership decision (d_1).

Suppose also that the ownership decision (which type of car to own) depends on some set of observable variables Z, which might include Y, some elements of X, and/or the quality difference (Q_1-Q_2):

---

[44] For a review of these and other GEV models, see Koppelman and Sethi (2000).

[45] See Train (1986, Ch. 5), Mannering and Hensher (1987), and Washington, Karlaftis and Mannering (2003, Ch. 12).

$$\begin{aligned} &d_1 = 1 \text{ if } U_1 > U_2, \quad 0 \text{ otherwise;} \\ &U_1 - U_2 = \beta_Z' Z + \varepsilon . \end{aligned} \tag{2.46}$$

This equation defines a binary probit model if ε is assumed normal, binary logit if ε is assumed logistic.

Selection bias is present if u and ε 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 ε, it is also correlated with the car-type indicator d_1, via (2.46), and therefore with car quality Q, via (2.45). This biases the estimated coefficients in (2.44) — especially that of β_Q, — because Q is an explanatory variable for usage and thus needs to be uncorrelated with the error term for usage.

If we can find an exogenous proxy for Q, we can use it instead of Q in estimating (2.44) and solve the problem. This can be accomplished using the following two-step procedure proposed by Heckman (1979).

Step 1 consists of estimating (2.46). Its explanatory variables are presumed exogenous so there is no bias at this stage. 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 (2.44) 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 (2.45). 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 X, Y, and Z. (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 X, Y, and Z.

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

Step 2 Version (b): Add a "correction term" to the error term in (2.44) 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 (2.44) would be appropriate if (2.44) 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 (2.46), from u; the remaining error term is independent of Q and so (2.44) is purged of selectivity bias. A recent application of this technique in transportation is West's (2004) model of automobile type choice and amount of use.

The conditional expectation just mentioned can be computed explicitly for binary probit and logit models.[46] We write the result as a term γC to be added to (2.44), where γ is a parameter to be estimated and C is a "correction variable" computed from the results of Step 1. The correction variable is included in an ordinary regression as though it were a real variable, and its coefficient is an estimate of γ. The estimated value of γ will be proportional to the correlation between u and ε, which we denote by ρ. It is this correlation that causes the problem, so we can test for selection bias by testing whether γ is different from zero. Furthermore, error term u may be regarded as consisting of γC plus a new error that is uncorrelated with Q; hence the other coefficients in the model are now estimated without bias.

Table 2.1 gives formulas for the correction variable $C = d_1C_1+(1–d_1)C_2$; it also shows how parameter γ is related to correlation ρ. In this table, ϕ denotes the probability density

[46] 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 ε_j, $j \neq i$). However they do not discuss how to pool the data with observations of individuals who choose other alternatives.

function of a standard normal random variable, σ_u is the standard deviation of u, and $\hat{P}_2 = 1 - \hat{P}_1$. Sometimes data are lacking on people 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.[47]

Table 2.1. Selectivity Correction Terms $\gamma \cdot (d^1 C_1 + d^2 C_2)$ for (2.44)

Model	*Correction Variable*		*Coefficient*
	C_1	C_2	γ
Probit	$\dfrac{\phi(\hat{\beta}_Z Z)}{\hat{P}_1}$	$-\dfrac{\phi(\hat{\beta}_Z Z)}{\hat{P}_2}$	$\rho\sigma_u$
Logit	$-\left[\dfrac{\hat{P}_2 \ln \hat{P}_2}{1-\hat{P}_2} + \ln \hat{P}_1\right]$	$\left[\dfrac{\hat{P}_1 \ln \hat{P}_1}{1-\hat{P}_1} + \ln \hat{P}_2\right]$	$\left(\sqrt{6}/\pi\right) \cdot \rho\sigma_u$

What this procedure does is add correction γC_1 for those individuals in the sample who chose a new car and γC_2 for the others. Note that in each row, C_1 is positive and C_2 is negative. Thus if ρ is positive, indicating that people choosing new cars are likely to drive more for unobserved reasons, the adjustment γC is positive for those individuals who choose new cars and negative for those who choose used cars — exactly the pattern we want for γC to replace the part of error term u that is correlated with Q.

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

[47] 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 Φ^{-1} denotes the inverse of the standard normal cumulative distribution function.

When observations from individual respondents are collected repeatedly over time, the set of respondents is called a *panel* and the information on them is called *panel data* or *longitudinal data*. A good example is the Dutch National 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 cross-sectional time series are applicable to disaggregate panel 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.[48]

2.4.4 Random Parameters and Mixed Logit

In the random utility model of (2.10)-(2.11), 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 (2.19) 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 (2.10). 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 (2.11) within another integral over the density function of that distribution. Subsequently, this simple idea was

[48] See Kitamura (2000) for a general review and Pendyala and Kitamura (1997) or Brownstone and Chu (1997) specifically on attrition.

