**January 23, 2001Draft: do not quote without permission! Comments Welcome!**

# Discrete Choice Modeling for Transportation

David Brownstone
Dept. of Economics
3151 Social Science Plaza
University of California
Irvine, CA 92697-5100
Email: dbrownst@uci.edu

Paper prepared for 9th IATBR Travel Behavior Conference, Australia, July 2000.

Acknowledgements: Justin Tobias provided many useful references and comments on the third section of this paper. David Hensher, Clifford Winston, and Kenneth Small also provided many useful comments and references. Of course, none of these people are responsible for any of the errors or omissions in this paper. Financial support was provided by the University of California Transportation Center.

# 1 Introduction

This paper discusses important developments in discrete choice modeling for transportation applications. Since there have been a number of excellent recent surveys of the discrete choice literature aimed at transportation applications (see Bhat, 1997 and 2000a), this paper will concentrate on new developments and areas given less weight in recent surveys. Small and Winston (1999) give an excellent review of the transportation demand literature that includes many examples of how discrete choice models have been used in demand analysis.

Discrete choice modeling is closely related to activity-based modeling of travel demand and duration modeling. Since I have little to add to the excellent recent surveys on these topics by Bhat (2000b) and Bhat and Koppelman (2000), I have restricted this paper to "pure" unordered discrete choice modeling.

The next section discusses recent developments in flexible discrete choice modeling. Note that I define flexible to mean that the parametric model family is rich enough to arbitrarily approximate any discrete choice process consistent with random utility maximization, and I concentrate on the mixed logit model. There is also a relatively new literature which seeks to estimate discrete choice models without making parametric functional form assumptions (see Savin, 2001, Horowitz, 1998, and Koop and Poirier, 2000 for a Bayesian approach). Since this literature is currently limited to binary discrete choice, I have not included it in this paper.

Although the flexible discrete choice models discussed in Section 2 have improved our ability to estimate realistic disaggregate transportation models, there are still difficult problems with inference and model selection. Section 3 argues that these problems can only be solved by adopting a Bayesian perspective, and it reviews Bayesian discrete choice modeling. Measurement error in either the dependent or independent variables causes serious problems for discrete choice modeling. Multiple imputations is a new general technique for dealing with measurement error, and it is described in Section 4.

Measurement error in transportation demand models is typically caused by imputing key travel time and cost variables from network models. The example in Section 4 shows how multiple imputations can be used to correct for this type of measurement error when a small validation study is available to model the measurement error process.

Almost all transportation demand surveys are stratified on mode choice, since it is typically very expensive to collect a simple random sample with sufficient observations taking each mode. This choice-based sampling causes few problems for estimating the conditional logit model, but may be a serious problem for the more flexible models discussed in Section 2. Section 5 reviews the difficulties caused by choice-based sampling and their current solutions.

As panel surveys become more common, transportation demand analysts are going to need to specify and estimate dynamic discrete choice models. Although formally a special case of the flexible models discussed in Section 2, dynamic models have some special characteristics that are discussed in Section 6.

Finally, transportation demand analysts are rarely interested in making inferences about the individual parameters in discrete choice models. Typically value of time estimates are complicated nonlinear functions of the underlying model parameters. Section 7 describes simple bootstrap methods for generating valid confidence regions for these nonlinear functions.

## 2 Flexible Discrete Choice Models

Since McFadden (1973) pioneered disaggregate discrete choice modeling of travel behavior in the 1970s researchers have been concerned about the Independence from Irrelevant Alternatives (IIA) property implied by the conditional logit model. Of course, the IIA property is also implied by any discrete choice model with independent and identically distributed unobserved utility terms. McFadden's Nested Logit (1978) model provided a generalization that could handle the types of unobserved error corelations frequently encountered in transportation applications. Nested Logit is the most popular

member of the wider class of generalized extreme value models. Small (1987) derived the ordered generalized extreme value model, and Chu (1981) derived the paired combinatorial logit model (see also Koppelman and Wen, 2000).

None of these generalized extreme value models are flexible enough to approximate arbitrary discrete choice models, and recent work by Bhat (1998a and b, and 1999) and Brownstone and Train (1999) have demonstrated cases where Nested Logit is not sufficiently flexible to model travel behavior. The only models that are flexible enough to approximate any discrete choice model are Multinomial Probit and Mixed Logit. I will discuss Multinomial Probit in Section 3, since most of the interesting new developments are Bayesian.

At the last IATBR meetings in Austin there was considerable "buzz" about mixed logit models, and Section 2.1 reviews current developments. Although there have been some new applications from new investigators (Calfee, *et. al.*, 1998, and Hensher, 2000), most of the recent work on mixed logit has come from the same authors cited in Bhat (1997). One practical problem with mixed logit is that the initial software packages were computationally slow. This problem has been substantially improved by the application of powerful new methods for drawing "pseudo-random" numbers. These new numerical techniques (see Section 2.2) are also useful for simulated moment or simulated likelihood estimators for probit models, but so far they have only been implemented for mixed logit models.

The most fundamental problem with applying flexible discrete choice models is the difficulty of identifying error correlations from discrete choice data. Massive amounts of data are required to accurately estimate all but the simplest departures from IIA. It is therefore not surprising that all applications impose many restrictions (such as those implied by simple nesting or error component structures) on unobserved error correlations. For many practitioners, the only practical impact of flexible models is to justify various specification tests for the IIA assumption. The lagrange multiplier tests

for mixed logit models described below in Section 2.1 represent an important improvement over earlier tests.

Flexible models are particularly important for forecasting demand. Average coefficient values, which are all that is needed for value of time or willingness to pay estimation, are typically unchanged from MNL estimates. Brownstone and Train (1999) and Brownstone, *et. al.* (2000) give a number of examples where mixed logit showed substantial departures from IIA, but the average willingness to pay estimates were very similar to those obtained from misspecified conditional logit models. However, the mixed logit forecasts of market shares for new alternatives were very different from the conditional logit forecasts.

### *2.1 Mixed Logit Models*

A person faces a choice among J alternatives, which will be modeled using a random utility framework. I assume that the person's utility from any alternative can be decomposed into a nonstochastic, linear-in-parameters part that depends on observed data, a stochastic part that is perhaps correlated over alternatives and heteroskedastic, and another stochastic part that is independently, identically distributed over alternatives and people. In particular, the utility to person n from alternative i is denoted

$$U_{in} = \beta' x_{in} + [\eta_{in} + \varepsilon_{in}] \qquad (2.1)$$

where $x_{in}$ is a vector of observed variables relating to alternative i and person n; $\beta$ is a vector of structural parameters which characterizes choices by the overall population; $\eta_{in}$ is a random term with zero mean whose distribution over people and alternatives depends in general on underlying parameters and observed data relating to alternative i and person n; and $\varepsilon_{in}$ is a random term with zero mean that is independent and identically distributed over alternatives and does not depend on underlying parameters or data. For any specific modeling context, the variance of $\varepsilon_{in}$ may not be identified separately from $\beta$, so it is normalized to set the scale of utility.

Stacking the utilities, we have: $U = \beta'X+[\eta+\varepsilon]$ where $V(\varepsilon)=\alpha I$ with known (i.e., normalized) $\alpha$ and $V(\eta)$ is general and can depend on underlying parameters and data. For the standard conditional logit model, each element of $\varepsilon$ is independent and identically distributed extreme value, and, more importantly, $\eta$ is zero, such that the unobserved portion of utility (i.e., the term in brackets) is independent over alternatives. Taken together, these assumptions give rise to the IIA property and its restrictive substitution patterns.

The Mixed Logit class of models assumes a general distribution for $\eta$ and an iid extreme value distribution for $\varepsilon$. Denote the density of $\eta$ by $f(\eta|\Omega)$ where $\Omega$ are the fixed parameters of the distribution. (The density f may also depend upon explanatory data for people and alternatives, but in what follows this is suppressed for notational convenience.) For a given value of $\eta$, the conditional choice probability is simply logit, since the remaining error term is iid extreme value:

$$L_i(\eta) = \exp(\beta'x_i + \eta_i) / \sum_j \exp(\beta'x_j + \eta_j). \qquad (2.2)$$

Since $\eta$ is not given, the (unconditional) choice probability is this logit formula integrated over all values of $\eta$ weighted by the density of $\eta$:

$$P_i=\int L_i(\eta)\, f(\eta|\Omega)d\eta \qquad (2.3)$$

Models of this form are called "mixed logit" because the choice probability is a mixture of logits with f as the mixing distribution. The probabilities do not exhibit IIA, and different substitution patterns are attained by appropriate specification of f.

The choice probability cannot be calculated exactly because the integral does not have a closed form in general. The integral is approximated through simulation. For a given value of the parameters $\Omega$, a value of $\eta$ is drawn from its distribution. Using this draw, the logit

formula $L_i(\eta)$ is calculated. This process is repeated for many draws, and the average of the resulting $L_i(\eta)$'s is taken as the approximate choice probability:

$$SP_i = (1/R) \sum_{r=1,...,R} L_i(\eta^r) \qquad (2.4)$$

where R is the number of replications (i.e., draws of $\eta$), $\eta^r$ is the r-th draw, and $SP_i$ is the simulated probability that the person chooses alternative i. By construction, $SP_i$ is an unbiased estimate of $P_i$ for any R; its variance decreases as R increases. It is strictly positive for any R, so that $\ln(SP_i)$ is always defined, which is important when using $SP_i$ in a log-likelihood function (as below). It is smooth (i.e., twice differentiable) in parameters and variables, which helps in the calculation of elasticities and especially in the numerical search for the maximum of the likelihood function. The simulated probabilities sum to one over alternatives, which is useful in forecasting.

