Environment and Planning A 2005, volume 37, pages 525 – 550

DOI:10.1068/a36137

Willingness-to-pay estimation with mixed logit models: some new evidence

Mauricio Sillano, Juan de Dios Ortúzar¶
Department of Transport Engineering, Pontificia Universidad Católica de Chile, Casilla 306, Santiago 22, Chile; e-mail: M.Sillano@sdgworld.net, jos@ing.puc.cl
Received 25 June 2003; in revised form 3 June 2004

Abstract. Mixed-logit models are currently the state of the art in discrete-choice modelling, and their estimation in various forms (in particular, mixing revealed-preference and stated-preference data) is becoming increasingly popular. Although the theory behind these models is fairly simple, the practical problems associated with their estimation with empirical data are still relatively unknown and certainly not solved to everybody's satisfaction. In this paper we use a stated-preference dataset—previously used to derive willingness to pay for reduction in atmospheric pollution and subjective values of time—to estimate random parameter mixed logit models with different estimation methods. We use our results to discuss in some depth the problems associated with the derivation of willingness to pay with this class of models.

1 Introduction

Since the dawn of discrete-choice modelling in the 1960s, when binary logit and probit models became useful tools to derive values of time, we have come a long way—and increasingly faster in the last few years. We have seen almost three decades of unchecked rule by the multinomial (MNL) and nested logit (NL) models, with the more powerful and flexible multinomial probit (MNP) being left aside because of the difficulties involved with its use in real-life problems. Today, when computing power and better numerical techniques have made possible its use in practical applications, MNP has been overshadowed again by the equally flexible and/or powerful but less unyielding, mixed logit (ML) model. Both approaches have the ability to treat correlated and heteroscedastic alternatives, as well as random taste variations through the estimation of random rather than fixed parameters.

In this paper we discuss a number of issues related to the interpretation of results and the use of this exciting model in real-life applications. In particular, we dig deeper into the use of the model to estimate measures of willingness to pay (WTP), such as the value of time or the value of a statistical life (Rizzi and Ortúzar, 2003).

The WTP for a unit change in a certain attribute can be computed as the marginal rate of substitution (MRS) between income and the quantity expressed by the attribute, at constant utility levels (Gaudry et al, 1989). The concept is equivalent to computing the compensated variation (Small and Rosen, 1981), as one usually works with a linear approximation of the indirect utility function. Thus, point estimates of the MRS represent the slope of the utility function for the range where this approximation holds. Furthermore, as income does not enter in the truncated indirect utility function, the MRS is calculated with respect to minus the cost variable (Jara-Díaz, 1990). In this way, the WTP in a linear utility function simply equals the ratio between the parameters of the variable of interest (that is, time in the case of the subjective value of time, SVT) and the cost variable (that is, the marginal utility of income, which itself has to follow certain properties in a well-specified model).

¶ Corresponding author.

A nontrivial problem is that the estimated parameters—even in the case of models which assume fixed population coefficients, such as the MNL and NL—are random variables with a probability distribution (an asymptotic normal distribution in the MNL). Therefore, it is not clear which is the distribution of the ratio WTP (Armstrong et al, 2001). This is naturally further compounded in the case of random coefficient models, such as the ML and MNP models. Fortunately, as we will see, the fact that the ML model can yield individual-based coefficient estimates as well as estimates of the parameters governing their population distribution, helps somewhat in this quest.

The rest of this paper is organised as follows. In section 2 we present the ML model as a fairly general random utility model, and contrast it with the popular but restrictive MNL model. In section 3 we briefly explain two methods available to estimate the ML model: the classical (Revelt and Train, 2000; Train, 1998) and Bayesian approaches (see, for example, Huber and Train, 2001). In section 4 we succinctly describe a stated-preference (SP) experiment designed to obtain WTP estimates for reductions in atmospheric pollution, and present results based on the estimation of MNL models. In section 5 we analyse in some detail the results of estimating random parameter ML models with this dataset, by means both of the classical and of the Bayesian approaches. In section 6 we discuss several issues associated to WTP estimation using ML population parameters, and in section 7 do the same but for the richer case of individual-level parameters. In section 8 we present our main conclusions.

2 A flexible random utility model

An individual n facing an alternative A_j in a choice situation t, will perceive a utility level U_{njt} which is completely deterministic to him or her, and so will proceed to select that option with the greatest utility. However, the modeller is forced to assume that the Us are random variables, as otherwise we cannot explain why apparently equal individuals (that is, equal in all attributes which can be observed or measured) choose different options. A standard form for U_{njt} is

$$U_{njt} = \boldsymbol{\beta} \boldsymbol{X}_{njt} + \varepsilon_{njt} , \tag{1}$$

where $\boldsymbol{X}_{njt}$ is the observed attribute vector and $\boldsymbol{\beta}$ is a vector of marginal utility parameters. The term ε_{njt} is white noise that can, for example, usefully be assumed to distribute independently and identically (iid) Gumbel (0, σ_ε), providing the remaining unobserved variability of the model. This last assumption leads to an MNL probability function, for the individual n choosing alternative A_i in situation t:

$$\mathrm{P}(i|\boldsymbol{X}_n) = \frac{\exp(\lambda \boldsymbol{\beta} \boldsymbol{X}_{in})}{\sum_{j=1}^{J} \exp(\lambda \boldsymbol{\beta} \boldsymbol{X}_{jn})},$$

where the scale parameter λ is inversely related to the unknown standard deviation σ_ε, and in practical model applications it is standardised (that is, taken as one) as it cannot be estimated separately from the taste parameters $\boldsymbol{\beta}$ (Ortúzar and Willumsen, 2001).

The MNL formulation has important limitations, mainly because of the rigidity of its error structure (a diagonal covariance matrix with equal variances). To overcome these limitations we can restate the random utility expression for U_{njt} in a more general form:

$$U_{njt} = \boldsymbol{\beta}_n \boldsymbol{X}_{njt} + \boldsymbol{\Omega}_{njt} \boldsymbol{Y}_{njt} + \varepsilon_{njt} ,$$

where $\boldsymbol{\beta}_n$ is now the vector of marginal utility parameters for individual n; $\boldsymbol{Y}_{njt}$ is a vector of loadings that map the error components according to the desired model structure, and $\boldsymbol{\Omega}_{njt}$ is a vector of stochastic components which follow a distribution

specified by the analyst, with zero mean and unknown variance. The terms $\boldsymbol{X}_{njt}$ and ε_{njt} are the same as in equation (1).

An adequate specification of the $\boldsymbol{Y}_{njt}$ vector allows us to treat different error structures, such as heteroscedasticity, correlation, cross-correlation, dynamics, and even autoregressive error components. For good revisions and discussions of the literature related to this general form and its applications, see Hensher and Greene (2003), Train (2003), and Walker (2001).

Because the estimation of WTP values deals with parameter ratios, hereafter we will deal only with a random-parameters structure in which the marginal utility parameters are individual specific (that is, different for each sampled individual n), but the same across choice situations. This last assumption may be relaxed if choice situations are significantly separated along time, as taste parameters could then be altered. The results stated below can be extended, however, to more complex error structures, but this would just make their computation more involved. Hence, hereafter the general random utility model will be presented in a more concise form:

$$U_{njt} = \boldsymbol{\beta}_n \boldsymbol{X}_{njt} + \omega_{njt} ,$$

where $\omega_{njt} = \boldsymbol{\Omega}_{njt} \boldsymbol{Y}_{njt} + \varepsilon_{njt}$.

The terminology of random-parameter logit arises from the way in which taste heterogeneity (that is, individual taste parameters) has been treated to allow estimation (Algers et al, 1999; Revelt and Train, 1998; Train, 1998). In a departure from the popular but rigid specification of the MNL model, we can state that the model parameters are not fixed across the population but, rather, are random variables with a certain distribution specified by the analyst according to prior knowledge of the utility structure. The random-utility model may thus be written as

$$U_{njt} = \boldsymbol{X}_{njt} \boldsymbol{\beta}_n + \varepsilon_{njt} \rightarrow \boldsymbol{\beta}_n \sim \mathrm{f}(\boldsymbol{b}, \boldsymbol{\Sigma}) , \tag{2}$$

where $\boldsymbol{b}$ is the vector of population means of the parameters, and $\boldsymbol{\Sigma}$ is their covariance matrix over the population. In expression (2), each individual-level parameter is considered as a conditional draw from the frequency distribution of the population parameter. In other words, we acknowledge that every individual has a distinct set of taste coefficients and that these follow a certain frequency distribution over the population.

3 Estimation procedures

The two estimation procedures presented below yield the same type of results for two groups of parameters: (1) the mean and standard deviation of the parameter distributions over the population; and (2) individual-level marginal utility parameters.