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 ε_{jn}) is unnecessary because, conditional on the values of random parameters, it yields the logit formula (2.14). 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 not difficult to write out. Using the logit formulation of (2.14) and (2.16), the choice probability conditional on random parameters is

$$P_{in|\beta} = \frac{\exp(\beta' z_{in})}{\sum_j \exp(\beta' z_{jn})}. \tag{2.47}$$

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

$$P_{in} = \int P_{in|\beta} \cdot f(\beta \mid \Theta) d\beta . \tag{2.48}$$

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

$$P_{in}^{sim} = (1/R)\sum_{r=1}^{R} P_{in|\beta}^{r} . \tag{2.49}$$

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

Brownstone and Train (1999) demonstrate how one can shape the model to capture anticipated patterns by specifying which parameters are random and what form their distribution takes — in particular, whether some of them are correlated with each other.[49] In their application,

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

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:

$$V = -0.264\cdot[p/\log(inc)] + 0.517\cdot range + (1.43+7.45\phi_1)\cdot size + (1.70+5.99\phi_2)\cdot luggage + 2.46\phi_3\cdot nonE + 1.07\phi_4\cdot nonN + (other\ terms)$$

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

This model provides for observed heterogeneity in the effect of price p on utility, since $(\partial V/\partial p)$ varies with *inc*. It provides for random coefficients on *size* and *luggage*, and for random constants as defined by *nonE* and *nonN*. This can be understood by examining the results term by term.

The terms in parentheses involving ϕ_1 and ϕ_2 represent the random coefficients. The coefficient of *size* is random with mean 1.43 and standard deviation 7.45. Similarly, the coefficient of *luggage* has mean 1.70 and standard deviation 5.99. These estimates indicate a wide variation in people's evaluation of these characteristics. For example, it implies that many people 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 ϕ_3 and ϕ_4 represent random alternative-specific constants with a particular correlation pattern, predicated on the assumption that groups of alternatives share common features for which people have idiosyncratic preferences — very similar to the rationale for nested logit. Each of the dummy variables *nonE* and *nonN* is simply a sum of alternative-specific constants for those car models falling into a particular group. The two groups overlap: any gasoline-powered or methanol-powered car falls into both. If the coefficients of ϕ_3 and ϕ_4 had

turned out to be negligible, then these terms would play no role and we would have the usual logit probability conditional on the values of ϕ_1 and ϕ_2. But the coefficients are not negligible, so each produces a correlation among utilities for alternatives within 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 ε_{in} in (2.10). Thus the combined additive random term in utility,[50] $(\varepsilon_{in}+2.46\phi_{3n}\cdot nonE_i+1.07\phi_{4n}\cdot nonN_i)$, exhibits correlation across those alternatives *i* representing cars that are not electric and, by similar argument involving ϕ_4, across those alternatives representing cars that are not natural gas. The two alternatives falling into *both* the *nonE* and *nonN* groups, namely alternatives G and M, are even more highly correlated with each other. Note that because the distributions of ϕ_3 and ϕ_4 are centered at zero, this combined random term does not imply any overall average preference for or against various types of vehicles; such absolute preferences are in fact included in *other terms*.

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

[50] As Brownstone and Train (1999) point out, the terms in ϕ_1 and ϕ_2 also may 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.

because one needs to simulate not only the explicit integral in (2.48) but also the integral that, for probit, is part of the definition of the conditional choice probability $P_{in|\beta}$.

2.4.5 Endogenous Prices

We have already mentioned the biases that can be introduced by including, as an explanatory variable, one that is not really independent of what is being explained. For example, the level of automobile ownership does a great job of "explaining" mode choice, but is it really an independent factor or does it, at least partly, respond to the mode choice decision? If the latter, automobile ownership is *endogenous* to mode choice; including it to explain mode choice then overstates its effect and also biases the coefficients of other variables that are correlated with it. This is called *endogeneity bias*.

There is a more subtle source of endogeneity bias that can occur with a variable measuring the price of an alternative. This has long been recognized as a problem in aggregate demand studies, where the demand curve is understood to be just one side of a simultaneous system determining price and quantity. But it can also afflict disaggregate studies if price is determined in a market setting where unobserved quality attributes are important and if those unobserved attributes are not controlled for in the model.

For example, Train and Winston (2007) seek to explain individual consumers' choices among 200 different makes and models of cars. Naturally they want to include the price of the car as a variable. The trouble is, manufacturers' pricing decisions take account of the car's quality, including some aspects of quality that the investigators are unable to measure. If some makes and models have higher unmeasured quality than others, and their prices reflect this, it can appear as though consumers are drawn to high prices whereas they are really drawn to high unobserved quality. Thus, the strength of the usual negative effect of price on choice probability will tend to be underestimated or even not discerned at all.