The choice probabilities depend on parameters $\beta$ and $\Omega$, which are to be estimated. Using the subscript n to index sampled individuals, and denoting the chosen alternative for each person by i, the log-likelihood function $\sum_n \ln(P_{in})$ is approximated by the simulated log-likelihood function $\sum_n \ln(SP_{in})$ and the estimated parameters are those that maximize the simulated log-likelihood function. Lee (1992) derives the asymptotic distribution of the maximum simulated likelihood estimator based on smooth probability simulators with the number of replications increasing with sample size. Under regularity conditions, the estimator is consistent and asymptotically normal. When the number of replications rises faster than the square root of the number of observations, the estimator is asymptotically equivalent to the maximum likelihood estimator.

The gradient of the simulated log-likelihood function is simple to calculate, which is convenient for implementing the search for the maximum:

$$G(\beta) \equiv \partial \sum_n \ln(SP_{ni}) / \partial \beta = \sum_n [1/SP_{ni}](1/R) \sum_r L_{ni}(\eta_n^r)[\sum_j (d_{nj} - L_{nj}(\eta_n^r))x_{nj}] \qquad (2.5)$$

$$G(\Omega) \equiv \partial \sum_n \ln(SP_{ni}) / \partial \Omega = \sum_n [1/SP_{ni}](1/R) \sum_r L_{ni}(\eta_n^r)[\sum_j (d_{nj} - L_{nj}(\eta_n^r))( \partial \eta_n^r / \partial \Omega)]$$

where $d_{nj} = 1$ for j=i and zero otherwise. The derivative $\partial\eta_n^r/\partial\Omega$ depends on the specification of $\eta$ and f. Also, if the same parameters enter $\beta$ and $\Omega$ the gradient is adjusted accordingly.

Analytic second derivatives can also be calculated. However, in contrast to the standard MNL model with its globally concave log-likelihood function, the inclusion of the $\Omega$ structural parameters removes the guarantee of global concavity, and the Hessian matrix is not guaranteed to be positive definite. This creates a more complicated situation for the iterative search, e.g., Revelt and Train (1998) found that calculating the Hessian from formulas for the second derivatives resulted in computationally slower estimation than using the BHHH or other approximate-Hessian procedures. To address this problem, Brownstone *et. al.* (2000) implemented specialized estimation code using the Bunch, Gay, and Welsch (1993) optimization software. These methods are more robust, and generally converge in many fewer iterations than the more standard numerical procedures (see Bunch, 1988). Although the number of iterations makes little practical difference when estimating MNL models, this is not longer true when using computationally intensive simulation approaches for calculating choice probabilities and gradients.

Different types of mixed logit models have been used in empirical work; they differ in the type of structure that is placed on the model, or, more precisely, in the specification of f. Train (1995) and Ben-Akiva and Bolduc (1996) specify an error-components structure: $U_i = \beta' x_i + \mu' z_i + \varepsilon_i$ where $\mu$ is a random vector with zero mean that does not vary over alternatives and has density $g(\mu|\Omega)$ with parameters $\Omega$; $z_i$ is a vector of observed data related to alternative i; and $\varepsilon_i$ is iid extreme value. This is a mixed logit with a particular structure for $\eta$, namely, $\eta_i=\mu' z_i$. The terms in $\mu' z_i$ are interpreted as error components that induce heteroskedasticity and correlation over alternatives in the unobserved portion of utility: $E([\mu' z_i +\varepsilon_i]'[\mu' z_j +\varepsilon_j] ) = z_i' V(\mu) z_j$. Even if the elements of $\mu$ are uncorrelated such that $V(\mu)$ is diagonal, the unobserved portion of utility is still correlated over alternatives.

In this specification, the choice probabilities are simulated by drawing values of $\mu$ from its distribution and calculating $\eta_i=\mu' x_i$. If as the number of error components (i.e., the dimension of $\mu$) is smaller than the number of alternatives (the dimension of $\eta$), placing an error-components structure on a mixed logit reduces the dimension of integration and hence simulation that is required for calculating the choice probabilities.

Different patterns of correlation, and hence different substitution patterns, are obtained through appropriate specification of $z_i$ and g. For example, an analog to nested logit is obtained by specifying $z_i$ as a vector of dummy variables -- one for each nest taking the value of 1 if i is in the nest and zero otherwise -- with $V(\mu)$ being diagonal (thereby providing an independent error component associated with each nest, such that there is correlation in unobserved utility within each nest but not across nests). Restricting $V(\mu)=\sigma I$ is analogous to restricting the log-sum coefficients in a nested logit model to be the same for all nests. Importantly, McFadden and Train (1998) have shown that any random utility model can be approximated by a mixed logit with an error-components structure and appropriate choice of the $z_i$'s and g.

Most recent empirical work with mixed logits has been motivated by a random-parameters, or random-coefficients, specification (Bhat, 1998a and b; Mehndiratti, 1996; Revelt and Train, 1998; Train 1998). The difference between a random-parameters and an error-components specification is entirely interpretation. In the random-parameters specification, the utility from alternative i is $U_i = b' x_i + \varepsilon_i$ where coefficients b are random with mean $\beta$ and deviations $\mu$. Then $U_i = \beta' x_i + [\mu' x_i + \varepsilon_i]$, which is an error-components structure with z = x. Elements of x that do not enter z can be considered variables whose coefficients do not vary in the population. And elements of z that do not enter x can be considered variables whose coefficients vary in the population but with zero means. In different contexts one or the other interpretation will seem more natural.

McFadden and Train (1998) also give Lagrange Multiplier tests for the presence of significant random error components in conditional logit models. These tests work by constructing artificial variables:

$$z_{in} = (x_{in} - x_{Cn})^2, \text{ with } x_{Cn} = \sum_j x_{jn} P_{jn} \qquad (2.6)$$

and $P_{jn}$ is the conditional logit choice probability. The conditional logit model is then re-estimated including these artificial variables, and the null hypothesis of no random coefficients on attributes *x* is rejected if the coefficients of the artificial variables are significantly different from zero. The actual test for the joint significance of the *z* variables can be carried out using either a Wald or Likelihood Ratio test statistic.

These Lagrange Multiplier tests can be easily carried out in any software package that estimates the conditional logit model. Our experience with these tests shows that they are easy to calculate and appear to be quite powerful omnibus tests. However, they are not as good for identifying which error components to include in a more general mixed logit specification. These Lagrange Multiplier tests also provide an alternative to standard Nested Logit specification tests that require estimating the model on a subset of the alternatives. In practice this frequently results in using only a small portion of the data, and this will lead to poor power properties. The Lagrange Multiplier tests are computed over the full set of alternatives and observations.

### *2.2 Low-dispersion sequences*

Computation of mixed logit choice probabilities in equation (2.3) typically requires Monte Carlo integration as in equation (2.4). Similar issues also arise in computing choice probabilities for multinomial probit models. The basic ingredient in this computation is the generation of "pseudo-random sequences" that are intended to mimic independent draws from a uniform distribution on the unit interval. Although these pseudo-random sequences cannot be distinguished from draws from a uniform distribution, they are not spread uniformly over the unit interval. Bhat (2000c) proposes replacing these pseudo-random sequences with sequences constructed from number theory to be more uniformly spread over the unit interval.