First we briefly present the classical approach, incorporating the latest developments in the field of estimation via simulated maximum-likelihood methods (Bhat, 2001; Garrido and Silva, 2004; Train, 2003), including the framework by which population-distribution parameters combined with information from individual choices can lead to consistent estimates of individual partworths (Revelt and Train, 2000). Second, we present the hierarchical Bayes estimation procedure, which has undergone remarkable development in recent years (Allenby and Rossi, 1999; Andrews et al, 2002; Huber and Train, 2001; Lahiri and Gao, 2001; McCulloch and Rossi, 1994; Sawtooth Software, 1999).

3.1 Classical estimation

By 'classical estimation' we mean the maximum-likelihood procedure commonly used to estimate this kind of model (Train, 2003). Following standard arguments, let a person's sequence of T choices be denoted by $\boldsymbol{y}_n = (y_{1n}, ..., y_{Tn})$, where $y_{tn} = i$ if $U_{nit} > U_{njt}, \forall j \neq i$. The conditional probability of observing an individual n stating a

sequence $\boldsymbol{y}_n$ of choices, given fixed values for the model parameters $\bar{\boldsymbol{\beta}}_n$, is given by the product of logit functions:

$$\Lambda(\boldsymbol{y}_n|\boldsymbol{\beta}_n) = \prod_{t=1}^{T}\left(\frac{\exp(\lambda\bar{\boldsymbol{\beta}}_n\boldsymbol{X}_{nti})}{\sum_{j=1}^{J}\exp(\lambda\bar{\boldsymbol{\beta}}_n\boldsymbol{X}_{ntj})}\right)^{g_{nit}}, \tag{3}$$

where g_{nit} equals one if $y_{tn} = i$, and zero otherwise. Now, as $\boldsymbol{\beta}_n$ is unknown, the unconditional probability of choice is given by the integration of equation (3) weighted by the density distribution of $\boldsymbol{\beta}_n$ over the population:

$$\mathrm{P}(\boldsymbol{y}_n) = \int \Lambda(\boldsymbol{y}_n|\boldsymbol{\beta}_n)\mathrm{f}(\boldsymbol{\beta}_n|\boldsymbol{b}, \boldsymbol{\Sigma})\mathrm{d}\boldsymbol{\beta}_n,$$

where $\mathrm{f}(\bullet)$ is the multivariate distribution of $\boldsymbol{\beta}_n$ over the sampled population. If covariance terms are not specified, $\boldsymbol{\Sigma}$ is a diagonal matrix.

The log-likelihood function in $\boldsymbol{b}$ and $\boldsymbol{\Sigma}$ is given by

$$\mathrm{l}(\boldsymbol{b}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N}\ln \mathrm{P}_n(\boldsymbol{y}_n),$$

but, as the probability P_n does not have a closed form it is approximated through simulation (SP_n), where draws are taken from the mixing distribution $\mathrm{f}(\bullet)$ weighted by the logit probablity, and then averaged up (McFadden and Train, 2000). The issue of how many draws should be performed and how they should be taken is discussed below.

The simulated log-likelihood function is given by

$$\mathrm{sl}(\boldsymbol{b}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N}\ln \mathrm{SP}_n(\boldsymbol{y}_n). \tag{4}$$

Conveniently, the simulator for the choice probabilities is smooth and unbiased. Different forms of 'smart' drawing techniques (that is, Halton and other low-discrepancy sequences, antithetic, quasi-random sampling, etc) can be used to reduce the simulation variance and to improve the efficiency of the estimation (Bhat, 2001; Garrido and Silva, 2004; Hajivassiliou and Ruud, 1994; Hensher and Greene, 2003).

Numerical procedures are used to find maximum-likelihood estimators for $\boldsymbol{b}$ and $\boldsymbol{\Sigma}$. These parameters define a frequency distribution for the $\boldsymbol{\beta}_n$ over the population. To obtain actual point estimates for each $\boldsymbol{\beta}_n$ a second procedure, described by Revelt and Train (2000), is required as follows.

The conditional density $\mathrm{h}(\boldsymbol{\beta}_n|\boldsymbol{y}_n, \boldsymbol{b}, \boldsymbol{\Sigma})$ of any $\boldsymbol{\beta}_n$ given a sequence of T_n choices $\boldsymbol{y}_n$ and the population parameters $\boldsymbol{b}$ and $\boldsymbol{\Sigma}$, may be expressed by Bayes's rule as

$$\mathrm{h}(\boldsymbol{\beta}_n|\boldsymbol{y}_n, \boldsymbol{b}, \boldsymbol{\Sigma}) = \frac{\mathrm{P}_n(\boldsymbol{y}_n|\boldsymbol{\beta}_n)\mathrm{f}(\boldsymbol{\beta}_n|\boldsymbol{b}, \boldsymbol{\Sigma})}{\mathrm{P}_n(\boldsymbol{y}_n|\boldsymbol{b}, \boldsymbol{\Sigma})}. \tag{5}$$

The conditional expectation of $\boldsymbol{\beta}_n$ results from integrating over the domain of $\boldsymbol{\beta}_n$. This integral can be approximated by simulation, averaging weighted draws $\boldsymbol{\beta}_n^r$ from the population-density function $\mathrm{f}(\boldsymbol{\beta}_n|\boldsymbol{b}, \boldsymbol{\Sigma})$. The simulated expectation SE is given by

$$\mathrm{SE}(\boldsymbol{\beta}_n|\boldsymbol{y}_n, \boldsymbol{b}, \boldsymbol{\Sigma}) = \frac{\sum_{r=1}^{R}\boldsymbol{\beta}_n^r\mathrm{P}_n(\boldsymbol{y}_n|\boldsymbol{\beta}_n^r)}{\sum_{r=1}^{R}\mathrm{P}_n(\boldsymbol{y}_n|\boldsymbol{\beta}_n^r)}.$$

Revelt and Train (2000) also propose, but do not apply, an alternative simulation method to condition individual-level choices. Consider the expression for $h(\boldsymbol{\beta}_n | \boldsymbol{y}_n, \boldsymbol{b}, \boldsymbol{\Sigma})$ in equation (5). The denominator is a constant value as it does not involve $\boldsymbol{\beta}_n$, so a proportionality relation can be established as

$$h(\boldsymbol{\beta}_n | \boldsymbol{y}_n, \boldsymbol{b}, \boldsymbol{\Sigma}) \propto P_n(\boldsymbol{y}_n | \boldsymbol{\beta}_n) f(\boldsymbol{\beta}_n | \boldsymbol{b}, \boldsymbol{\Sigma}) .$$

Draws from the posterior $h(\boldsymbol{\beta}_n | \boldsymbol{y}_n, \boldsymbol{b}, \boldsymbol{\Sigma})$ can then be obtained using the Metropolis–Hastings algorithm (Chib and Greenberg, 1995), with successive iterations improving the fit of the $\boldsymbol{\beta}_n$ to the observed individual choices. During this process, the prior $f(\boldsymbol{\beta}_n | \boldsymbol{b}, \boldsymbol{\Sigma})$, that is, the parameter distribution obtained by maximum likelihood, remains fixed; it provides information about the population distribution of $\boldsymbol{\beta}_n$. After a number of burnout iterations to ensure that a steady state has been reached [typically, a few thousands (Kass et al, 1998)], only one of every m of the sampled values generated is stored to avoid potential correlation among them; m is obtained as a result of the convergence analysis (Raftery and Lewis, 1992). From these values a sampling distribution for $h(\boldsymbol{\beta}_n | \boldsymbol{y}_n, \boldsymbol{b}, \boldsymbol{\Sigma})$ can be built, and inferences about the mean and standard deviation values can be obtained (Arora et al, 1998; Sawtooth Software, 1999). In this paper we favoured this last procedure for implementation purposes.(1)

The outcome of the estimation process is two sets of parameters: $\boldsymbol{b}$ and $\boldsymbol{\Sigma}$, the population parameters obtained by simulated maximum likelihood and $\boldsymbol{\beta}_n$, the individual parameters for $n = 1, ..., N$, estimated via conditioning the observed individual choices on the estimated population parameters.

3.2 Bayesian estimation

Use of the Bayesian statistic paradigm for the estimation of ML models has gained much interest in recent years (Huber and Train, 2001; Sawtooth Software, 1999; Train, 2001). The ability to estimate individual partworths appeared initially as its main appeal, but it has shown further advantages with respect to the estimation procedure. The Bayesian approach considers the parameters as stochastic variables so, applying Bayes's rule of conditional probability, a posterior distribution for $\boldsymbol{\beta}_n$ conditional on observed data and prior beliefs about these parameters can be estimated.