Unobserved quality attributes are not a problem if the model includes a full set of alternative-specific dummy variables, because then the attributes are accounted for by the coefficients of those dummy variables (*i.e.*, by the alternative-specific constants). But that solution is not always satisfactory for at least two reasons. First, the data set might be too small to estimate all the alternative-specific constants, especially when there are many of them; for example, if it happens that some alternative is not chosen by anyone in the sample, its

alternative-specific constant cannot be estimated. Second, including all alternative-specific constants makes it impossible to include any other variable that varies across alternatives but not across sample members.[51] Thus, for example, if for any given product all sample members face the same price, the price alone cannot be included as an explanatory variable along with a full set of alternative-specific dummies; price can only be included by interacting it with some characteristic of consumers such as income, so that it varies in some manner different from what can be absorbed into the alternative-specific constants. Thus the alternative-specific constants, if included in the model, may hide important information.

A solution to this problem is available whether or not alternative-specific constants have been estimated. The key is to either estimate or calibrate alternative-specific constants and then regress them on price (and any other variables that vary only by alternative) using an *instrumental-variables estimator* — a procedure noted briefly in Section 2.1.6. If the alternative-specific constants are not estimated statistically, one can obtain them using the calibration procedure described earlier for updating a model to a new location (Section 2.2.7). That procedure finds the values that equate the choice shares predicted by the discrete-choice model to the observed choice shares. (Presumably the latter are based on large enough samples so that they are all strictly positive.)[52]

We now describe the instrumental-variables estimator more fully, using the example of an endogenous price variable in Train and Winston (2007). The instrumental variables are any set of variables that do a good job of predicting price but are independent of unmeasured quality attributes that might be correlated with price. The instrumental-variables estimator effectively replaces "price" in the regression equation by its predicted value based on all exogenous variables in the model, including the instrumental variables. In this way price is purged of its endogeneity but otherwise still carries information relevant to understanding consumer choice.

[51] Formally, this is because the variable would be perfectly collinear with the alternative-specific dummies: that is, some linear combination of them adds to zero. The inability to include perfectly collinear variables is a well-known limitation on any regression model. Intuitively, the problem is that two different variables (or combinations of variables) would be competing to explain the same dimension of variation in the data, and so their separate effects could not be distinguished.

[52] An alternative solution is proposed by Petrin and Train (2004): adding artificial variables to the model to absorb the part of error term that is correlated with price, an extension of the approach of Heckman described as "Step 2 Version (b)" in Section 2.4.2.

The procedure works only in a linear model, which is why it cannot be applied directly to the discrete-choice model itself.

The full Train-Winston procedure, then, looks like this. The conditional indirect utility function V_{in} is decomposed two parts, one varying across consumers and the other not:

$$V_{in} = \beta' z_{in} + \left(\gamma' x_i + \xi_i\right) \tag{2.50}$$

where z_{in} and x_i are vectors of explanatory variables, the latter varying only across alternatives, not individuals. The first part of this decomposition, $\beta' z_{in}$, has exactly the same meaning as in the mixed logit model of (2.47), including the fact that β can vary randomly across consumers. The second part, which we write as

$$\delta_i = \gamma' x_i + \xi_i \tag{2.51}$$

varies only across products and has a constant parameter vector γ. Price is included among variables x; it may also be included in z if it is interacted with a consumer characteristic such as income.

The model is estimated iteratively. Starting with a guess or a preliminary estimate of constants $\{\delta_i^{old}\}$, β is estimated using the mixed-logit estimator with utility (2.50), holding the part in parentheses (δ_i) constant. Predicted choice shares $\hat{s}_i$ (conditional on $\{\delta_i^{old}\}$) are then obtained by aggregating the mixed-logit probabilities (2.49) over the sample. Finally, using the observed choice shares s_i, δ_i is updated according to the rule described in a footnote to Section 2.2.7:

$$\delta_i^{new} = \delta_i^{old} + \log(s_i) - \log(\hat{s}_i) \ .$$

These steps are iterated until they converge.[53] Finally, γ is estimated by regressing the resulting values of $\{\delta_i\}$ on x_i according to (2.51), using instrumental variables.

What sort of variable would make a good instrument for price? The value of such a variable, for a given make/model of automobile, needs to be reasonably well correlated with the price of that make/model but not with its unmeasured characteristics. A key insight of Berry (1994) is that a good instrument can be constructed as an index of the measured characteristics of *other* makes and models. This is because automobile prices are presumed to be determined in an oligopoly setting, in which each firm sets product prices while taking account of all the other

[53] Proof of convergence is given by Berry, Levinsohn, and Pakes (1995, Appendix I).

products in the market, including its own. Hence price is influenced by the instrumental variable, as required. Furthermore, because there are many other makes and models, it is reasonable to assume that an index of such characteristics is not significantly influenced by the price of the make/model under consideration, thus meeting the other requirement for an instrument.[54]

Fortunately, this procedure is unnecessary for most problems in urban transportation because either the model contains alternative-specific constants or prices can be considered exogenous (for example they may be set by a public agency). The procedure has been used mainly to study the demand for automobiles, computers, and other products characterized by significant and measurable product variety.[55]

2.5 Activity Patterns and Trip Chaining

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. For example, early theories based on shopping strategies in the face of multiple products and storage costs yielded rich insights but no practical predictive models, whereas *ad hoc* empirical models were theoretically unsatisfactory and also not very accurate (Thill and Thomas 1987). Nevertheless, important progress has been made. Descriptive information on activity patterns is now extensive, often showing surprising similarities across nations.[56] Surveys now elicit multi-day diaries describing all activities and travel undertaken during a period of time. And newer theoretical work more fully integrates shopping and trip chaining into conventional consumption theory (Anas 2007). One intriguing result: making it easier to chain trips together may result in more destinations visited rather than less travel.