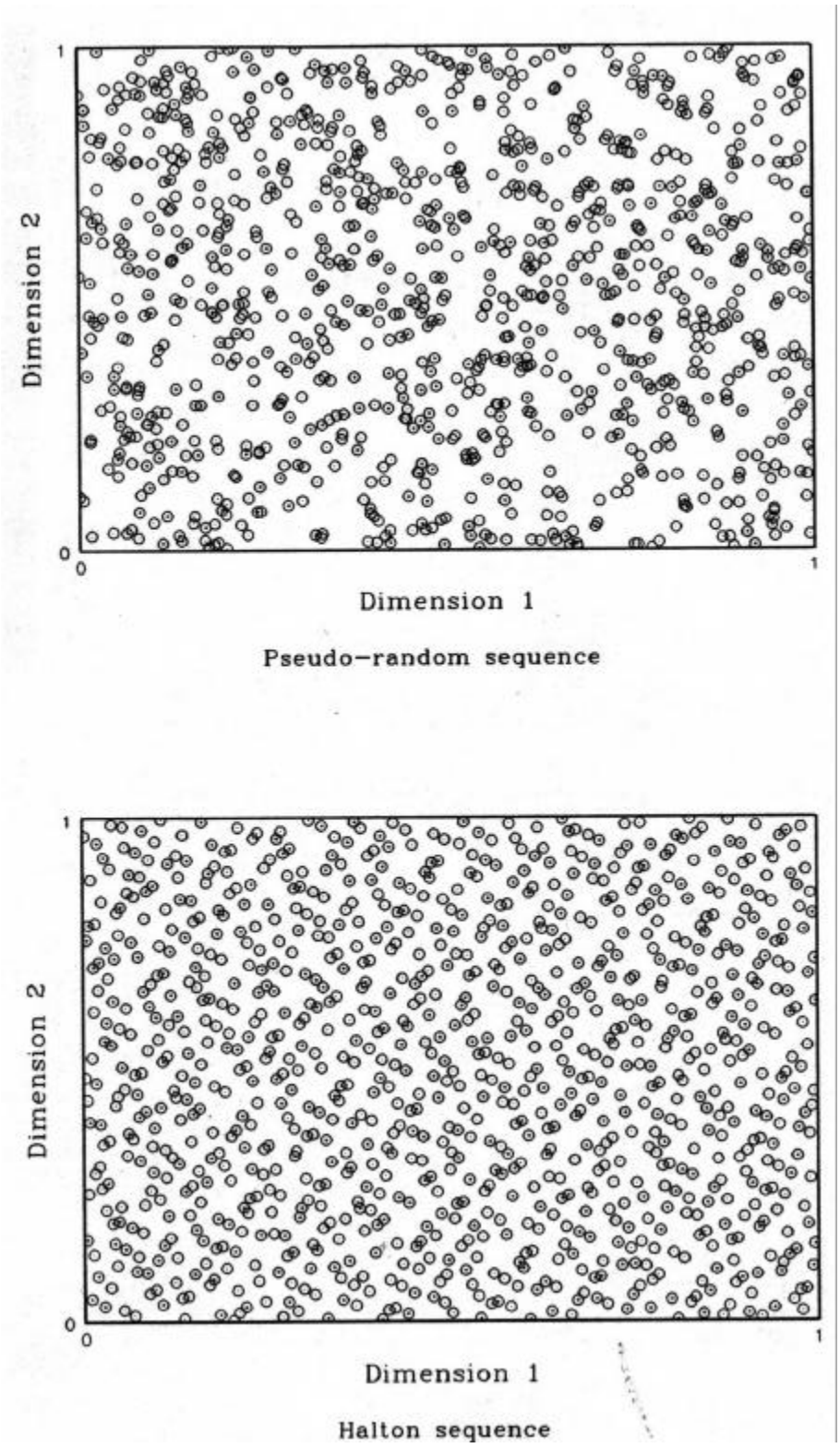



**Figure 1 1000 Draws on the Unit Square** (from Bhat (2000c))

These sequences, called low-dispersion sequences by mathematicians, yield much more accurate approximations in Monte Carlo integration relative to standard pseudo-random sequences. The reason for the superior performance of these sequences is shown in Figure 1. Even with 1000 draws, the pseudo-random sequences leave noticeable holes in the unit square, while the Halton sequence used by Bhat gives very uniform coverage.

Bhat (2000c) gives results from a Monte Carlo study of simulated maximum mixed logit models to compare the performance of the Halton sequence and the standard pseudo-random sequence. For four and five dimension integrals (as used in Brownstone *et. al.*, 2000) the Halton sequence methods required 125 draws to achieve the same accuracy as 2000 draws with the standard pseudo-random number sequences. As a result, the computation time required to estimate the mixed logit model using Halton sequences was 10% of the time required for the standard methods. Train (1999) and Revelt and Train (1999) have also reported similar large reductions in computation time using Halton sequences for mixed logit estimation.

These results clearly demonstrate the promise of these new numerical methods for estimating mixed logit models. Moreover, using these new methods to estimate multinomial probit models should result in similar improvements in computational speed and accuracy.

## 3 Bayesian Models

Although Bayesian methods are attracting increasing attention (Malakoff, 1999), there have been very few Bayesian discrete choice models in transportation. Applied researchers in other disciplines are adopting Bayesian techniques because they provide a principled approach for incorporating non-sample prior information, and they avoid asymptotic approximations. These advantages apply to discrete choice models used in transportation research, so this section will argue that transportation researchers should adopt Bayesian techniques.

There are many examples in transportation research where researchers have useful prior information. Mode choice modelers would all agree that the coefficients on travel time and price should be negative, and many would agree that the implied value of travel time should lie between zero and 150% of the respondent's wage. It is quite difficult to impose even simple non-negativity constraints on standard discrete choice model estimators, and if these constraints are imposed then non-standard inference procedures must be used (see Andrews, 1999). As will be shown later in this section, Bayesian methods incorporate prior information in a much simpler fashion.

A practical reason for slow adoption of Bayesian techniques has been their computational difficulty. Until the last decade numerically evaluating the complex integrals in realistic discrete choice models has been a daunting task. However the same advances in numerical algorithms and computing hardware that has enabled application of flexible discrete choice models in Section 2 have also enabled Bayesians to handle very complex models.

### *3.1 Bayesian versus classical inference*

This section is designed to introduce Bayesian methods and compare them to classical methods currently used in applied discrete choice modeling. A good introduction to modern Bayesian analysis is Carlin and Louis' (1996) textbook. More details on the computational aspects of Bayesian analysis are in Chen, Shao, and Ibrahim (2000). Geweke (1999) reviews Bayesian methodology as applied to econometric models, and he also describes current software for carrying out Bayesian analysis.

The key difference between Bayesian and classical statistics is that Bayesians treat parameters as random variables. Bayesians are therefore led to summarize their prior knowledge about parameters $\theta$ by a *prior* distribution, $p(\theta)$. The sampling distribution, or likelihood function, is given by $f(x|\theta)$. After observing some data, the information about $\theta$ is given by the *posterior* distribution:

$$p(\theta \mid x) = f(x \mid \theta)p(\theta) \Big/ \int f(x \mid \theta)p(\theta)d(\theta) \qquad (3.1)$$

Note that all inference is based on this posterior distribution. In many circumstances (under quadratic loss) the optimal Bayes estimator is the mean of the posterior distribution, and Bayesian confidence bands are typically given by the smallest region of the posterior distribution with the specified coverage probability. Bayesian confidence regions are interpreted as fixed regions containing the random parameter $\theta$ with the specified coverage probability (called Highest Posterior Density regions). This is very different from the classical confidence region, which is a region with random endpoints that contain the true value $\theta$ with the specified probability over independent repeated realizations of the data. Classical inference therefore depends on the distribution of unobserved realizations of the data, whereas Bayesian inference conditions on the observed data. Bayesian inference is also exact and does not rely on asymptotic approximations.

The Bayesian approach also requires the *a priori* specification of a prior distribution for all of the model parameters. In cases where this prior is summarizing the results of previous empirical research, specifying the prior distribution is a useful exercise for quantifying previous knowledge. There are many circumstances where the prior distribution cannot be fully based on previous empirical work, and the resulting specification of prior distributions based on the investigator's subjective beliefs is the most controversial part of Bayesian methodology. Poirier (1988) argues that the subjective Bayesian approach is the only approach consistent with the usual rational actor model adopted by economists and transportation researchers to explain consumers' choices under uncertainty. More importantly, the requirement to specify a prior distribution enforces intellectual honesty on Bayesian practitioners. All empirical work is guided by prior knowledge and the subjective reasons for excluding some variables and observations are usually only implicit in the classical framework. Bayesians are therefore forced to carry out sensitivity analysis across other reasonable prior distributions to convince others that their empirical results are not just reflections of their prior beliefs.

The simplicity of the formula defining the posterior distribution hides some difficult computational problems. Computing the posterior distribution typically requires integrating over $\theta$, and this can be difficult for the number of parameters frequently encountered in applied transportation work. Until recently Bayesians solved this problem by working with *conjugate families*. These are a family of prior distributions linked to a family of likelihood functions where the posterior distribution is in the same family as the prior distribution. For example, the Beta family is a conjugate prior for the binomial with fixed number of trials. Koop and Poirier (1993) have developed and applied a conjugate prior for the conditional (and multinomial) logit model, but there do not appear to be tractable conjugate priors for other GEV discrete choice models. Poirier (1996) shows how the prior family he developed for the conditional logit model can be extended to the Nested Logit model.

Poirier's analysis of the Nested Logit model highlights the ease with which Bayesians can cope with model uncertainty. For many applications of Nested Logit there are competing correlation structures (commonly associated with "trees"). These competing models cannot be nested in a larger Nested Logit model, and applied researchers frequently choose a "correct" model with little guidance from the data. The Bayesian approach to this problem does not require the choice of a correct model. Inference can be carried out unconditional on model choice. Suppose there are *M* competing models indexed by *m* with likelihood $f_m(x|\theta)$ and prior density $p_m(\theta)$. Let $\pi_m$ be the prior probability that model *m* is correct. If we define the marginal data density for model *m* by:

$$f_m(x) = \int f_m(x|\theta)p_m(\theta)\,\mathrm{d}\theta \qquad (3.2)$$

then the posterior probability that model *m* is correct is given by:

$$\bar{p}_m = p_m f_m(x) \Big/ \sum_{j=1}^{M} p_j f_j(x) \qquad (3.3)$$

These posterior probabilities might suggest that there is an obvious correct model, but in any case the unconditional posterior distribution for $q$ is then given by:

$$\bar{p}(q \mid x) = \sum_{j=1}^{M} \bar{p}_j \, p_j(q \mid x) \qquad (3.4)$$

In the common case where there is uncertainty about the correct model, then averaging over models as in equation (3.4) will almost always yield better results than arbitrary choice of a "correct" model.

Since transportation researchers or econometricians do not commonly use Bayesian analysis, it is useful to illustrate the key ideas with a simple concrete example. I will take a highly simplified version of the travel time measurement problem discussed in more detail in Section 4 of this paper. Suppose we are trying to measure the minutes required ($q$) to travel on a segment of a highway on a particular day and time. There are loop detectors under a few points along this segment, but there are 2 competing algorithms for converting the loop detector signals into speeds and travel times. Method 1's algorithm applied to this case can be summarized by a normally distributed prior distribution for $q$ with mean of 6 minutes and variance equal to 1. Method 2 similarly yields a normally distributed prior distribution for $q$ with mean of 10 minutes and variance equal to 4. To resolve these incompatible prior estimates, we conduct a small floating car experiment that involves driving 10 cars down the highway. The drivers of these cars use stopwatches to record their travel times, and we know from past experience with these drivers that the standard deviation of their measurements are equal to 1 and are normally distributed.