Let $\Psi(\boldsymbol{b}, \boldsymbol{\Sigma})$ be the analyst's prior knowledge about the distribution of $\boldsymbol{b}$ and $\boldsymbol{\Sigma}$, and consider a likelihood function for the observed sequence of choices conditional on fixed values of $\boldsymbol{b}$ and $\boldsymbol{\Sigma}$. By Bayes's rule, the posterior distribution for $\boldsymbol{\beta}_n$, $\boldsymbol{b}$,and $\boldsymbol{\Sigma}$ is proportional to

$$\prod_{n=1}^{N} \Lambda(\boldsymbol{y}_n | \boldsymbol{\beta}_n) f(\boldsymbol{\beta}_n | \boldsymbol{b}, \boldsymbol{\Sigma}) \Psi(\boldsymbol{b}, \boldsymbol{\Sigma}) .$$

Draws for $\boldsymbol{b}$ and $\boldsymbol{\Sigma}$ can be obtained by use of Gibbs sampling, and draws for $\boldsymbol{\beta}_n$ are taken by means of the Metropolis–Hastings algorithm; a detailed sequential procedure has been described by Sawtooth Software (1999). A crucial element that has not been mentioned in recent applications is the need to test for the convergence of the series and lack of correlation among the steady-state regime values (Cowles and Carlin, 1996).

Train (2001) discusses how the posterior means from the Bayesian estimation can be analysed from a classical perspective. This is thanks to the Bernstein–von Mises theorem, which states that, asymptotically, the posterior distribution of a Bayesian

(1) The approach was coded in WinBUGS, a software package developed by the MRC Biostatistics Unit at the University of Cambridge and the Imperial College School of Medicine at St Mary's, London. The program is free for downloading from their website: http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml

estimator converges to a normal distribution which is the same as the asymptotic distribution of the maximum-likelihood estimator (that is, the standard deviation of the posterior distribution of the Bayesian estimator can be taken as the classical standard error of a maximum-likelihood estimator). This means that classical statistical analysis (for example, the construction of t-statistics to analyse the significance of an estimated parameter) can be performed on Bayesian estimators without compromising the interpretation of the results.

Bayesian estimation has certain advantages over the classical approach.

(a) No numerical maximisation routines are necessary; rather, draws from the posterior distribution are taken until convergence is achieved.

(b) As the number of attributes considered in the utility expression grows, the number of elements in the covariance matrix $\boldsymbol{\Sigma}$ rises exponentially, increasing computation time in the classical approach. However, the Bayesian method can handle a full covariance matrix almost as easily as a restricted one, with computation time rising only with the number of parameters.

(c) Identification issues are related to the lack of orthogonality in the effects of the random variables, and not to the number of independent equations representing these. This means that an identification problem may arise when the effect of a certain variable in the structural utility formulation is confused with the effect of another variable, but not because of insufficient sample points.

The Bayesian estimation procedure was also implemented in WinBUGS (Spiegelhalter et al, 2001). This package incorporates Gibbs sampling protocols and the Metropolis–Hastings sampling algorithm[2], but lacks a convergence analysis, which has to be performed separately. Both estimation procedures were applied to the atmospheric pollution reduction valuation stated-preferences experiment described below, and the main results compared.

4 The stated-preference experiment

A residential-location-based stated-preference experiment was undertaken in Santiago to assess the valuation of atmospheric-nuisance reductions. A full description of the microeconomic formulation, survey design, and main results derived from MNL models has been presented (Ortúzar and Rodriguez, 2002). It is important to mention that a great deal of effort was spent in defining an air-pollution attribute that would be understandable and representative. The number of days per year with an alert status associated with the air quality of a particular dwelling zone was eventually selected (DA—days of alert). Other attributes considered in the formulation were: travel time to work (TTW); travel time to study (TTS); and a dummy variable (δCURRENT) that attempted to capture the inertia effect associated with the current residential location.

Selected households[3] were asked to rank once ten alternative residential locations (that is, nine arising from a fractional factorial design considering only main effects, plus their current location).[4] One beauty of the exercise was the family discussion about alternative locations and, as a result, the serious intent with which the ranking task was performed. The rank data were later converted into nine independent-choice situations per household, as is common in this type of study. After lexicographic and

[2] A GAUSS code for Bayesian estimation written by Kenneth Train was also tried out, but not used in the final estimations. We are grateful to him for sharing his code.

[3] Families renting a flat who had moved to their dwelling during the previous year; the idea was that these people would find a residential-location-based stated-preference experiment easier to handle.

[4] The survey design was a straightforward application of agreed stated-preference principles (Louviere et al, 2000).

Table 1. Estimation results (with *t*-statistics shown in parentheses), excluding lexicographic and inconsistent responses.

Attribute	Parameter
Travel time to work (TTW) (minutes per week)	−0.00417 (−10.6)
Travel time to study (TTS) (minutes per week)	−0.00250 (−7.8)
Days of alert (DA) (days per year)	−0.27370 (−11.0)
Rent (RENT) (10^3 Ch $ a month)	−0.02641 (−12.5)
δCURRENT	0.89690 (5.9)
Log-likelihood	−849.6

inconsistent responses had been excluded in the usual way, the sample size consisted of 648 observations from a total of 75 households. Maximum-likelihood estimation results for a MNL model are shown in table 1.

All parameters are significant and have the expected sign. WTP values calculated as the ratio of all attribute parameters and the RENT coefficient[5] are shown in table 2, together with confidence intervals calculated according to Armstrong et al (2001).

It is worth mentioning that the subjective values of time below are in close agreement with values estimated both previously and afterwards in the country for rather different settings (Galilea and Ortúzar, 2004; Iragüen and Ortúzar, 2004; Ortúzar et al, 2000; Pérez et al, 2003). This gives us great confidence about the quality of the data used.

Table 2. Point estimates, with 95% confidence intervals shown in parentheses, for subjective valuation of attributes.

Attribute	Subjective value
TTW (Ch$ per minute)	36 (29 – 45)
TTS (Ch$ per minute)	22 (16 – 28)
DA (Ch$ per days of alert per year)	124 362 (100 818 – 152 301)

5 Estimation of random-parameter logit

5.1 Classical estimation: population parameters

As we saw above, the classical approach to the estimation of ML models has two stages. In the first, simulated maximum likelihood yields estimates of the population distribution of the parameters. Estimation results for a ML model with iid normal parameters are presented in table 3 (over), together with the previous MNL-model results for the sake of comparison.[6] In this model, ML1, the nine choices from each household were considered, correctly, as repeated-choice observations. Although the rank transformation assumed independent choices for each family, nevertheless, the whole set is correlated in relation to the choices made by other households.

[5] To obtain the subjective value of time figures (Ch$ per minute), the ratios of the parameters of time and rent were multiplied by the factor (12/52) 1000. To obtain the WTP for the DA attribute (Ch$ per day), the ratio of the parameters DA and RENT was multiplied by 12 000. At the time of the survey 1US $ = 490 Ch$.

[6] Maximum-likelihood estimation was conducted with the aid of a GAUSS code written by Train, Revelt, and Ruud at the University of Berkeley, CA. The code is available for downloading at Kenneth Train's web page: http://elsa.berkeley.edu/~train. We tested using multivariate normally distributed parameters but found nonsignificant covariances; so, for reasons of computing time saving, we stuck to independent distributions.

Table 3. Multinomial (MNL) and mixed logit (ML) model results (with *t*-statistics shown in parentheses) with four independently and identically distributed normal distribution parameters and one fixed.

Attribute		Parameter	
		MNL	ML1
TTW	mean	−0.00417 (−10.6)	−0.009924 (−7.9)
	standard deviation		0.005734 (4.5)
TTS	mean	−0.00250 (−7.8)	−0.005769 (−8.2)
	standard deviation		0.002656 (2.7)
DA	mean	−0.27370 (−11.0)	−0.478625 (−6.8)
	standard deviation		0.405665 (4.7)
RENT	mean	−0.02641 (−12.5)	−0.057396 (−7.0)
	standard deviation		0.047482 (6.2)
δCURRENT	mean	0.89690 (5.9)	1.053245 (5.5)
Log-likelihood		−849.6	−747.0

The inertia parameter (δCURRENT) was originally considered to vary over the population, but its estimated deviation was statistically negligible (*t*-test 0.66), so in the final estimation round it was considered fixed.