---

[54] Specifically Train and Winston (2007) create four instrumental variables. The first two are the sums of differences in measured characteristics between the make/model in question and (1) all other makes/models by the same manufacturer, or (2) all other makes/models by other manufacturers. The other two are formed the same way but using the *squares* of differences in characteristics.

[55] The procedure was first developed for aggregate data, using demand functions built up from disaggregate discrete-choice decisions by Berry (1994); Berry, Levinsohn and Pakes (1995); and Bresnahan, Stern and Trajtenberg.(1997). In those cases the mixed-logit part of the algorithm is simplified because consumers, not being observed individually, must be treated as observationally identical. Applications to disaggregate data include Berry, Levinsohn and Pakes (2004) for automobiles and Goolsbee and Petrin (2004) for television reception.

[56] For example, (Timmermans *et al.* 2002) compare data on Japan, Canada, The Netherlands, the US, and UK.

As for models that are fully activity-based, two main classes have emerged.[57] One consists of econometric models that 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. The other consists of simulation models that enumerate feasible combinations of activities, based on logical constraints relating activities at different locations to travel between those locations.

Turning to the first class of models, 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 with this is that it leads to enormous numbers of possible choice 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.

Fundamental to describing trips are the locations and the starting and ending times of activities that the trips are intended to connect. Yet few if any formal activity models, whether of the econometric or simulation variety, have been able to satisfactorily account for the varying degrees of flexibility in locations and times of day. Moreover, to fully understand the processes generating travel, one needs to model the substitution between in-home and out-of-home activities, adding further to the sheer number of possibilities to consider.

As an example of what can be accomplished, 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, just as one would expect. But the policy *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. This result is consistent with several studies of telecommuting, which

[57] See Ettema and Timermans (1997), Ben-Akiva and Bowman (1998), or Bhat and Koppelman (2003) for reviews. A collection of recent research appears in Miller (2005).

have found only a very small net reduction in travel; indeed many types of telecommunication appear to be complements to, rather than substitutes for, travel.[58]

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 on 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 main reason these quantities are so important is that they account for a large portion of the benefits (positive or negative) of changes in the transportation environment. For this reason, analysts are generally not satisfied to merely have these benefits captured implicitly as part of consumer surplus as, for example, in equation (2.17); rather, they would like to separate them explicitly in order to illuminate the nature of the changes being considered. In this section, we consider the theory behind consumers' response to time and reliability and our empirical knowledge about those responses; in Chapter 5, we discuss how that knowledge is used in the evaluation of projects or policies.

2.6.1 Value of Time: Basic Theory

The most natural definition of value of time is in terms of "compensating variation" (Varian 1992). 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."[59]

[58] Choo, Mokhtarian and Salomon (2005), Plaut (1997).

[59] 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. These claims are based on the idea that travelers judge changes in the transportation system relative to some reference point, taken to be the situation before a proposed change. But

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 it can 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 maximized subject to several constraints. First, the usual budget constraint requires that expenditures are no greater than the sum of 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 be within total time available, $\overline{T}$. Finally, certain activities (such as travel) have technological features (such as maximum speeds) that impose a minimum $\overline{T}_k$ on time T_k spent in activity k. (We will consider later an extension where T_w is also so constrained.) We assume that the first two constraints are binding, so they can be expressed as equalities.

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

---

that idea ignores the fact that travel patterns and other factors affecting travel are in constant flux. This flux means that within a few months of a change, most travelers will no longer view the "before" situation, as defined by an analyst, as a reference point. It is more plausible that travelers will adjust their behavior to their needs at any point in time, subject to the situation that faces them (as measured by current conditions) and not to some arbitrary previous situation. Studies based on consistent definitions have generally not found such dependencies (MVA Consultancy *et al*. 1987, pp. 65-68), and theory refutes the alleged rationale for them (Mackie, Jara-Díaz and Fowkes 2001).

$$\Lambda = U\left(G, T_w, \{T_k\}\right) + \lambda \cdot \left[Y + wT_w - G\right] + \mu \cdot \left[\overline{T} - T_w - \sum_k T_k\right] + \sum_k \phi_k \cdot \left[T_k - \overline{T}_k\right], \quad (2.52)$$

where λ, μ, and $\{\phi_k\}$ are Lagrange multipliers that indicate how tightly each of the corresponding constraints limits utility. The first-order condition for maximizing (2.52) with respect to one activity time T_k is

$$U_{Tk} - \mu + \phi_k = 0 \quad (2.53)$$

while that with respect to T_w is

$$U_{Tw} + \lambda \cdot \left[w + T_w \cdot (dw / dT_w)\right] - \mu = 0\,, \quad (2.54)$$