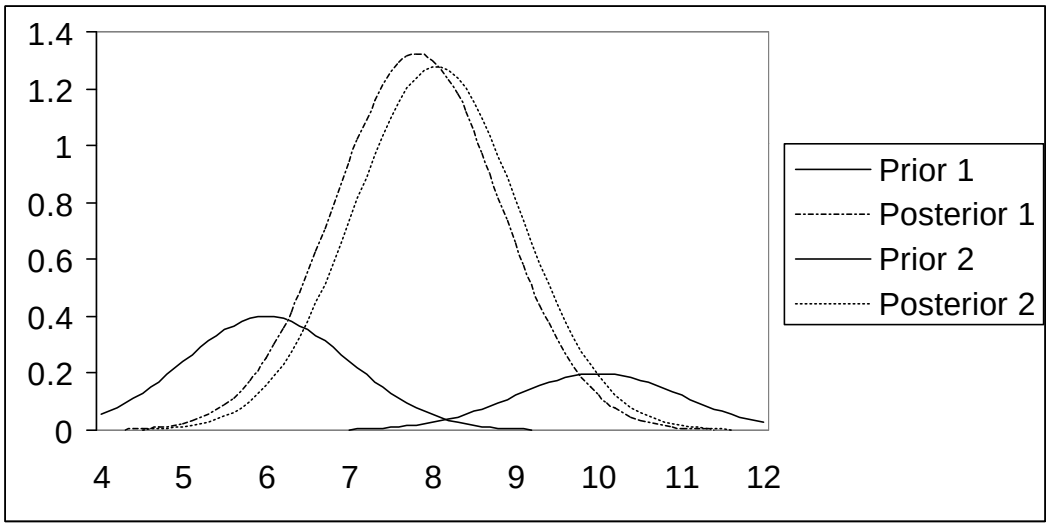
The likelihood function for the $n$=10 floating car measurements is given by a $N(\theta, \sigma^2)$ distribution, and the prior distribution for method $i$ is given by a $N(\mu_i, \tau_i^2)$ distribution. Simple algebra shows that resulting posterior distribution is given by a:

$N\left( \frac{\sigma^2 \mu_i}{\sigma^2 + n\tau_i^2} + \frac{n\tau_i^2 \bar{X}}{\sigma^2 + n\tau_i^2}, \frac{n\tau_i^2}{\sigma^2 + n\tau_i^2} \right)$ distribution. Note that since the prior and posterior distributions are both members of the Normal family, the Normal distribution is a conjugate prior for itself. The mean of the posterior distribution (which is the optimal Bayes point estimate under a squared error loss function) is given by a weighted average of the prior mean $(\mu_i)$ and the maximum likelihood estimate $(\bar{X})$. As the sample size ($n$) and/or the prior variance $(\tau_i^2)$ increases, the posterior mean approaches the maximum likelihood estimate. This implies that asymptotically the Bayes estimator is equal to the maximum likelihood estimator. Unlike standard asymptotic theory, however, Bayes inference is not approximate and gives exact finite sample results. Although I have illustrated these properties for this particular example, they apply to all Bayesian models.

If the sample mean $(\bar{X})$ of the floating car measurements is equal to 8, then Figure 2 shows the prior and posterior distributions for this example. The posterior distribution for method 1 is $N$(7.8, .09) and the posterior distribution for method 2 is $N$(8.05, .10). Although there is very little overlap between the prior distributions in this example, the posterior distributions are very close.

Figure 2 suggests that these data do not allow us to discriminate between the different methods for computing speeds from loop detector data. If we start out with equal prior probabilities that the two models are correct, then the posterior calculation in equation (3.3) gives a posterior probability of Method 1 being correct of .58. It is clear that more data would be required to discriminate between these methods, but if we are only interested in estimating $\theta$ from the floating car data it doesn't matter which prior is chosen. If we did want to collect more data, then we could just use the posterior distributions in Figure 2 as the prior distributions. All of the relevant information from the first experiment is contained in the posterior distributions.

**Figure 2**




### *3.2 Bayesian discrete choice models*

This section will concentrate on the multinomial probit model (equation 2.1 with the errors $\eta_{in} + \varepsilon_{in}$ following a multivariate normal distribution). Koop and Poirier (1993) and Poirier (1996) have developed Bayesian methods for the conditional and nested logit models, but I believe that most Bayesian applications will use multinomial probit because of the computational advantages relative to classical multinomial probit analysis and the greater flexibility of the multinomial probit model. If the latent utilities are observed, then the multinomial probit model just becomes a system of linear regression equations which can easily be analyzed using standard conjugate prior distributions. When the latent utilities are not observed, the Bayesian analyst faces the same problem as the classical statistician - computing the choice probabilities in high dimensional problems.

As computers have become more powerful and readily available, Bayesians have used simulation methods to calculate posterior distributions in complex problems with many parameters. In fact, many of the methods used to simulate likelihood functions and choice probabilities for multinomial probit and mixed logit models were developed by Bayesians. Classical statisticians need simulation to help maximize complex likelihood functions, while Bayesians need simulation to calculate complex posterior distributions.

As long as there is some way to simulate draws from a posterior distribution, Bayesian inference can be carried out as accurately as necessary. The most useful class of these simulation algorithms are called Markov Chain Monte Carlo methods. These methods, which include the Gibbs sampler described below, have the property that the successive draws come from a Markov chain whose stationary distribution is the joint posterior distribution.

McCulloch and Rossi (1994) and Geweke, *et. al.* (1997) use the Gibbs sampler (Gelfand and Smith, 1990) and data augmentation (Tanner and Wong, 1987) to carry out Bayesian inference for the multinomial probit model. Albert and Chib (1993) used similar methods for binary and ordered choice models. To provide a description of their algorithm in generic notation, let $U, \theta$, and $Y$ denote vectors of latent utilities, model parameters (including the slope parameters $\beta$ and the covariance parameters in the error distribution), and observed choice data ($Y_{ij} = 1$ if $U_{ij} = \max_k U_{ik}$ and 0 otherwise). Let $p(\theta, U \mid Y)$ denote the joint posterior density function for $\theta$ and $U$ conditional on $Y$. Suppose there is a partition of the parameter vector $\theta$ into $B$ subvectors, $\theta = (\theta_{(1)}, ..., \theta_{(B)})$, such that the conditional posterior densities $p(\theta_{(i)} \mid \theta_{(j)}, j \neq i, U, Y)$ and $p(U \mid \theta, Y)$ are of sufficiently simple form that it is practical to draw random subvectors $\theta_{(i)}$ and $U$ from these conditional densities. The Gibbs algorithm starts with an initial value $(\theta^{(0)}, U^{(0)})$, and then draws in turn each of the subvectors $U, \theta_{(1)}, ..., \theta_{(B)}$ from the appropriate conditional density, conditioned on the most recent values of the remaining parameters. After each draw, the corresponding initial value subvector is replaced by the new subvector, until after a complete iteration an updated vector $(\theta^{(1)}, U^{(1)})$ is obtained. After the $m$th iteration we obtain the draw $(\theta^{(m)}, U^{(m)})$. Under regularity conditions (see Roberts and Smith, 1993) the sample of $(\theta, U)$ draws converges in distribution to the joint posterior distribution as $m$ grows larger. Posterior inference on $\theta$ can be carried out using the corresponding draws from the Gibbs sampler. In particular, these draws can be used to carry out exact inference on the ratios of elements of $\theta$ as required in value of time estimation.

The Gibbs sampling algorithm works well for multinomial probit since the conditioning required to implement the sampler is easy to do with the underlying joint normality of the latent utilities and normal prior distribution for $\theta$. Note that the latent variables $U$ are added to the problem in the data augmentation step. This makes the conditioning on $Y$ in $p(\theta_{(i)} | \theta_{(j)}, j \neq i, U, Y)$ superfluous, so this step just draws from the conditional distribution of the utility index parameters $\theta$. Draws of the latent $U$ from $p(U | \theta, Y)$ are done by drawing from the conditional distributions of each component of $U$ conditioned on the remaining components of $U$, $\theta$, and $Y$. Choice probabilities can be recovered by numerically integrating over the draws of $U$ from the Gibbs sampler draws. These choice probabilities can either be computed for fixed values of $\theta$ or averaged over the posterior distribution of $\theta$.

Geweke *et. al.* (1997) carried out a Monte Carlo study comparing this Bayesian method (using the posterior mean as the point estimator of $\theta$) with classical simulated maximum likelihood and simulated method of moments using the GHK probability simulator. They found that the Bayesian Gibbs sampler method was more reliable (i.e. did not suffer from failure to converge), more accurate (especially for models with high error corelations), and required approximately the same computation time. Note that the Bayesian approach to multinomial probit does not require numerical optimization. The only way it can fail is if the Gibbs sampler doesn't converge to a stable equilibrium.

Allenby and Rossi (1999) have developed a Bayesian version of the mixed probit model. Although mixed logit is easier to compute in a classical setting, it is much easier to implement Gibbs sampling for the mixed probit model. Given the same correlation structure, these models are very similar. Using the same notation as in Section 2, Allenby and Rossi specify:

$$p(\eta_i, \beta, \theta | x) \propto f(x | \eta_i, \beta) p(\eta_i | \theta) p(\theta) p(\beta). \qquad (3.5)$$

Here $f(x \mid \eta_i, \beta)$ is the likelihood for an independent probit model, $\eta_i$ is the random effect for observation $i$, $\theta$ are the parameters of the distribution of $\eta_i$ over the sample, and $\theta$ and $\beta$ are independent. This is an example of a hierarchical Bayes model which have become increasingly popular in Bayesian analysis of linear panel data models. Unlike the classical mixed probit model which only estimates the population parameters $\beta$ and $\theta$, this Bayesian formulation permits inference for the individual random effects terms as well. These individual effects can be important in marketing applications applications. Revelt and Train (1999) have specified similar models using mixed logit and non-Bayesian methods. Revelt and Train point out that there are circumstances where the computational burden of their method will be less than Allenby and Rossi's Bayesian methods, but their inferences are only asymptotically valid while the Bayesian inferences are exact.

One important advantage of Bayesian analysis of flexible discrete choice models can be clarified by considering the all too frequent case where at least some of the covariance parameters are poorly identified. This means that the likelihood function will be almost flat along the dimensions corresponding to these parameters, and classical methods will therefore have problems converging to the optimum. As long as proper prior distributions are used, the Gibbs sampler will have no trouble converging, but it is very likely that the resulting posterior distribution for the poorly identified parameters will have the same shape as the prior distribution for these parameters. However, it is still possible to carry out informative Bayesian inference on other parameters of interest. More generally, comparison of the information conveyed in the prior and likelihood distributions is an excellent way to quantify the relative importance of these two inputs into Bayesian inference.