The estimation of an ML model results in a substantial improvement of fit over the MNL model, which is a common result in mixed-logit applications (Hensher, 2001a; Train, 1998) because of the increased explanatory power of the specification. However, attention must be paid to the ML model results as unwelcome effects may arise from its unconstrained formulation. The frequency distribution of the parameters over the population accounts for taste variations and unobserved heterogeneity, and this has been proven to exist beyond socioeconomic characterisation (Iragüen and Ortúzar, 2004; Morey and Rossman, 2002; Ortúzar et al, 2002; Rizzi and Ortúzar, 2004). However, a normally distributed parameter will yield individual values with both negative and positive signs, as its domain covers all real values. This means that implausible positive values for the RENT, TTW, TTS, and DA parameters could be obtained for some observations.

In fact, the portion of the population for which the model assigns an incorrect parameter sign can be estimated as the cumulative mass function of the frequency distribution of the parameter over the population evaluated at zero (that is, for supposedly negative parameters, the area under the frequency curve between zero and positive infinity). In this case, model ML1 would account for 4% of the population having positive TTW parameters, 1% of the population having positive TTS parameters, 12% of the population having positive DA parameters, and 11% of the population having positive RENT parameters. Although this problem may be overcome in various ways most of these methods introduce further problems; hence, the issue is discussed in more depth below.[7]

Another significant effect of the ML model is the considerably larger mean values for the attribute parameters compared with those in the MNL model. This stems from the fact that the ML model decomposes the unobserved component of utility and normalises the parameters through the scale factor λ.

[7] For example, by the use of a log-normal distribution, but this is not the only way to constrain parameter estimates to a positive domain. One could define other distributions and truncate them to the positive range. Furthermore, the log-normal carries undesirable effects, such as a biased mean value caused by its long tail. The distribution is discussed below to keep consistency with other studies cited here, but note that recent research discusses the application of a truncated normal distribution in a ML model estimation with Bayes (Train and Sonnier, 2003).

Assume a utility structure given by

$$U_{njt} = \boldsymbol{\beta} \boldsymbol{X}_{njt} + \varepsilon_{njt},$$

which is a standard MNL specification. The variance of U is a result of the iid Gumbel term, and is computed as

$$\text{var}(U_{njt}) = \text{var}(\varepsilon_{njt}) = \frac{\pi^2}{6\lambda^2}.$$

However, part of the variance is treated explicitly as a separate error component in the ML model:

$$U_{njt} = (\boldsymbol{b} + \boldsymbol{s}\boldsymbol{\gamma}_n)\boldsymbol{X}_{njt} + \varepsilon_{njt},$$

where $\boldsymbol{s}$ is the vector of standard deviations of the model parameters over the population, $\boldsymbol{b}$ is the vector of means and $\boldsymbol{\gamma}_n$ is a vector of standard random perturbations that may be distributed, for example, normally. In this case the variance of U would be computed as

$$\text{var}(U_{njt}) = \boldsymbol{s}^2 \boldsymbol{X}_{njt}^2 + \text{var}(\hat{\varepsilon}_{njt}) = \boldsymbol{s}^2 \boldsymbol{X}_{njt}^2 + \frac{\pi^2}{6\hat{\lambda}^2},$$

and because the model variance is independent of its specification (it depends exclusively on the data), it is easy to see that the scale parameter $\hat{\lambda}$ ($\hat{\lambda} > \lambda$), means larger mean parameter estimates for the ML model.

This issue deserves special consideration, particularly when model estimates are used for valuation purposes as the rescaling process may result in mean estimates that are relatively higher for some attributes than for others. For example, the rescaling of model ML1 relative to the MNL model yields enlargement factors that range from 1.17 (for the δCURRENT parameter) to 2.4 (for the TTW parameter), determining different directions of change for the parameter ratios of model ML1 relative to MNL.

Then, the rescaling effect is driven by the reduced unobserved variance, but a different mechanism determines the uneven nature of the enlargement factors for each parameter. An intuitive explanation for this would be that the explicit treatment of parameter variation over the population into the systematic utility portion is equivalent to the incorporation of an explanatory variable previously left out in the original (MNL) model. This is analogous to one of the misspecification problems discussed by Horowitz (1981), and would lead to the restructuring of the utility parameters to compensate for the extra explanation accounted for.[8] In any case, the point is that the direction of parameter rescaling relative to the MNL model has to be considered a potential source of model misspecification, just as the omission of explanatory variables is. As will be shown below, this issue may have some repercussions on WTP estimations.

5.2 Classical estimation: individual-level parameters

Individual-level parameters were calculated using the simulated maximum likelihood estimates and conditioning them with individual household choices, as shown in section 3. Frequency charts for the 75 individual household parameters are shown in figure 1 (over). The charts reflect the actual frequency distribution of the parameters over the population, which is in fact discrete as the sample has only a finite number of 'individuals' (that is, 75 households). The figures show that given the sample size, the frequency distributions do not resemble smooth normal distributions, as we assumed

[8] Comments by Joan Walker on this issue are greatly appreciated.

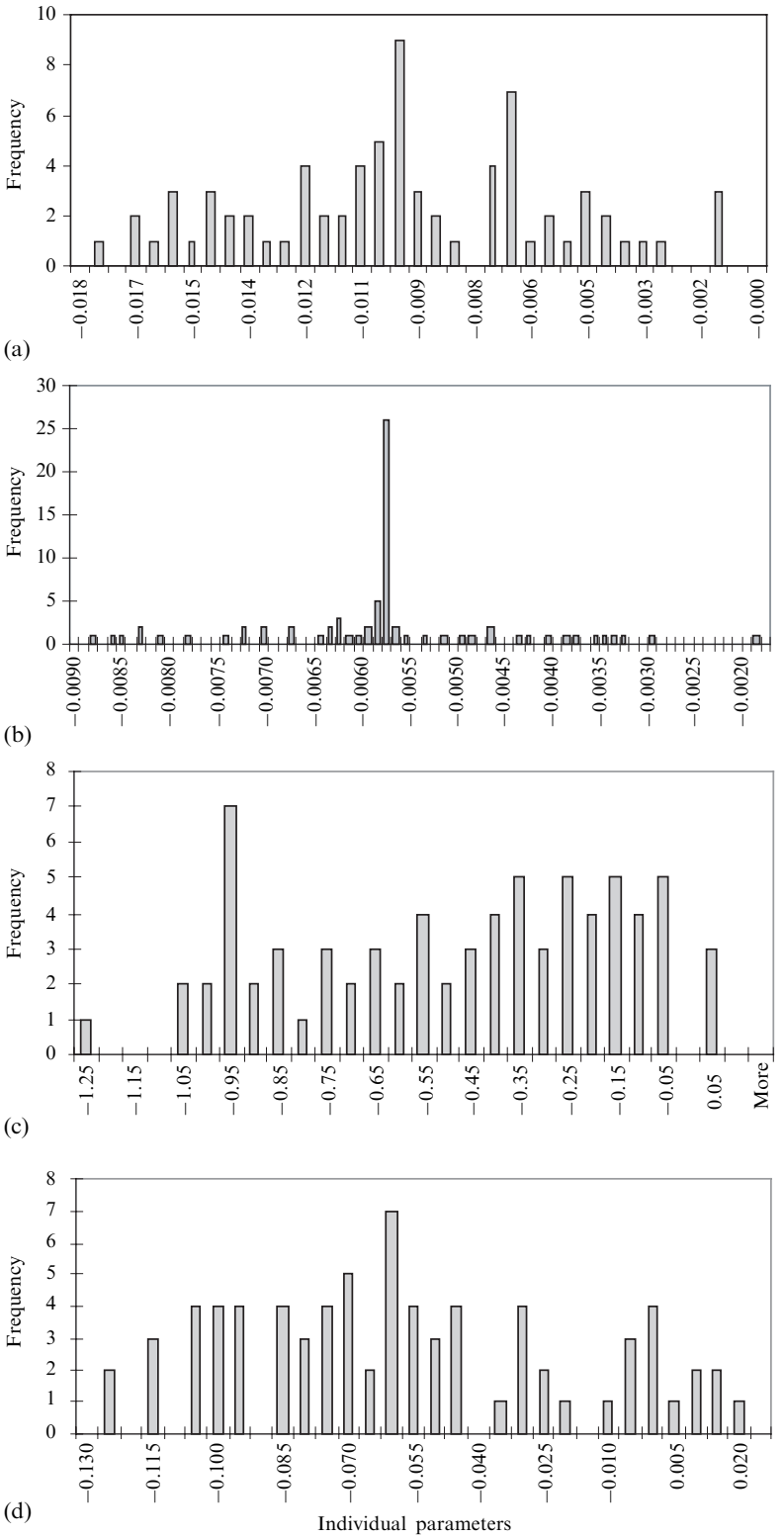

Figure 1. Histograms of (a) TTW, (b) TTS, (c) DA, (d) RENT point estimates for sampled population.

for the estimation process. This means that a certain amount of error must be expected when analysing a discrete set of values using a continuous distribution.