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, $\overline{T}$, wage schedule $w(T_w)$, and minimum activity times $\{\overline{T}_k\}$. The rate at which utility increases as the k-th minimum-time constraint is relaxed is given by its Lagrange multiplier, ϕ_k; the increase of utility with respect to unearned income is λ. Hence the marginal value of time for the k-th time component is their ratio:

$$v_T^k \equiv \left(\frac{\partial Y}{\partial \overline{T}_k}\right)_V = -\frac{\partial V / \partial \overline{T}_k}{\partial V / \partial Y} = \frac{\phi_k}{\lambda}\,. \quad (2.55)$$

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

Equations (2.53)-(2.55) imply that for a travel activity k,

$$v_T^k = \frac{\mu}{\lambda} - \frac{U_{Tk}}{\lambda} = w + T_w \cdot \frac{dw}{dT_w} + \frac{U_{Tw}}{\lambda} - \frac{U_{Tk}}{\lambda}\,. \quad (2.56)$$

This equation decomposes the value of travel-time savings into the opportunity cost of time that could be used for work, μ/λ, less the value of the marginal utility of time spent in travel. The opportunity cost μ/λ is both pecuniary (the first two terms after the last equality) and non-pecuniary (the third term, which could be positive or negative).

Most of the theoretical literature assumes that the wage rate is fixed, in which case equation (2.56) 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 v_T^k to rise with total trip time because the total time constraint in (2.52) will bind more tightly, causing μ, the marginal utility of leisure, to rise.[60]

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 δ_k is one if travel activity k is chosen and zero otherwise. The indirect utility function has the same derivatives with respect to exogenous variables c_k and $\overline{T}_k$ as does the Lagrangian function, which under the assumptions just stated can be written as:

$$\frac{\partial V}{\partial c_k} = -\lambda\delta_k ; \qquad \frac{\partial V}{\partial \overline{T}_k} = -\phi_k\delta_k .$$

Equivalently, writing V_k as the conditional indirect utility function:

$$\frac{\partial V_k}{\partial c_k} = -\lambda ; \qquad \frac{\partial V_k}{\partial \overline{T}_k} = -\phi_k . \tag{2.57}$$

Then our definition of value of time in (2.55) is identical to that in (2.19):

$$v_T^k \equiv \frac{\phi_k}{\lambda} = \frac{\partial V_k / \partial \overline{T}_k}{\partial V_k / \partial c_k} .$$

Note also that the first of equations (2.57) is identical to (2.24) since we assume just one trip per time period. (It is easy to generalize to allow for an endogenously chosen number of trips per time period.)

[60] According to (2.54), this could happen by the consumer adjusting so as to raise the wage rate w and/or the marginal enjoyment of work U_{Tw}. This reflects the idea that if time is scarce, one is more choosy about the kind of job one will accept. Academics with consulting opportunities will often say that when they are exceptionally busy, they only accept consulting jobs with very high wages or that they find especially interesting, in either case helping them meet the marginal condition (2.54). Similarly, some people busy with small children will take on part-time work only if it pays well or is especially enjoyable.

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 the right-hand side of (2.56) disappears. Then (2.56) implies that the value of time is a fraction of the wage rate, as for example with specification (2.15) with β_3=0. Alternatively, one might think that work enjoyment varies nonlinearly with the observed wage rate: perhaps negatively due to wage differentials that compensate for working conditions, or perhaps positively due to employers' responses to an income-elastic demand for job amenities. Then (2.56) implies that value of time is a nonlinear function of the wage rate, which could suggest using (2.15) with a non-zero term β_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 (2.52) can lead to operational specifications for the conditional indirect utility function.

2.6.3 Extensions

Human behavior is complex, and many additional factors may affect how people allocate their time. No analytical model can account for all of them, but we can mention several interesting extensions to the theory just presented. They account for constraints on work hours, variable commuting times, technologies of home production, and psychological bias against losses from the status quo.

First, consider work-hour constraints. People cannot always change the amount of time they spend at work, perhaps because they are locked into a particular job with fixed hours. To some extent this is handled by allowing w to depend on T_w; but we could also consider a stricter constraint that T_w be fixed, say at $\overline{T}_w$. This adds a term $\phi_w \cdot \left[T_w - \overline{T}_w\right]$ to (2.52), where ϕ_w is another Lagrangian multiplier whose sign indicates whether this person would prefer to work fewer (ϕ_w>0) or more (ϕ_w<0) hours. This modification adds a term ϕ_w/λ to the value of time as given by (2.56), thus raising or lowering the value of time depending on the sign of ϕ_w. Indeed, MVA Consultancy *et al.* (1987, pp. 149-150) find that people who are required to work extra hours at short notice have 15-20 percent higher values of travel time than other workers, suggesting that for them such a model may apply with ϕ_w>0.