## 4 Measurement Errors

Discrete choice applications in transportation are plagued by serious measurement errors. It is very difficult to directly observe key variables for unchosen alternatives, so it is common practice to impute travel times and costs from network models. Unfortunately

this practice yields inconsistent parameter estimates and overstates the precision of these estimates. The standard approaches to measurement error in linear models are instrumental variables and joint modeling of the measurement error and behavioral processes. Unfortunately both of these approaches are very difficult to implement for the complex discrete choice models described in this chapter. This section describes a relatively new method which has been applied to discrete choice models with measurement error (Brownstone, 1998).

Rubin's (1987) multiple imputation methodology can be motivated as a method for consistent inference with imputed values for missing or erroneous observations (which are treated as having missing values for the correct data). If the imputed values are somehow produced to match the first two moments of the correct unobserved values, then standard estimation methods that treat the imputed values as if they are correct will yield consistent parameter estimates. Unfortunately the standard errors produced by this approach will be inconsistent and downward biased because they ignore the errors introduced by the imputation process. Rubin proposed solving this problem by independently drawing multiple imputed values. The component of variance due to the imputation error is then estimated by the variability of the estimates across the different imputed data sets. Typically drawing these multiple imputed values is the hard part of this methodology, so I will first describe Rubin's methods for combining results from multiply imputed data. Although Rubin developed the theoretical properties of this methodology for Bayesian models, Rubin (1996 and 1987, Chapter 4) show that these results apply asymptotically to classical statistical models.

Suppose we are interested in estimating an unknown parameter vector $\theta$. If no data are missing or measured with error, then we would use the estimator $\tilde{\theta}$ and its associated covariance estimator $\tilde{\Omega}$. If we have a model for predicting the missing (or erroneous) values conditional on all observed data, then we can use this model to make independent simulated draws for the missing data. If *m* independent sets of missing data are drawn and

*m* corresponding parameter and covariance estimators, $\tilde{\theta}_j$ and $\tilde{\Omega}_j$, are computed, then Rubin's Multiple imputation estimators are given by

$$\hat{\theta} = \sum_{j=1}^{m} \tilde{\theta}_j / m \qquad (4.1)$$

$$\hat{\Sigma} = U + \left(1 + m^{-1}\right) B, \qquad (4.2)$$

where

$$B = \sum_{j=1}^{m} \left(\tilde{\theta}_j - \hat{\theta}\right)\left(\tilde{\theta}_j - \hat{\theta}\right)' / (m-1) \qquad (4.3)$$

$$U = \sum_{j=1}^{m} \tilde{\Omega}_j / m. \qquad (4.4)$$

Note that *B* is an estimate of the covariance among the *m* parameter estimates for each independent simulated draw for the missing data, and *U* is an estimate of the covariance of the estimated parameters given a particular draw. *B* can also be interpreted as a measure of the covariance caused by the nonresponse (or measurement error) process.

Rubin (1987) shows that for a fixed number of draws, $m \geq 2$, $\hat{\theta}$ is a consistent estimator for $\theta$ and $\hat{\Sigma}$ is a consistent estimator of the covariance of $\hat{\theta}$. Of course *B* will be better estimated if the number of draws is large, and the factor $(1 + m^{-1})$ in equation (4) compensates for the effects of small *m*. Rubin (1987) shows that as *m* gets large, then the Wald test statistic for the null hypothesis that $\theta = \theta^0$,

$$\left(\theta - \theta^0\right)' \hat{\Sigma}^{-1} \left(\theta - \theta^0\right), \qquad (4.5)$$

is asymptotically distributed according to an F distribution with *K* (the number of elements in $\theta$) and $\boldsymbol{n}$ degrees of freedom. The value of $\boldsymbol{n}$ is given by:

$$\boldsymbol{n} = (m - 1)(1 + r_m^{-1})^2 \text{ and} \qquad (4.6)$$

$$r_m = (1 + m^{-1}) \text{ Trace}(BU^{-1})/K.$$

This suggests increasing *m* until $\boldsymbol{n}$ is large enough (e.g. 100) so that the standard asymptotic Chi-squared distribution of Wald test statistics applies. Meng and Rubin (1992) show how

to perform likelihood ratio tests with multiply-imputed data. Their procedures are useful in high-dimensional problems where it may be impractical to compute and store the complete covariance matrices required for the Wald test statistic (equation 4.5).

The key to successful implementation of multiple imputation is to use a *proper* imputation procedure. The full definition of a proper imputation procedure is given in Rubin (1987, pp. 118-119). Loosely speaking, if the estimates computed with the true values of the missing data ($\ddot{\theta}$ and $\ddot{\Omega}$) are treated as fixed, then $\hat{\theta}$ and $U$ must be approximately unbiased estimators of $\ddot{\theta}$ and $\ddot{\Omega}$. In addition $B$ must be an approximately unbiased estimator of the variation in $\hat{\theta}$ caused by the non-response mechanism. The safest way to generate proper imputation procedures is to explicitly draw from the (Bayesian) posterior predictive distribution of the missing values under a specific model. There are other proper multiple imputation procedures that require no explicit Bayesian calculations, and one such is described below. Any proper imputation procedure must condition on all observed data, and different sets of imputed values must be drawn independently so that they reflect all sources of uncertainty in the response process.

I will illustrate Rubin's multiple imputation methodology using an example from Brownstone *et. al.* (1999). We use new data from the San Diego congestion pricing demonstration project (referred to as FasTrak, see Kazimi *et. al.*, 1999). This project allows solo drivers to pay to use an eight-mile stretch of reversible high occupancy vehicle (HOV) lanes along Interstate Route 15 (I-15). The combination of free HOV use and priced solo driver use is generally referred to as high occupancy toll (HOT) lanes. In this demonstration project, HOT lane users must travel the entire eight-mile length before exiting. The per-trip fee for solo drivers is posted on changeable message signs upstream from the entrance to the lanes, and may be adjusted every six minutes to maintain free-flowing traffic conditions in the HOT lanes. Solo drivers who subscribe to the FasTrak program are issued windshield-mounted transponders used for automatic vehicle identification. Each time they use the lanes, their accounts are automatically debited the per-trip fee. This represents a dynamic form of voluntary congestion pricing, where solo

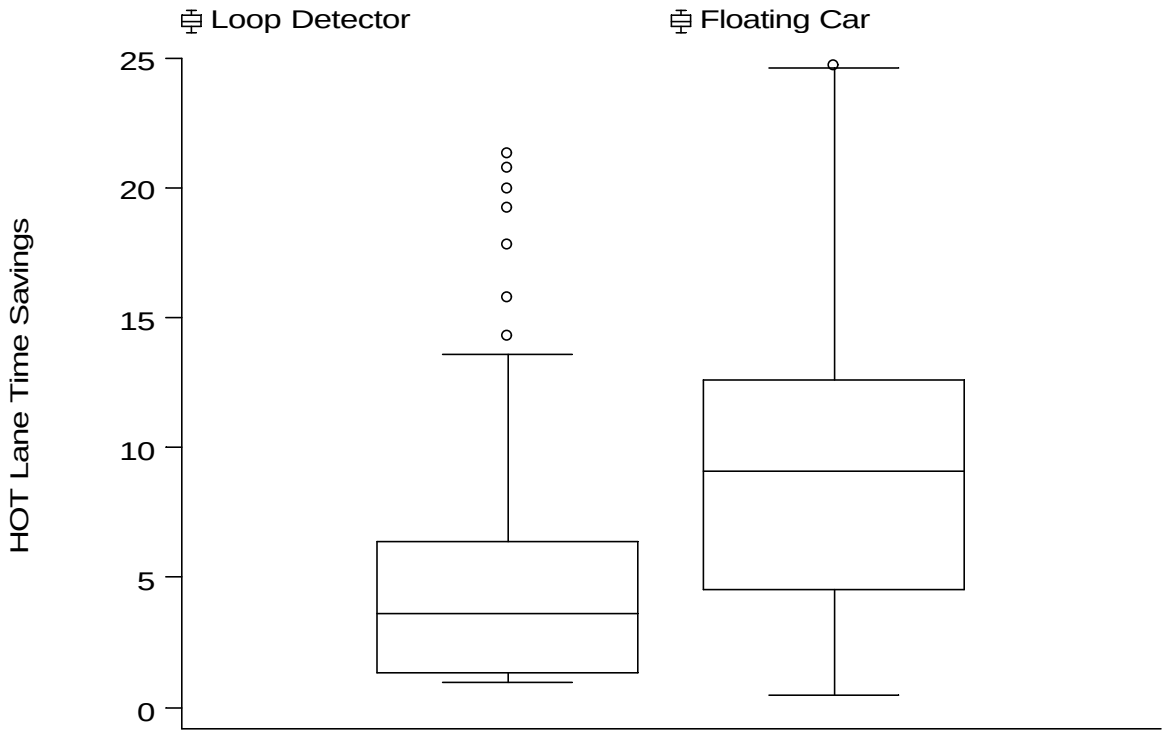
drivers can choose to pay to reduce their travel time, and the payment is generally related to the level of congestion. Carpoolers can still use the HOT lanes with no charge.