The distributions of household parameters reveal that only small percentages of the sample have values with theoretically incorrect signs. In fact, it is obvious that the previously calculated values for the expected percentages with wrong sign were overestimated. The actual percentages with incorrect sign are given in table 4, along with the original estimations; the actual values were computed simply by counting individual cases with a theoretically incorrect sign (that is, a positive value). The original values had been calculated as the cumulative mass function evaluated from zero to positive infinity.

Table 4. Percentage of individual parameters with incorrect sign.

Attribute	Percentage with incorrect sign	
	population distribution	individual parameters
TTW	4	0
TTS	1	0
DA	12	4
RENT	11	8

The percentages for the DA and RENT parameters correspond to just three and six sampled households, respectively. Furthermore, in all nine cases the *t*-tests results of the individual parameters were below one, suggesting that the incorrect-sign parameters were not statistically significant. Hence they could be considered as null values for those exclusive households, and the sign assumptions could be maintained.

This finding has another consequence worth noticing: the parameter signs were basically correct even when an unconstrained (normal) distribution was imposed on them. This could be considered a case-specific situation, but it suggests that forcing the parameters to follow a log-normal distribution, for example, may not be necessary and hence a potential problem with that function could be avoided. As is known, log-normal distributions tend to produce likelihood functions that are extremely flat around the maximum, making convergence hard to achieve (Algers et al, 1999; Hensher and Greene, 2003).

An interesting result arises if we evaluate the log-likelihood function for the individual-level parameters instead of the population parameters [that is, the log-likelihood value calculated at convergence with equation (4)]. In this case the log-likelihood value for the estimated model shows a substantial improvement in fit: from -747.0 for the log-likelihood based on population parameters, to -512.9 for the value calculated from individual-level coefficients. This is not a surprise, as the individual-level parameters characterise the log-likelihood function more precisely than do the mean and standard deviation of the population, resembling more accurately the observed household choices.

5.3 Bayesian estimation

In this case, the use of the combination of Gibbs sampling and the Metropolis–Hastings algorithm leads to the simultaneous estimation of the two sets of parameters described above (population and individual-based parameters). The first set, which are the comparable ones, is presented in table 5 (over) (model ML2) together with those of model ML1. Although the values are similar, the ML2 parameters are larger in magnitude but again not in a constant scale. It is worth noting that for larger samples (that is, samples of more than 300 individuals, or around 3000 observations) we have

Table 5. Hierarchical Bayes and maximum-likelihood estimators for mixed logit (ML) model population parameters, with *t*-statistics shown in parentheses.

Attribute		Parameter	
		ML1	ML2
TTW	mean	−0.009924 (−7.9)	−0.01141 (−6.7)
	standard deviation	0.005734 (4.5)	0.01133 (8.2)
TTS	mean	−0.005769 (−8.2)	−0.00783 (−4.6)
	standard deviation	0.002656 (2.7)	0.01025 (7.5)
DA	mean	−0.478625 (−6.8)	−0.56960 (−8.3)
	standard deviation	0.405665 (4.7)	0.46920 (7.0)
RENT	mean	−0.057396 (−7.0)	−0.06974 (−8.9)
	standard deviation	0.047482 (6.2)	0.05339 (5.6)
δCURRENT	mean	1.053245 (5.5)	1.16800 (5.8)
Log-likelihood		−747.0	−474.9

found closer differences in scale between Bayesian and classical estimates, as have other analysts (Huber and Train, 2001).

The observed substantial improvement in fit relates to the fact that the Bayesian log-likelihood function, unlike that for the classical approach, is constructed as the summation of the logarithm of the individually calculated choice probabilities with their actual individual parameters, and not with averaged simulated probabilities.[9] So, even though both values essentially express the fit of the model to the data, they are calculated differently, and cannot be compared directly. To obtain a value equivalent to the classically obtained log-likelihood from the Bayesian procedure, we inserted the Bayesian estimates as initial values in the maximum-likelihood procedure; in this way we aimed to get a simulation of the choice probabilities based on the Bayesian solution. The log-likelihood value for the Bayesian estimates was −769.0 (that is, worse than the classical value), and the process later converged to −747.0, showing that the maximum-likelihood procedure was invariantly reaching a global maximum.[10]

On the other hand, if we compute the log-likelihood value as the multiplication of the logarithms of individual choice probabilities (based on individual parameters), there is a discrepancy between classically obtained individual values and the Bayesian results. The 'classical' values yield a log-likelihood of −512.9, whereas the Bayesian results give the value of −474.9.

To sum up, classical estimation yields better results in terms of fit for population parameters, whereas the Bayesian procedure appears to be considerably more powerful for individual-level parameters (at least for a small sample). This result has an intuitive explanation: the maximum-likelihood procedure seeks a mean value that best represents the choices of the sampled population, plus a dispersion value that emulates the variability around this mean. On the other hand, the Bayesian procedure is aimed directly at satisfying the choices of each sampled person, and the population parameters are estimated taking this into account. In the classical approach the fact that the sampled population is finite and discrete is conveniently forgotten for the sake of simplicity, and the individual-level models are conditioned from parameters of an infinite population.

Now we move to consider the second set of estimated parameters: the individual $\boldsymbol{\beta}_n$. Frequency distributions for these values are plotted in figure 2. Again, the actual

[9] The simulated probabilities are also individual based, but they are random outcomes which bear no relation to the information provided by each household.

[10] We are grateful to Kenneth Train for having pointed this out to us.

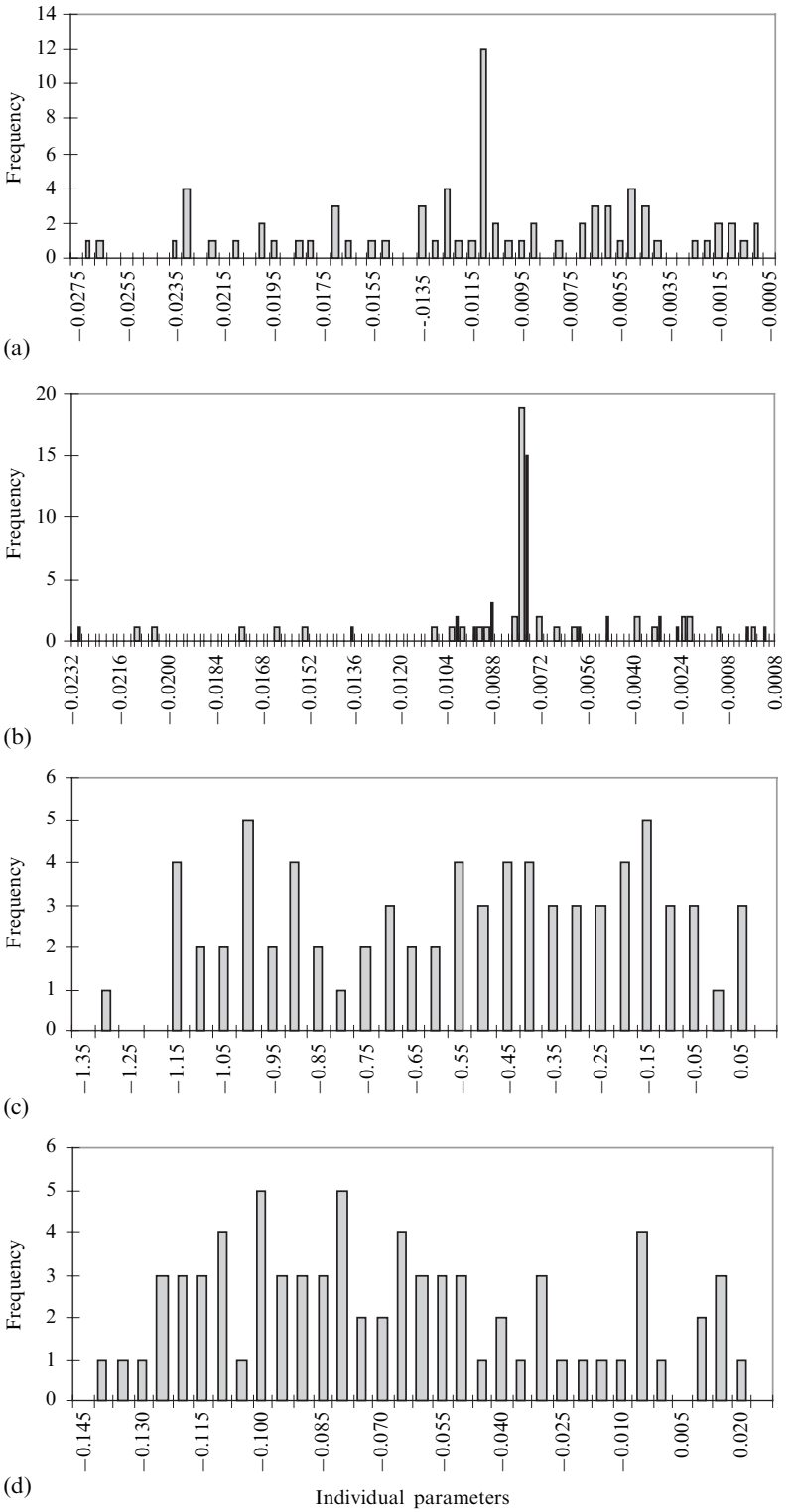

Figure 2. Histograms of (a) TTW, (b) TTS, (c) DA, (d) RENT individual point estimates for sampled population.

percentages of households with wrong-sign parameters are significantly lower than the proportions estimated according to the distribution mass defined by the population parameters. This comparison is shown in table 6.