Second, consider the situation where the amount of time devoted to consuming goods is proportional to the amount of goods— as for example occurs if the "good" is watching a movie. For simplicity, let's label this as activity 0 and consider it a generalized leisure activity, whose consumption time is proportional to all goods consumption G. Then the term $\phi_0 \cdot \left[T_0 - \overline{T}_0\right]$ in (2.52) must be changed to $\phi_0 \cdot \left[T_0 - \ell \cdot G\right]$, where ℓ is the unit time requirement for consumption. The constraint is binding if $\phi_0>0$. While this modification does not alter the formulas derived for value of time, it does change the meaning of λ (the marginal utility of income) in those formulas, as shown by Jara-Díaz (2000, 2003). Previously, the solution required that $\lambda=U_G$, the marginal utility of consumption, as is easily seen from the first-order condition for maximizing (2.52) with respect to G. But now that first-order condition implies $\lambda=U_G-\phi_0\cdot\ell$. Since λ 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. Consumers are pressed for time in all their activities because of the time needed for ordinary consumption.

Third, suppose that working more hours requires spending more time commuting. This would be the case, for example, if it requires a secondary worker entering the work force or a part-time worker increasing the number of days worked. Following De Borger and Van Dender (2003), suppose commuting is activity c and commuting time is proportional to the amount of time worked: $T_c = a \cdot T_w$ for fixed parameter a. Substituting this equality into the overall time constraint in (2.52), the value μ in (2.54) becomes multiplied by (1+a) and all terms on the right-hand side of (2.56) are therefore divided by (1+a). Thus greater commuting time (larger a) causes the value of time to *decrease* — opposite to the effect noted earlier from a rising marginal utility of leisure. One way to understand this is that a larger value of a reduces the effective hourly wage rate, *i.e.* the wage rate net of commuting cost. De Borger and Van Dender suggest in numerical simulations that the effect can be quite large.

Finally, consider the well-known phenomenon of loss aversion. People often display asymmetric preferences regarding gains or losses from the status quo (more generally, from a reference situation), assigning far more weight to losses than to gains. Kahneman and Tversky (1979) develop a general theory of such behavior known as prospect theory. This behavior may create a large gap between "willingness to pay" (WTP) for a time savings and "willingness to accept" (WTA) for an identical time increase, each being defined as the amount of money paid

or received so as to leave the traveler indifferent to the change. De Borger and Fosgerau (2006) apply the following "reference-dependent" utility function, adapted from the literature on prospect theory:

$$V(c,t) = V^c(-c) + V^t(-v_T t)$$

where c and t are deviations in cost and time, respectively, from the reference situation, v_T is a fixed parameter for a given traveler known as the "reference-free value of time," and the "value function" $V^i(\cdot)$ is an increasing function of its argument with $V^i(0)=0$, $i=c,t$. Loss aversion is represented by assuming that V^i slopes downward from zero more steeply than it slopes upward from zero: $V^i(x)<-V^i(-x)$ for $x>0$. Specifically, De Borger and Fosgerau assume the following value functions postulated by Tversky and Kahneman (1991):

$$V^i(x) = S(x)e^{-\eta_i S(x)}|x|$$

where $S(x)\equiv x/|x|$ is the sign of x, with $S(0)=0$, and η^c and η^t are fixed parameters. We have loss aversion in V^i if $\eta^i>0$. We define willingness to pay $WTP(t)$ as the amount the consumer would pay and remain indifferent in light of a time saving of $t>0$; *i.e.* it is the solution to

$$V^c(-WTP)+V^t(v_T t)=0.$$

Similarly $WTA(t)$ is the amount the consumer would have to receive to be willing to accept a time increase of t:

$$V^c(WTA)+V^t(-v_T t)=0.$$

These definitions imply that $WTP(t) = v_T e^{-(\eta^c+\eta^t)\cdot t}$ and $WTA(t) = v_T e^{(\eta^c+\eta^t)\cdot t}$, again for $t>0$. If there is loss aversion ($\eta>0$), we immediately see that $WTP<WTA$ and both WTP and WTA differ from $v_T t$. In terms of our previous definitions, the marginal value of time $(\partial V/\partial t)/(\partial V/\partial c)$ falls short of v_T for a time saving and it rises above v_T for a time increase. Furthermore, if values of (WTP/t) and (WTA/t) can be measured from experimental data, then v_T can be determined simply as their geometric mean. De Borger and Fosgerau, using a considerably more general model than this one, estimate values of $\eta^c=0.24$ and $\eta^t=0.50$ from stated preference data from over 2000 individuals, implying that WTA exceeds WTP by a factor of four.