We are interested in modeling commuter's mode choice (solo drive, carpool, or FasTrak), but there is substantial measurement error in the time saved by using the HOT lanes. HOT lane users, and especially solo drivers paying to use the carpool lanes, tend to report unrealistically high values of time savings. While it is certainly possible that their mode choice decisions are based on their perceptions rather than the objective time savings, any useful policy model needs to be sensitive to actual time savings. Objective measurements of time savings are available from two types of data on speeds. First, floating car observations were obtained by driving cars down the corridor at frequent intervals and recording the actual travel times. Due to the high costs of collecting floating car data, they are only available for 5 consecutive days in the middle of our two-month survey period. The second type of data on travel times, point speeds derived from magnetic loop detectors placed along the corridor for general traffic counting purposes, are available during the entire data collection period, but Figure 3 shows that these data are subject to large errors. The loop detector data generally understate the actual time savings by 50%.

Since our mode choice model requires accurate time savings data for the entire two-month sample period, we used the loop detector data, toll (which is related to congestion and time savings in this experiment), and time of day to predict the missing floating car data. Table 4.1 shows the best fitting linear regression model for predicting floating car HOT lane time savings. To avoid unreasonable predictions we first transform both time savings measures to keep them bounded between zero and 35 minutes, which is the maximum observed loop detector time savings. The exact transformation for both time savings variables is given by the following transformed logit:

$$\log\left(\left(\frac{t}{35}\right)\Big/\left(1-\frac{t}{35}\right)\right). \qquad (4.7)$$

**Figure 3: Distribution of HOT Lane Time Savings**




We tried a number of different specifications including higher order terms in loop detector time savings and toll variables, but none of them significantly improved the fit of the model. We also experimented with lagged values, but the cubic polynomial in time effectively removes the autocorrelation in the time savings measures. Since the purpose of this model is accurate prediction, we are looking for the most parsimonious model with the best fit. Although the variables involving the tolls are not individually significant, they are jointly significantly different from zero at the one percent level. If they are excluded from the model, then the $R^2$ drops slightly to .89. However, excluding the loop detector data reduces the $R^2$ to .82 and increases the root MSE of the residuals to .46.

**Table 4.1: Imputation Model for Floating Car HOT Lane Time Savings**

| Dependent Variable: Logit of Floating Car Time Savings | | | $R^2$ = 0.90<br>Root MSE = 0.36 |
|---|---|---|---|
| Independent Variables: | Coef. | Std. Err. | *t*-Stat. |
| Logit of Loop Detector Time Savings × Minutes Past 5:00 A.M. | 0.0029 | 0.00031 | 9.3 |
| Minutes Past 5:00 A.M. | 0.222 | 0.0149 | 14.8 |
| (Minutes Past 5:00 A.M.)$^2$ | -0.00138 | 0.000121 | -11.4 |
| (Minutes Past 5:00 A.M.)$^3$ | 2.73E-06 | 2.91E-07 | 9.38 |
| Toll | -0.229 | 0.188 | -1.22 |
| Toll × Minutes Past 5:00 A.M. | 0.00222 | 0.00126 | 1.77 |
| Constant | -11.4 | 0.52 | -22.1 |

To draw one set of imputed values for the missing floating car data, first draw one set of slope and residual variance parameters from the asymptotic distribution of the linear regression estimators from Table 4.1. The slope parameters are drawn from the joint normal distribution centered at the parameter estimates with covariance given by the usual least squares formula ($s^2(X'X)^{-1}$). The residual variance, $s_*^2$, is drawn by dividing the residual sum of squares by a draw from an independent $c_d^2$ distribution, where *d* is the residual degrees of freedom. An imputed residual vector is then drawn from independent normal distributions with mean zero and variance equal to $s_*^2$. The imputed values are then computed by adding this imputed residual to the predicted value from the regression using the imputed slope parameters. Additional sets of imputed values are drawn the same way beginning with independent draws of the slope and residual variance parameters. Observations where floating car data are observed are fixed at these observed values across all imputations. This imputation method, which Schenker and Welsh (1988) call the "normal imputation" procedure, is equivalent to drawing from the Bayesian predictive posterior distribution from a standard linear regression model with uninformative priors.

We used the multiply imputed time savings data to fit a multinomial logit mode choice model. The final estimates were calculated from equations 4.1 – 4.4. Relative to the downward biased standard errors calculated by treating the imputed values as correct, multiple imputation standard errors from equation 4.2 were 30 - 50% larger. Of course, this bias in the standard errors would be much larger if the imputation model was less accurate.

The model specification includes interaction terms with the time savings variable, so the implied value of time saved by taking the HOT lane varies over the sample. Table 4.2 shows the distribution of this value of time for the model estimated on the loop detector data and the multiply imputed model using the corrected time savings data. There are substantial differences in the lower part of the distribution, and the mean and median are about one third lower for the corrected estimates.

**Table 4.2: Implied Value of Time Saved from Mode Choice Model**

| **Value of Time ($/hour)** | Corrected | Loop Data |
|---|---|---|
| 95$^{th}$ Percentile | 108.70 | 105.60 |
| 90$^{th}$ Percentile | 72.12 | 73.63 |
| 75$^{th}$ Percentile | 31.30 | 35.27 |
| 50$^{th}$ Percentile | 18.71 | 23.37 |
| 25$^{th}$ Percentile | 10.30 | 16.55 |
| 10$^{th}$ Percentile | -20.72 | 14.43 |
| 5$^{th}$ Percentile | -83.02 | 14.08 |
| Mean | 25.63 | 32.64 |

The multiple imputation approach is computationally quite simple. All of the calculations in Brownstone *et. al.* (1999) were done using the STATA system. Perhaps the largest advantage of multiple imputations is that it allows the imputations to be made once and then used for a variety of analyses. This allows agencies or researchers who collect data to represent the uncertainty in their data by including multiply imputed values for key variables. The U.S. Federal Reserve Board now provides multiply imputed income and wealth variables in the public release of its Survey of Consumer Finances, and the U.S. Bureau of Labor Statistics is experimenting with multiply imputing income and durable expenditures in its Consumer Expenditure Survey. Note

that these imputations can take advantage of confidential information (such as precise location) which are not normally released in public use data sets.

## 5 Choice-Base Sampling

Many travel demand data are collected by stratifying on the mode choice variable. These "choice-based samples" are particularly useful when some of the modes have small market shares since the sample scheme allows "on-board" surveying for those choosing rare modes. Maximizing a random-sample likelihood function with a choice-based sample will generally yield inconsistent parameter estimates. McFadden (see proof in Manski and Lerman, 1977) shows that for the conditional logit model with a full set of mode-specific constants only the parameters associated with these mode-specific constants are inconsistent. A relatively simple estimator that yields consistent estimates under choice-based sampling was developed by Manski and Lerman (1977). Their Weighted Exogenous Sample Maximum Likelihood Estimator (WESMLE) is the maximand of the weighted likelihood function:

$$\sum_n w_n L_n(q, x_n), \tag{5.1}$$

where $L_n$ is the log likelihood function for the $n^{th}$ observation and the sampling weight, $w_n$, is the inverse of the probability that the $n^{th}$ observation (individual) would be chosen from a completely random sample of the population. Of course, if the sampling scheme were completely random, then all of the sampling weights would be equal and the WESMLE would simply be the usual maximum likelihood estimator. For a simple choice-based sample, the WESMLE weights are just given by the ratio of the population mode share divided by the sample mode share. This is just the inverse of the sampling probability multiplied by the sample size divided by population size to make the sum of the weights equal the sample size.

Manski and Lerman (1977) show that the WESMLE is consistent and asymptotically normal, but not fully efficient (see Imbens, 1992 for fully efficient alternative estimators).

Manski and Lerman's proof actually shows that the WESMLE's properties hold for any likelihood function (subject to regularity conditions) as long as the sampling weights are known with certainty. DuMouchel and Duncan's (1983) regression estimator for non-random samples is just equation (5.1) with the likelihood function given by the standard normal regression model. The asymptotic covariance of the WESMLE is given by:

$V = \Psi^{-1}\Lambda\Psi^{-1}$, where

$$\Psi = -E\left(\partial^2 w_n \, L_n(q, x_n) \Big/ \partial q \partial q'\right) \quad \text{and} \tag{5.2}$$

$$\Lambda = E\left(\left(\partial w_n L_n(q, x_n) \Big/ \partial q\right)\left(\partial w_n L_n(q, x_n) \Big/ \partial q'\right)\right).$$

This covariance matrix can be consistently estimated by replacing the expectations in equation (5.2) with sample moments evaluated at the WESMLE estimates.

A major advantage of the WESMLE is that it can be computed easily by modifying existing maximum likelihood programs. The WESMLE for both the linear regression model and the conditional logit model can be computed by appropriately weighting the variables and applying standard maximum likelihood programs. Unfortunately, this procedure yields downward biased standard error estimates, but the consistent estimates given by equation (5.2) are straightforward to compute. This downward bias can be substantial in common applications. The incorrect standard errors for the models in Section 6 are typically downward biased by 50 percent relative to the correct standard errors in equation (5.2).[1]

Note that if the weights are small for a rare mode with most of the variation in key variables, then it will be difficult to get accurate estimates with the Manski-Lerman estimator. Cosslett (1981) noted that the WESMLE is inefficient, and proposed an

[1] A STATA program for computing the WESMLE and the correct standard errors for the conditional logit model is available from the author.

efficient alternative. Unfortunately, Cosslett's estimator is very difficult to compute, and it has only been applied by McFadden *et. al.* (1985). More recently, Imbens (1992) has developed an efficient weighted generalized method of moments estimator for choice-based samples. Imbens adds the restriction that the weighted mode shares must equal the (known) population shares in addition to the moment conditions equivalent to the first order conditions for unweighted maximum likelihood estimation. Although Imben's estimator is simpler to compute than Cosslett's estimator, it cannot be computed by simply weighing the data as in the WESMLE. Wooldridge (1999) shows how these results carry over to other non-random sampling schemes frequently encountered in transportation surveys, and Lancaster (1997) analyzes choice-based sampling from a Bayesian perspective.