The incorrect-sign percentage for the case of TTS corresponds to a single household which had a completely insignificant parameter ($t = 0.02$). Those for the DA and RENT parameters correspond to the same households who received a wrong-sign parameter in the classical estimation, and the values were again not significantly different from zero.

Table 6. Percentage of individual parameters with incorrect sign.

Attribute	Percentage with incorrect sign	
	population distribution	individual parameters
TTW	15.7	0.0
TTS	22.2	1.3
DA	11.2	4.0
RENT	9.6	8.0

5.4 Comparison

In previous literature it has been maintained that the two approaches—classical and Bayesian—lead to equivalent and similar results (Huber and Train, 2001; Revelt and Train, 2000). As the two estimation methods are similar in spirit (that is, they share the same behavioural assumptions), the comparative advantages to the analyst (that is, ease of implementation and analysis) must be taken into account in deciding the preferred procedure.

To compare the approaches overcoming their scale problem, a correlation analysis of both sets of individual-level parameters was conducted (table 7). The results suggest that both procedures explain the variability of the coefficients over the population in a fairly similar way.

Table 7. Correlation between parameters estimated through classical and Bayesian techniques.

Attribute	Correlation
TTW	0.968
TTS	0.900
DA	0.996
RENT	0.985

In the end, we found the Bayesian approach preferable for the following reasons.

(a) Implementation of the estimation procedures was easier in WinBUGS, as it incorporates both the Gibbs sampler and the Metropolis–Hastings algorithm as internal functions.

(b) As Bayesian methods do not involve maximisation procedures, the problem of multiple solutions (that is, the case of the ML log-likelihood function) is eliminated and, with a sufficiently high number of simulations, convergence is assured.

(c) Bayesian methods are known to work well even with small samples (Lenk et al, 1999); this was also the case here, as evidenced by the substantially better fit of the model estimated through hierarchical Bayes. Therefore, the Bayesian estimates have to be considered more reliable.

6 Estimation of willingness to pay from population parameters

Most ML applications have been limited to estimating the first set of parameters discussed above: the mean and spread of the distribution of population parameters over the sample. In general, the estimation of WTP values involves taking ratios of stochastic variables, even for models with fixed coefficients (Armstrong et al, 2001). In the ML case this problem is compounded by the fact that not only the estimates but also the parameters themselves are random variables, and this is not a trivial issue (Meijer and Rouwendal, 2000).

In this section we discuss some econometric aspects of four different methods which may be used to achieve WTP estimates from the parameter distributions.[11] Although the methods can be applied to jointly distributed parameters, in this case only independent distributions were used. However, we checked that the results were indeed coincidental.

6.1 Ratios of population means

The simplest way to derive WTP values is to take the ratio of the means of the parameter distributions involved. In other words, if

$$\theta_u \sim \mathrm{f}(\mu_u, \sigma_u) \wedge \theta_d \sim \mathrm{g}(\mu_d, \sigma_d),$$

then

$$\frac{\theta_u}{\theta_d} \rightarrow \frac{\mu_u}{\mu_d}.$$

This is not the mean value of the WTP, but a WTP value derived from the coefficients of the 'average individual' for each parameter. Therefore, this interpretation should not be used in cost–benefit analysis, and the calculation of this index may only be used as a means of testing model specification.

The ratios of population means for the ML2 model are presented in table 8, along with the MNL model estimates. In this case all the distributions are normal. Confidence intervals for the ratios were again calculated using the *t*-test formula proposed by Armstrong et al (2001).

In previous sections we opined that the ML parameters had a tendency to grow in magnitude over the MNL parameters. Nevertheless, the parameter ratios tend to be quite stable. In fact, the parameter ratios for models MNL and ML2 lie within each other's confidence intervals, except for the WTP value for DA reductions, where the mean value for model ML2 is not included in the interval for model MNL. Even though the opposite does happen, a distinction has to be made and the analyst should acknowledge the superior explanatory power of the ML model specification in order to select a final WTP value.

Table 8. Willingness to pay (WTP) as ratio of population means, with confidence intervals shown in parentheses, for multinomial (MNL) and mixed logit (ML) models.

Attribute	Willingness to pay	
	MNL	ML2
TTW (Ch$ per minute)	36 (29–45)	37 (25–54)
TTS (Ch$ per minute)	22 (16–28)	25 (14–40)
DA (Ch$ per days of alert per year)	124 362 (100 818–152 301)	98 009 (70 127–135 793)

[11] We are indebted to Kenneth Train for proposing the ideas that gave birth to this discussion. However, any errors in our arguments are our sole responsibility.

In addition, it is worth recalling that, as the method disregards the rest of the distribution, it considers a unique value for the parameters—neglecting all information about heterogeneity in the population. In the end, the model is treated almost as an MNL model; in some ways, making the extra estimation effort worthless.

6.2 Simulation

This method has been applied in the past to construct confidence intervals (Armstrong et al, 2001; Ettema et al, 1997), and has been used to derive WTP values from ML models by Hensher and Greene (2003) and Espino et al (2004). It is a first approach to construct a WTP distribution over the population with the use of information neglected by the previous method.

In this method, random draws for each parameter are taken from its distribution and their ratio is computed. This is repeated a large number of times, allowing frequencies to be computed sampling the WTP distribution. Mean and standard-deviation values can then be inferred, as well as cumulative values from the resulting distribution. An important feature of this method is that no assumptions are needed about the resulting distribution of the parameter ratios. In particular, the ratio of two normally distributed variables may turn out to be an unstable distribution (Meijer and Rouwendal, 2000). For example, the ratio of two standard normal distributions is a Cauchy distribution, and for this the first two moments cannot be estimated analytically. The ratio of independent multivariate normal distributions has been studied by Fieller (1932) and Hinkley (1969).

The simulation results for the WTP distribution derived from the population parameters of model ML2 are shown in table 9. Confidence intervals for the mean value of the resulting distribution cannot be computed in this case. The standard errors used for computing the confidence intervals in table 8 correspond to the standard deviations of the asymptotic distributions of the estimators, which are normally distributed, and yield boundaries where the ratios of means lie within a 95% confidence level. The standard-deviation values presented in table 9 are indicators of the variance of the parameter ratios over the simulated population; the construction of a confidence interval from these values would yield boundaries within which the parameter ratios of, say, 95% of the population lie.

As can be seen, the spread of the distributions is extremely large. This is related to the fact that the simulation process involves drawing values that may be close to zero. When these correspond to the RENT parameter, the ratio tends to infinity yielding inordinately large WTP values. To overcome such inconveniently extreme values (both positive and negative), small and equal percentages were cut off from each tail of the sampled distribution: 1% off each tail in WTP for both TTW and TTS reduction distributions, and 3% off each tail in the WTP for DA reduction distribution.

Table 9. Simulated willingness-to-pay (WTP) distributions in multinomial (MNL) and mixed logit (ML) models.

Attribute		Simulated WTP	
		MNL	ML2
TTW	mean	36	36
(Ch$ per minute)	standard deviation		134.6
TTS	mean	22	26
(Ch$ per minute)	standard deviation		20.8
DA	mean	124 362	94 774
(Ch$ per days of alert per year)	standard deviation		161 280

Hensher and Greene (2003) discuss the effect of removing parts of the simulated distributions of WTP, and compare this action with constraining the distributions. But in relation to the validity of this method, the real issue is not whether, or how, to constrain the distribution to make it theoretically correct. Hensher and Greene (2003) acknowledge that the mere fact of applying statistic distributions—which are already analytical constructs—to behavioural parameters governed by an unknown logic makes constraining (or removing parts of) the parameters or WTP distributions no better and no worse than an unconstrained distribution, unless there is an underlying theoretical rationale.