We caution that the kind of loss aversion applied to an individual, in a hypothetical situation with a very clear reference scenario (a recent actual trip), need not apply to a proposed change to a transportation system affecting thousands of people in varying and changing circumstances. If you improve a traffic signal and wait a few months, many users' situations will

have changed since before the improvement and their reference situations are unclear; indeed, other events simultaneously affecting their travel might make them quite unaware of the signal improvement, even if they pay close attention to the overall time of their commute. Therefore it is not appropriate to use distinct values for *WTP* and *WTA* in most welfare analyses of public policies. Rather, we view the model just described as useful for interpreting stated-preference results. Nevertheless it illustrates that the standard assumption of rational behavior need not hold literally in all situations, and sometimes this must be taken into account in empirical work.

Other theoretical extensions to the theory of value of time include those showing how it depends on tax rates (Forsyth 1980) and on scheduling considerations (Small 1982).

2.6.4 Value of Reliability: Theory

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.[61] 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 trip-scheduling choice presented in equation (2.30). 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:

$$C(t_d, T_r) = \alpha \cdot T + \beta \cdot SDE + \gamma \cdot SDL + \theta \cdot DL \tag{2.58}$$

where $\alpha \equiv v_T/60$ is the per-minute value of travel time, β and γ are per-minute costs of early and late arrival, and θ is a fixed cost of arriving late. Travel time T is disaggregated into a value T_f that represents the lowest possible travel time for this particular trip, plus a random (unpredictable) component of travel time, $T_r \geq 0$. Since T_f is simply a given for the traveler, we use the functional notation $C(t_d,T_r)$ to focus attention on two variables: departure time t_d and stochastic delay T_r. Specifically, our objective is to measure the increase in expected cost C due

[61] See for example MVA Consultancy *et al.* (1987), p. 149; Small, Noland, Chu and Lewis (1999); and Hensher (2001).

to the dispersion in T_r, given that t_d is subject to choice by the traveler. Letting C^* denote this expected cost after the user chooses t_d optimally, we have

$$C^* = \underset{t_d}{Min}\, E\left[C(t_d, t_r)\right] = \underset{t_d}{Min}\left[\alpha \cdot E(T) + \beta \cdot E(SDE) + \gamma \cdot E(SDL) + \theta \cdot P_L\right] \tag{2.59}$$

where E denotes an expected value taken over the distribution of T_r, conditional on t_d; and where $P_L \equiv E(DL)$ is the probability of being late, again conditional on t_d. This equation can form the basis for specifying the reliability term in a model like (2.31). 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 T_f and hence E(T) are independent of departure time. To find the optimal departure time, let $f(T_r)$ be the probability density function for T_r and let t^* be the preferred arrival time at the destination. The next to last term in the square brackets of (2.59) can then be written as

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

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

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

where P_L^* is the probability of being late given the optimal departure time.[62] Similarly, differentiating the term involving β in (2.59) yields $-\beta \cdot \left(1 - P_L^*\right)$. Finally, differentiating the last term yields $-\theta f^0$ where $f^0 \equiv f(\tilde{t} - t_d^*)$ is the probability density at the point where the traveler is neither early nor late. What these three derivatives tell us is that departing later will lower the expected cost of early arrival but raise the expected costs of late arrival (involving γ and θ). Combining all three terms and setting them equal to zero gives the first-order condition for optimal departure time:

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

$$P_L^* = \frac{\beta + \theta f^0}{\beta + \gamma} . \tag{2.60}$$

In the special case θ=0, equation (2.60) yields the very intuitive rule $P_L^* = \beta/(\beta+\gamma)$, noted by Bates *et al.* (2001, p. 202).

Equation (2.60) is only an implicit equation for the optimal departure time, t_d^*, because both P_L and f^0 depend on t_d. The equation can be regarded as a rule for setting a "buffer" to allow for occasional delays, a buffer whose size balances the aversions to early and late arrival. If T_r has a tight distribution (low variance), then the desired probability P_L^* can be achieved by a small time buffer; but as the distribution of T_r becomes more dispersed (more unreliability), a larger time buffer is required, causing the usual early arrivals to be of greater magnitude and therefore to incur greater costs. Lomax, Turner and Margiotta (2003) estimate buffers in 21 US cities using an assumed value P_L^*=0.95, obtaining results that are substantial fractions of average travel time.

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, (2.60) again simplifies to a closed form:

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

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

$$v_R R = \left(\frac{\beta\gamma}{\beta + \gamma} \right) \cdot \frac{b}{2} . \tag{2.61}$$

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 (2.61) indicates that reliability cost derives from the combination of costly scheduling mismatches and dispersion in travel time. The specific value for v_R depends on how unreliability R is defined; if it is defined as half the possible range of travel times, then v_R is just the term in parentheses in (2.61).

More generally, the last two terms in (2.59) are potentially important if $\gamma > \beta$ or if θ is large, conditions that are in fact true according to the empirical findings in (2.30). Because they

contain $E(SDL)$ and P_L, these terms depend especially on the shape of the distribution of T_r in its upper ranges, which governs 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 (2.61) 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 and with b now defined as the headway between transit vehicles.[63] Although under that interpretation $v_R R$ 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 these literatures by others.

Waters (1996) reviews 56 value-of-time estimates from 14 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 value for personal travel by automobile equal to 50 percent of the average wage rate.