Applied researchers have not paid much attention to improved choice-based sample estimators because most use the conditional logit model. McFadden showed that using unweighted maximum likelihood with choice-based samples only causes inconsistency in the alternative-specific constants (although this does require a full set of alternative-specific constants!). As more flexible discrete choice models are used in applied work, researchers will need to pay more attention to efficient estimation with choice-base and other non-random samples.

## 6 Dynamic Discrete Choice Models

As panel data have become more widely utilized in transportation (see Golob, *et. al.*, 1997, and Raimond and Hensher, 1997), researchers have been confronted with the need to model dynamic discrete choice data. Heckman (1981) presented a general framework for modeling, and Chamberlain (1984, especially Section 3) provided links to the more-established methods for dynamic linear panel data models. This section will concentrate on methods for modeling the autocorelation in the unobserved utilities resulting from observing repeated choices from the same individual or household over time. I will also restrict attention to multinomial discrete choice settings since these are ubiquitous in transportation.

Following advances in simulation-based inference approaches (McFadden, 1989, Pakes and Pollard, 1989, Keane, 1994) there have been a number of applications of the multinomial multiperiod probit model. These applications include decision to work (Keane, 1994), brand choice (Elrod and Keane, 1995, and McCulloch and Rossi, 1994), and residential location (Hajivassiliou *et. al.*, 1996). These models take the general form:

$$U_{itn} = \beta_n x_{itn} + \varepsilon_{itn}\,, \qquad (6.1)$$

where t denotes time. In some stated preference applications, t indexes successive choice experiments given to the same respondent. The subscript on $\beta$ denotes that there may be random coefficients in the model. Both the random coefficients and the error terms are assumed to follow multivariate normal distributions. If we just restrict attention to the error terms, then allowing free correlations across time and alternatives will yield a large number of parameters that will be hard to identify with real data. Therefore most applications with more than two time points use a simple first-order autoregressive process to model correlation across time.

The mixed logit model can also be used to model dynamic discrete choice. For the same reasons mentioned in Section 2.1, mixed logit can have substantial computational advantages over multinomial probit when there are many alternatives and relatively few error components. A fairly general mixed logit specification is given by:

$$U_{itn} = b_{tn}'x_{itn} + \varepsilon_{itn}\,, \qquad (6.2)$$

where $\varepsilon_{itn}$ are independent and identically distributed according to a standard Weibull distribution, $b_{tn} = \rho\, b_{(t-1)n} + \nu_{tn}$ where $\nu_{tn}$ are independent and identically distributed according to a $N(0, \sigma^2)$ distribution, and the starting conditions are such that $E(b_{0n}) = \beta$ and $Var(b_{0n}) = \upsilon$. This is a random coefficients model where the coefficients evolve according to a first-order autoregressive process. Brownstone *et. al.* (2000) estimated the special case given by $b_{tn} = b_{0n}$ and $\sigma^2 = 0$ for two repetitions of a stated preference alternative-fuel vehicle choice experiment. This model corresponds to each individual having their own preference parameters that remain fixed across repeated choice settings. This special case is very easy to implement since it just requires fixing the draw of the random coefficient for

different choices made by the same individual. Brownstone *et. al.* (2000) did not find any significant difference between this model and an independent choice model, but Hensher (1999) reports significant differences using similar autocorrelation structures for repeated stated preference data.

## 7 Inference for Discrete Choice

While some of the methods proposed in this paper are new, all discrete choice modelers know that the estimated coefficients are not individually useful. Because the scale of the error terms are not identified, the scale of the individual coefficients is also not identified. Therefore we typically look at ratios of the coefficients (usually identifying willingness to pay or value of time in the model), or use the coefficients to carry out demand simulations. Even though these are the quantities of interest for policy analysis, it is very rare that any confidence region is given. Judging from reading many applied papers, the implied assertion is that if the individual coefficients have high *t*-statistics, then any nonlinear combination of them must also have high *t*-statistics.

Of course this is obviously false. Even if the asymptotic normal approximation to the joint distribution of the parameter estimates is accurate, there is no reason why the ratios of any two of these coefficients would even have a mean or a variance. If the coefficient estimates are uncorelated, then the ratios will typically have a Cauchy distribution (which has no finite moments). This fact suggests that standard delta-method approximations (see Greene, 1997, pages 127 and 916) will not yield reliable inferences, although the resulting standard error estimates are certainly better than nothing!

A more reliable and general method is parametric bootstrapping. This requires drawing from the estimated asymptotic distribution of the parameter estimates, and computing the nonlinear function for each independent draw. If this process is repeated many times, any feature of the sampling distribution of the nonlinear function can be accurately estimated. Since the moments of these distributions may not exist, confidence regions should be estimated directly using percentiles of the sampling distribution. These calculations can

be carried out with minimal programming in most econometric software packages (including STATA and LIMDEP).

Of course, Bayesians don't need to carry out any additional calculations to generate confidence regions for nonlinear functions of choice model parameters. Modern Bayesian inference requires many draws from the posterior distribution, and these draws can easily be used to produce the posterior distribution of any nonlinear function of the underlying parameters.

Since it is not very difficult to produce confidence bands for value of time estimates, it is likely that they are not being computed because researchers believe that the resulting confidence bands will be very wide. This view is supported by a brief review of empirical value of time studies. Based on his review, Small (1992) suggests that 50 percent of gross wage rate is a reasonable value of time estimate. On the higher end of previous studies, Cambridge Systematics (1977) estimate that value of time for commuters in Los Angeles is 72 per cent of gross hourly wage. These previous studies are based upon mode choice models that consider differences between transit and automobile travel, and to the extent that differences between crowded transit and private automobiles are not captured, the results will be biased. In more recent work, Calfee and Winston (1998) attempt to avoid this problem by using stated preference data that only considers the tradeoff between travel by automobile in slower, free lanes and travel by automobile in faster, priced lanes. Their results indicate that commuters place a lower value on time saving than previously estimated (roughly $3.50 to $5.00 per hour or 15 to 25 percent of hourly wage). Calfee and Winston rely upon stated preference data because they lack revealed preference data for the choices involved with congestion pricing. Hensher (1999) shows that value of time estimates from complex stated preference data are sensitive to the specification of unobserved effects.

Brownstone *et. al.* (1999) used revealed preference data from a new congestion pricing experiment in San Diego (see Section 4). Their mean value of time estimate in Table 4.2 is about 50% of gross wage, which is consistent with Small's findings. However, the

standard error of this estimate is 150% of gross wage, which means that all known estimates are within one standard error of the "consensus" value. While this is not a pleasant result, it focuses attention on improving our basic data and models.

## 8 Conclusion

Although there have been significant advances in traditional discrete choice methodology (see Section 2), this paper argues that it is time to consider different approaches. In particular, the problems of model selection and measurement error are ubiquitous in applied transportation demand analysis. These problems are much easier to handle using a Bayesian paradigm reviewed in the third section of this paper. Although the fourth section of this paper described the multiple imputations methodology for measurement error in a classical framework, it is also fundamentally a Bayesian methodology. Major advances in Bayesian computation have allowed Bayesian analysis of very complex multinomial discrete choice models, and these same computational advances have also been instrumental in classical discrete choice modeling. It is now time apply these methods to transportation demand analysis.

## 9 References


Albert, S. H. and S. Chib (1993). Bayesian analysis of binary and polychotomous response data. Journal of the American Statistical Association, 88, 669-679.

Allenby, G. M. and P. E. Rossi (1999). Marketing Models of Consumer Heterogeneity. Journal of Econometrics, 89, 57-78.

Andrews, D. W. K. (1999). Estimation when a parameter is on a boundary. Econometrica, 67, 1341-1384.

Ben-Akiva, M. and D. Bolduc (1996). Multinomial probit with a logit kernel and a general parametric specification of the covariance structure. Working paper, Department of Civil and Environmental Engineering, MIT.

Bhat, C. (1997). Recent Methodological Advances Relevant to Activity and Travel Behavior Analysis. Presented at the 8th Meeting of the International Association for Travel Behavior Research, September 21-25, Austin, TX. To appear in H.S. Mahmassani, (ed.), Recent Developments in Travel Behavior Research. Pergamon, Oxford, in press.

Bhat, C. (1998a). Accomodating flexible substitution patterns in multidimensional choice modeling: formulation and application to travel mode and departure time choice. Transportation Research, 32B, 425-440.

Bhat, C. (1998b). Accomodating variations in responsiveness to level-of-service measures in travel mode choice modeling. Transportation Research, 32A, 495-507.

Bhat, C. (1999). Incorporating observed and unobserved heterogeneity in urban work travel mode choice modeling. Transportation Science, forthcoming.

Bhat, C. (2000a). Flexible model structures for discrete choice analysis. Forthcoming in Handbook of Transport I: Transport Modeling.

Bhat, C. (2000b). Duration modeling. Forthcoming in Handbook of Transport I: Transport Modeling.

Bhat, C. (2000c). Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model, Transportation Research B, forthcoming.

Bhat, C. and F. S. Koppelman (2000). Activity-based modeling of travel demand. Forthcoming in Handbook of Transportation Science.

Brownstone, D. (1998). Multiple Imputation Methodology For Missing Data, Non-Random Response, And Panel Attrition." In Theoretical Foundations of Travel Choice Modeling, eds. T. Gärling, T. Laitila and K. Westin, 421-450, Amsterdam: Elsevier.

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

Brownstone, D., T.F. Golob and C. Kazimi (1999). Modeling non-ignorable attrition and measurement error in panel surveys: an application to travel demand modeling. Presented at the International Conference on Survey Nonresponse, October 28-31, Portland, OR. To appear in Groves, R.M., D. Dillman, J.L. Eltinge and R.J.A. Little, eds. Survey Nonresponse. New York: Wiley.

Brownstone, D., D. S. Bunch, and K. Train (2000). Joint Mixed Logit Models of Stated and Revealed references for Alternative-fuel Vehicles. Transportation Research B, 34, 315-338.