A consistent rationale for cutting off the tails of the distributions is the following: there are no *real* people with such extreme values to fill in the tails we are removing. In fact, much larger percentages should be taken off each tail for the simulated WTP distribution to be plausible—maybe even 20% or 30%. So, when applying this method, the analyst must remember that the final goal is to estimate WTP values for the sampled population, and for sample sizes smaller than infinity this is a finite set of values. Therefore, the real problem with the simulation of WTP distributions from sampled values is not how to constrain them in a correct way but, rather, the fact that we are simulating countless numbers of values for people who do not even exist.

6.3 Log-normal distribution for WTP

The use of log-normal distributions for parameters over the population has been proposed by many authors. This would constrain their signs to be consistent and would yield an analytical expression for the resulting WTP distribution, as the ratio of two log-normal distributed variables is also log-normally distributed.

Consider a random variable x such that $x \sim \mathrm{N}(\mu_x, \sigma_x)$. Then a variable defined as $X = \exp(x)$ has a log-normal distribution with mean $\exp(\mu_x + \sigma_x^2/2)$, and standard deviation given by $\exp(\mu_x + \sigma_x^2)/2[\exp(\sigma_x^2) - 1]^{1/2}$. Now consider the ratio of two log-normal variables, say X/Y, then:

$$\frac{X}{Y} = \frac{\exp(x)}{\exp(y)} = \exp(x - y) = \mathrm{WTP},$$

where

$$\mathrm{WTP} \sim \ln \mathrm{N}\left\{ \exp\left(\mu_{wtp} + \frac{\sigma_{wtp}^2}{2}\right), \exp\left(\mu_{wtp} + \frac{\sigma_{wtp}^2}{2}\right)[\exp(\sigma_{wtp}^2) - 1]^{(1/2)} \right\}. \tag{6}$$

As x and y are normally distributed variables, their difference is also normally distributed with

$$(x - y) \sim \mathrm{N}(\mu_x - \mu_y, \sigma_x^2 + \sigma_y^2 - 2\sigma_{xy}).$$

As we are dealing only with independent parameters, in this case the covariance term disappears. Then, replacing the above expression in equation (6) we get an expression for the log-normal WTP distribution:

$$\mathrm{WTP} \sim \ln \mathrm{N}\left\{ \exp\left[(\mu_x - \mu_y) + \frac{\sigma_x^2 + \sigma_y^2}{2}\right], \exp\left[(\mu_x - \mu_y) + \frac{\sigma_x^2 + \sigma_y^2}{2}\right][\exp(\sigma_x^2 + \sigma_y^2) - 1]^{1/2} \right\}. \tag{7}$$

Expression (7) can be used to calculate cumulative proportions and confidence intervals. Table 10 (over) presents the results of an ML model (ML3-log) where all taste coefficients were specified as log-normal except for the δCURRENT parameter

Table 10. Log-normal distributed parameters (with *t*-statistics shown in parentheses) mixed logit (ML) models.

Attribute		Parameter	
		ML3-log	ML3
TTW	mean	0.010189 (5.5)	−8.78450
	standard deviation	0.678145 (2.4)	2.89762
TTS	mean	0.006323 (5.4)	−9.84639
	standard deviation	0.755265 (3.0)	3.09285
DA	mean	0.453518 (2.1)	−0.92269
	standard deviation	0.249252 (2.4)	0.51375
RENT	mean	0.062650 (3.5)	−4.84943
	standard deviation	0.496853 (2.4)	2.03926
δCURRENT	mean	1.06775 (8.6)	
Log-likelihood		−593.0	

which was taken as fixed.[12] The second column of the table shows the coefficients of the underlying normal distribution [that is, the μ_k and σ_k of equation (7)]. To compute the WTP values, two courses of action may be followed: first, take the ratio of the means of each attribute parameter to that of the RENT mean (which is analogous to the first method described above); and second, take the mean of the resulting WTP log-normal distribution parameters directly. Both set of results are presented in table 11.

The very considerable differences between the ratios of the mean and the means of the ratios introduce new evidence to the discussion. The ratios of the means do not yield the WTP for the mean individual household, but for a virtual one which perceives the mean marginal utility of the population for each attribute (that is, an 'individual household' which has the mean parameter for, say, the DA attribute and also the mean parameter for RENT). The existence of this household is not a fact but a mere coincidence, and even if such a household did exist, its WTP value would not be representative. So, again, this index may only be useful as a model specification search tool.

Table 11 shows that taking the ratio of the parameter means considerably underestimates the mean of the WTP distribution. Hensher and Greene (2003) simulated the resulting WTP log-normal distribution and also derived an unusually high mean. They managed to lower it to more plausible values by truncating the simulated distribution,

Table 11. Ratio of log-normal means and means of the log-normal willingness to pay (WTP).

Attribute		WTP for log-normal model	
		ratio of means	mean of WTP distribution
TTW	mean	37.5	2 401
(Ch$ per minute)	standard deviation		1 278 543
TTS	mean	23.3	1 490
(Ch$ per minute)	standard deviation		177 999
DA	mean	86 928	5 557 979
(Ch$ per days of alert per year)	standard deviation		50 424 646

[12] Usually the log-normal mean, median, and standard-deviation values are derived from the exponential of normal variables. In WinBUGS, however, log-normal distributions may be specified directly for the coefficients defined by their mean and standard deviation. Inference of mean and standard deviation of the exponentiated normal is done simply by inverting the process.

but found it very sensitive to this kind of constraint. So this phenomenon is not case specific and does not seem to depend on the data.

In fact, an analytical explanation for this underestimation can easily be derived. Consider two independently distributed log-normal structural parameters β and γ (for example, time and cost) with associated normal means b and c and variances s^2 and d^2, respectively. The ratio of their means can be expressed as a function of the coefficients of the underlying normal distributions:

$$\left.\begin{aligned} \bar{\beta} &= \exp\left(b + \frac{s^2}{2}\right) \\ \bar{\gamma} &= \exp\left(c + \frac{d^2}{2}\right) \end{aligned}\right\} \frac{\bar{\beta}}{\bar{\gamma}} = \exp\left(b - c + \frac{s^2 - d^2}{2}\right).$$

And from expression (7) we can express the mean of the WTP log-normal distribution in terms of the same coefficients:

$$\overline{\text{WTP}} = \exp\left(b - c + \frac{s^2 + d^2}{2}\right).$$

From here we can derive the relation

$$\overline{\text{WTP}} = \frac{\bar{\beta}}{\bar{\gamma}} \exp d^2.$$

Thus, the ratio of the means of log-normal parameters is equal to the mean WTP value deflated by the exponential of the variance of the normal distribution underlying the log-normal cost coefficient (that is, the parameter in the denominator of the WTP ratio). In other words, the WTP mean and the ratio of parameter means are scaled by a proportionality factor which, by the way, is fixed for the model (that is, the three attributes considered in this example are scaled by the same factor). The logic of this effect is as follows: the larger the variance of the cost coefficient, the larger the portion of the mass of the denominators that will be near to zero, and hence the mean WTP will grow larger.

The use of log-normal distributions for valuation purposes is not recommended. Their wide tail tends to give extremely large WTP values, with high probabilities yielding large portions of cumulative mass close to zero which distort the analysis. Its main appeal is that it allows constraining the parameters to be strictly positive (negative coefficients, enter with a negative sign in the utility formulation). However, as we have seen, the relative ease of the estimation with normal distributions may also lead to structural parameters with correct theoretical signs. Thus, it is not worthwhile undergoing the effort of estimating the model with log-normal distributed parameters, as even if the individual values show a large portion of incorrectly signed people, the right course of action should be to investigate them for consistency, and perhaps remove them from the sample.

6.4 Fixing the cost coefficient

A fourth method consists of fixing the cost coefficient and thus letting the WTP distribution follow the distribution of the numerator; if it follows a normal distribution, as in our example, the resulting WTP distribution is simply given by:

$$\left.\begin{aligned} \theta_{\text{att}} &\sim \text{N}(\mu_{\text{att}}, \sigma_{\text{att}}) \\ \theta_{\text{c}} &\text{ fixed} \end{aligned}\right\} \frac{\theta_{\text{att}}}{\theta_{\text{c}}} \sim \text{N}\left(\frac{\mu_{\text{att}}}{\theta_{\text{c}}}, \frac{\sigma_{\text{att}}}{\theta_{\text{c}}}\right),$$

where c indicates cost and att indicates an attribute other than cost. Revelt and Train (2000) cite three reasons for fixing the cost coefficient: (1) this effectively solves the problem under discussion; (2) the ML model tends to be unstable when all coefficients vary over the population, and identification issues arise (Ruud, 1996); and (3) the choice of an appropriate distribution for the cost coefficient is not straightforward, as the normal and other distributions allow for positive values, and the log-normal is both hard to estimate and gives values very close to zero—as discussed above.