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.[64] 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; the average of these two values is 51% of

[63] This equivalence is pointed out by Wardman (2004, p. 364); the equation is also derived by de Palma and Lindsey (2001).

[64] 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).

the relevant wage rate.[65] Gunn (2001) find that Dutch values used in 1988, differentiated by level of household income, compare well with 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 vehicles. Another Dutch study — using a novel methodology in which the "choice" is to leave a job rather than to pick a 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 ¥2333/hour for weekday automobile travel in 1999, which was 84 percent of the wage rate.[66]

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, somewhat of an approximation.[67] The easiest way to summarize this evidence is as 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 *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. There are also 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 2.5 times as high. Wardman actually

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

[66] 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).

[67] See for example MVA Consultancy *et al.* (1987, pp. 133-135, 150, 152) and Mackie *et al.* (2003).

gets a smaller differential, namely a ratio of 1.62, which is quite precisely estimated; nevertheless he joins Mackie *et al*. (2003) in recommending 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.*, 1987, 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 *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 Winston (1998) obtain values using SP surveys of car commuters of 19 percent and 20 percent, respectively, of the wage rate.[68] 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,[69] 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, we conclude that the value of time for personal journeys varies widely by circumstance, usually between 20 and 90 percent of the gross wage rate and averaging around 50 percent. Although it varies somewhat less than proportionally with income or wages, expressing it as a fraction of the wage rate is a good approximation and is more useful than

[68] This statement is based on Calfee and Winston's (1998) summary of the average over the entire sample (p. 91), and on Hensher's (1997) Table 3.7 (p. 274), panel for "private commute," using his preferred VOT of $4.35/hour.

[69] See their Table 1, rows 4-5, 13-14

expressing it 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.) The value of time is much higher for business travel, generally taken as 100 percent of total compensation including benefits. The value of walking and waiting time for transit trips is 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 VOT equal to 41-58 percent of the corresponding mean VOT.[70] The California studies reviewed by Brownstone and Small (2005) measure heterogeneity as the inter-quartile range (75^{th} minus 25^{th} percentile values) of the distribution of VOT or VOR. With that measure, they find that unobserved heterogeneity in VOT — that is, heterogeneity due just to random coefficients — is 55-144 percent of median VOT.

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 90^{th}

[70] This statement is based on Hensher's (2001) Table 3, Model 3a, the lower panel showing values of time in which the cost coefficient is that on a variable measuring the toll.

and 50th percentile of the travel-time distribution across days, or some similar measure. In these studies also, VOR tends to be of about the same magnitude as VOT. One of these 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.[71]

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 of automobile users reviewed by Brownstone and Small (2005) obtain high VOT for automobile users even when simultaneously measuring VOR.

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 place primary emphasis on inter-city travel. De Jong (2000) provides a recent review of studies; they suggest that for countries like The Netherlands, where a high proportion of travel is urban, values of time are quite high. This finding is consistent with the common belief that travel time for freight vehicles is viewed as similar to business time for the driver plus some inventory value for equipment and payload. Kawamura (2000), Wigan *et al.* (2000), and Fowkes *et al.* (2004) provide 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 behavior. This is necessary because the variety of purposes and available choices make travel behavior very 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 several different approaches.

Both aggregate and disaggregate models can be instructive, the choice between them depending 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 have performed well in many but

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

not all 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 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.

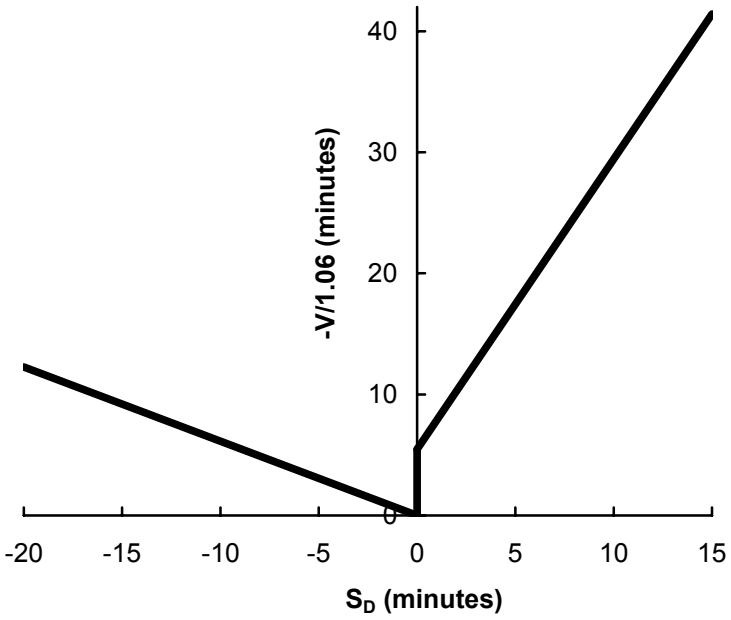

Figure 2.1 Disutility of Schedule Delay