Bunch, D. S. (1988). A comparison of algorithms for maximum likelihood estimation of choice models. Journal of Econometrics, 38, 145-167.

Bunch, D. S., D. M. Gay, and R. E. Welsch (1993). Algorithm 717: Subroutines for maximum likelihood and quasi-likelihood estimation of parameters in nonlinear regression models," ACM Transactions on Mathematical Software, 19(1), 109-130.

Calfee, J. and Winston, C. (1998). The Value Of Automobile Travel Time: Implications For Congestion Policy. Journal of Public Economics, 69, 83-102.

Calfee, J., C. Winston, and R. Stempski. (1998). Econometric issues in estimating consumer preferences from stated preference data: a case study of the value of

automobile travel time. Washington , DC: American Enterprise Institute working paper (December, 1998).

Cambridge Systematics, Inc. (1977). The Development of a Disaggregate Behavioral Work Mode Choice Model. Prepared for California Department of Transportation and Southern California Association of Governments. Cambridge, MA.

Carlin, B. P. and T. A. Louis (1996), Bayes and Empirical Bayes Methods for Data Analysis. New York: Chapman and Hall.

Chamberlain, G. (1984). Panel Data, in Z. Griliches and M.D. Intriligator (eds.), Handbook of Econometrics Vol. 2. Amsterdam: North-Holland, 1248-1318.

Chen, M-H., Q-M. Shao, and J. G. Ibrahim (2000). Monte Carlo Methods in Bayesian Computation. New York: Springer.

Chu, C. (1981). Structural issues and sources of bias in residential location and travel mode choice models. Unpublished Ph.D. dissertation. Dept. of civil Engineering, Northwestern University.

Cosslett, S. R. (1981). Efficient Estimation of Discrete Choice Models, in C. F. Manski and D. McFadden (eds.). Structural Analysis of discrete Data with Econometric Applications, Cambridge, MA: MIT Press, 51-111.

DuMouchel, William H. and Gregory J. Duncan (1983). Using Sample Survey Weights in Multiple Regression Analysis of Stratified Samples, Journal of the American Statistical Association 78, 535-543.

Elrod, T. and M. Keane (1995). A Factor-analytic Probit Model for Estimating Market Structure in Panel Data. Journal of Marketing Research 32, 1-16.

Gelfand, A. E. and A. F. M. Smith (1990). Sampling based approaches to calculating marginal densities. Journal of the American Statistical Association, 85, 398-409.

Geweke, J. F. (1999). Using simulation methods for Bayesian econometric models: inference, development, and communication (with discussion and reply). Econometric Reviews, 18(1), 1-126.

Geweke, J. F., M. P. Keane, and D. E. Runkle (1997). Statistical Inference in the Multinomial Multiperiod Probit Model. Journal of Econometrics, 80, 125-167.

Golob, T.F., R. Kitamura, and L. Long (eds.) (1997). Panels for Transportation Planning. Boston: Kluwer Academic Publishers.

Greene, W.H. (1997). Econometric Analysis, Third Edition. New Jersey: Prentice-Hall.

Hajivassiliou, V., D. McFadden, and P. Ruud (1996). Simulation of multivariate normal rectangle probabilities and their derivatives: theoretical and computational results. Journal of Econometrics, 72, 85-134.

Heckman, J.J. (1981). Statistical Models for Discrete Panel Data, in C. F. Manski and D. McFadden (eds.). Structural Analysis of Discrete Data with Econometric Applications, Cambridge, MA: MIT Press, 114-178.

Hensher, D. (2000). Measurement of the Valuation of Travel Time Savings. Journal of Transport Economics and Policy, forthcoming.

Hensher, D. (1999). The Sensitivity of the Valuation of Travel Time Savings to the Specification of Unobserved Effects. Transportation Research E, forthcoming.

Horowitz, J. (1998). Semiparametric Methods in Economics, New York: Springer-Verlag.

Imbens, G. (1992). An Efficient Method Of Moments Estimator For Discrete Choice Models With Choice-Based Sampling. Econometrica, 60, 1187-1214.

Kazimi, C., D. Brownstone, A. Ghosh, T.F. Golob, and D. van Amelsfort (1999). Willingness-to-Pay to Reduce Commute Time and Its Variance: Evidence from the San Diego I-15 Congestion Pricing Project. Working Paper UCI-ITS-WP-99-8, Irvine: Institute of Transportation Studies.

Keane, M. (1994). A Computationally Practical Simulation Estimator for Panel Data. Econometrica, 62, 95-116.

Koop, G. and D. J. Poirier (1993). Bayesian Analysis of Logit Models using Natural Conjugate Priors. Journal of Econometrics, 56, 323-340.

Koop, G. and D. J. Poirier (2000). Bayesian Variants of Some Classical Semiparametric Regression Techniques. University of California, Irvine, Dept. of Economics Working Paper 00-01-22, December.

Koppelman, F. S. and C-H Wen (2000). The paired combinatorial logit model: properties, estimation and application. Transportation Research B, 34, 75-89.

Lancaster, T. (1997). Bayes WESML posterior inference from choice-based samples. Journal of Econometrics, 79, 291-303.

Lee, L. (1992). On Efficiency of Methods of Simulated Moments and Maximum Simulated Likelihood Estimation of Discrete Response Models. Econometrica, 8, 518-552.

Malakoff, D. (1999). Bayes offers a “new” way to make sense of numbers. Science, 286 (Nov. 19), 1460-1464.

Manski, C.F. and S. Lerman (1977). The estimation of choice probabilities from choice-based samples. Econometrica, 45, 1977-1988.

McCulloch, R. and P.E. Rossi (1994). An Exact Likelihood Analysis of the Multinomial Probit Model. Journal of Econometrics 64, 207-240.

McFadden, D. (1973). Conditional Logit Analysis of Qualitative Choice Behavior, in P. Zarembka (ed.), Frontiers in Econometrics, New York: Academic Press.

McFadden, D. (1978). Modelling the Choice of Residential Location, in A. Karlqvist (ed.), Spatial Interaction Theory and Residential Location. Amsterdam: North-Holland, 75-96.

McFadden, D. (1989). A Method of Simulated Moments for Estimation of Discrete Response Models Without Numerical Integration. Econometrica, 57, 995-1026.

McFadden, D. and K. Train (1998). Mixed MNL models for discrete response. Department of Economics, UC Berkeley.

McFadden, D., C. Winston, and A. Boersch-Supan (1985). Joint estimation of freight transportation decisions under nonrandom sampling, in A. F. Daughety (ed.), Analytical Studies in Transport Economics. Cambridge University Press, 137-157.

Mehndiratta, S. (1996). Time-of-Day Effects in Inter-City Business Travel. Ph.D. thesis, Department of Civil Engineering, University of California, Berkeley.

Meng, X-l. and D. B. Rubin (1992). Performing Likelihood-Ratio Tests with Multiply-Imputed Data Sets. Biometrika, 79, 103-111.

Pakes, A. and D.Pollard (1989). Simulation and the Asymptotics of Optimization Estimators. Econometrica, 57, 1027-1058.

Poirier, D. J. (1988). Frequentist and Subjectivist Perspectives on the Problems of Model Building in Economics (with discussion and reply). Journal of Economic Perspectives, 2, 121-170.

Poirier, D. J. (1996). A Bayesian Analysis of Nested Logit Models. Journal of Econometrics, 75, 163-181.

Raimond, T. and D.A. Hensher (1997). A Review Of Empirical Studies And Applications. In Panels for Transportation Planning, eds. T.F. Golob, R. Kitamura and L. Long, 15-72. Boston: Kluwer Academic Publishers.

Revelt, D. and K. Train (1998). Incentives for appliance efficiency in a competitive energy environment: Random-parameters logit models of households' choices. Review of Economics and Statistics, 80, 647-657.

Revelt, D. and K. Train (1999). Customer-Specific Taste Parameters and Mixed Logit. Department of Economics, University of California, Berkeley, November.

Roberts, G. O. and A.F.M. Smith (1993). Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms. Stochastic Processes and their Applications, 49, 207-216.

Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: John Wiley.

Rubin, D. B. (1996). Multiple Imputation After 18+ Years. Journal of the American Statistical Association, 91, 473-489.

Savin, N. E. (2001). Binary Response Models: Logits, Probits, and Semiparametrics. Paper presented at the American Economic Association Annual Meetings, New Orleans, LA, January 6, 2001.

Schenker, N. and A.H. Welsh (1988). Asymptotic Results for Multiple Imputation. Annals of Statistics, 16, 1550-1566.

Small, Kenneth A (1987). "A Discrete Choice Model for Ordered Alternatives." Econometrica, 55, 409-424.

Small, K. (1992). Urban Transportation Economics. Switzerland: Harwood Academic

Publishers.

Small, K. and C. Winston (1999). The demand for transportation: models and applications. In J. A. Gomez-Ibanez, W. Tye, and C. Winston (eds.), Essays in Transportation Economics and Policy: A Handbook in Honor of John R. Meyer. Brookings Institution, Washington, D.C., 11-55.

Tanner, M. A. and W. H. Wong (1987). The calculation of posterior distributions by data augmentation (with discussion). Journal of the American Statistical Association, 82, 528-50.

Train, K. (1995). Simulation methods for probit and related models based on convenient error partitioning. Working Paper No. 95-237, Department of Economics, University of California, Berkeley.

Train, K. (1998). Unobserved taste variation in recreation demand models. Land Economics, 74(2), 230-239.

Train, K. (1999). Halton Sequences for Mixed Logit. Working paper, Department of Economics, University of California, Berkeley.

Wooldridge, J. M. (1999). Asymptotic properties of weighted M-estimators for variable probability samples. Econometrica, 67, 1385 – 1406.