Our models incorporate a fixed coefficient for the δCURRENT attribute, avoiding potential identification problems. The other two arguments are valid, but can be resolved by the use of individual-level WTP estimation, as proposed in section 3. Notwithstanding, there is one drawback of this method that needs attention.

Table 12 compares estimates of WTP derived from the MNL model with those of an ML model (ML4) with a fixed RENT coefficient. As can be seen, the means of the resulting WTP distributions are considerably higher than the MNL point estimates—a result that has also been reported by Algers et al (1999) and Revelt and Train (1998).

Table 12. Mean estimates of willingness to pay for fixed-cost coefficient mixed logit (ML) and multinomial (MNL) models.

Attribute		Willingness to pay	
		MNL	ML4
TTW	mean	36	51
(Ch$ per minute)	standard deviation		54.8
TTS	mean	22	31
(Ch$ per minute)	standard deviation		47.5
DA	mean	124 362	126 160
(Ch$ per days of alert per year)	standard deviation		107 430

Hensher (2001a; 2001b; 2001c) has also found higher mean WTP values for heteroscedastic and autoregressive specifications, which could indicate that mixed-logit models (with any error structure) tend to overestimate WTP values. But Hensher did not explore the possibility that constraining only part of the error structure could be causing an unbalanced growth in the model coefficients, hence producing higher welfare estimates.

In section 5 we explained why larger means for ML parameters, in relation to the MNL model, should be expected because of the extra variance explained by the random parameters; we also discussed possible reasons for obtaining uneven enlargement factors. In fact, constraining a taste coefficient to be fixed over the population may make it grow in a less-than-average proportion (that is, the parameters that are allowed to vary grow more than the parameters that *should* vary over the population, but are constrained to be fixed). Note that this is not the case with the δCURRENT parameter, because its standard deviation was originally estimated and was found not to be significant. This issue is best illustrated in table 13, where the different columns present the same model estimated with different parameters being fixed. In all cases, the coefficients with potential variability remain 'small' when fixed.

In this model fixing the RENT coefficient makes the denominator of the WTP smaller than it should be, causing an overestimation of the mean WTP (as well as of the whole WTP distribution). The inverse miscalculation can occur if a noncost coefficient is fixed: then the numerator remains smaller, and so does the WTP value. In table 13, the cells containing WTP values affected by constraining a given coefficient are shown in bold.

Table 13. Willingness to pay (WTP) (with *t*-statistics shown in parentheses) of mixed logit (ML) models estimated with different parameters being fixed.

Attribute		Parameters[a]				
		ML1	ML4 RENT fixed	ML5 TTW fixed	ML6 TTS fixed	ML7 DA fixed
TTW	mean	−0.01141 (−6.7)	−0.00966 (−6.2)	−0.00688 (−11.9)	−0.01036 (−6.7)	−0.01004 (−6.1)
	standard deviation	0.01133 (8.2)	0.01036 (8.3)	-	0.01021 (8.4)	0.01077 (8.2)
TTS	mean	−0.00783 (−4.6)	−0.00588 (−3.9)	−0.00672 (−4.2)	−0.00503 (−10.3)	−0.00708 (−4.3)
	standard deviation	0.01025 (7.5)	0.00898 (8.3)	0.00921 (8.0)	-	0.00961 (7.9)
DA	mean	−0.56960 (−8.3)	−0.45870 (−8.0)	−0.50060 (−8.0)	−0.51480 (−8.5)	−0.44540 (−13.1)
	standard deviation	0.46920 (7.0)	0.39060 (6.9)	0.42240 (6.8)	0.39380 (6.6)	-
RENT	mean	−0.06974 (−8.9)	−0.04363 (−13.9)	−0.06017 (−8.5)	−0.06010 (−8.7)	−0.06060 (−8.4)
	standard deviation	0.05339 (5.6)	-	0.04582 (7.2)	0.04479 (7.5)	0.04874 (7.5)
δCURRENT	mean	1.16800 (5.8)	1.01200 (5.6)	1.10000 (5.7)	1.07400 (5.9)	1.08500 (6.0)
Willingness to pay						
TTW	mean	36	**51**	**26.4**	39.8	38
(Ch$ per minute)	standard deviation		54.8			
TTS	mean	22	**31**	25.7	**19.3**	26.9
(Ch$ per minute)	standard deviation		47.5			
DA (Ch$ per days	mean	124 362	**126 160**	99 837	102 788	**88 198**
ofalert per year)	standard deviation		107 430			
Log-likelihood		−570.0	−698.0	−634.9	−609.1	−646.0

[a] The WTP for models ML5 to ML7 do not have a standard-deviation estimate as they are constructed as the ratio between a fixed parameter and another with a normal distribution.

7 Estimation of willingness to pay from individual-level parameters

In section 5 we discussed two econometric processes involved in the estimation of individual-level structural parameters: the use of a Bayesian approach and the conditioning of individual choices to the population parameters. As mentioned, we applied the Bayesian approach in this research. The estimation of individual taste parameters eliminates the issue of analysing the WTP distribution resulting from the division of two random variables over the population. Instead, individual-level WTP point estimates can be computed along with their individual confidence intervals.

Figures 3 and 4 present frequency charts for the valuation of the three attributes in our stated-preference experiment (TTW, TTS, and DA), derived from individual WTP point estimates obtained from model ML1. The charts show high concentrations on each edge of the axis, accounting for extremely large positive and negative WTP values. However, it is important to mention that, notwithstanding the sign of the WTP value, all implausibly large values belong to households with nonsignificant RENT parameters. That is, the denominator of the WTP ratio is statistically close to zero, yielding an inordinately large value.

It is also important to mention that in figures 3(a) and 3(b), the only negative WTP values are also associated with extreme cases. In fact, they correspond to the few observations with an incorrect sign for the RENT parameter; as this was also not significant in those cases, it caused the ratio to grow disproportionably.

As suggested above, special attention should be given to observations with a cost parameter statistically equal to zero. In these cases, the WTP ratio grows to implausibly large monetary valuations for reductions in the corresponding attribute. On the other hand, as the individual household does not place any weight on the cost

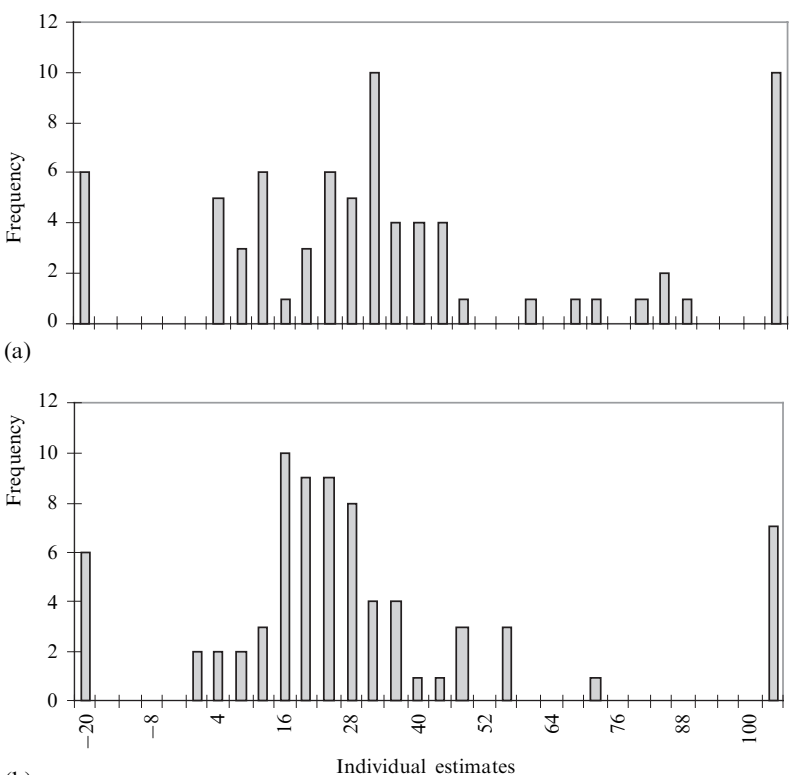

Figure 3. Individual-level point estimates of willingness to pay for reductions lin (a) TTW, (b) TTS.

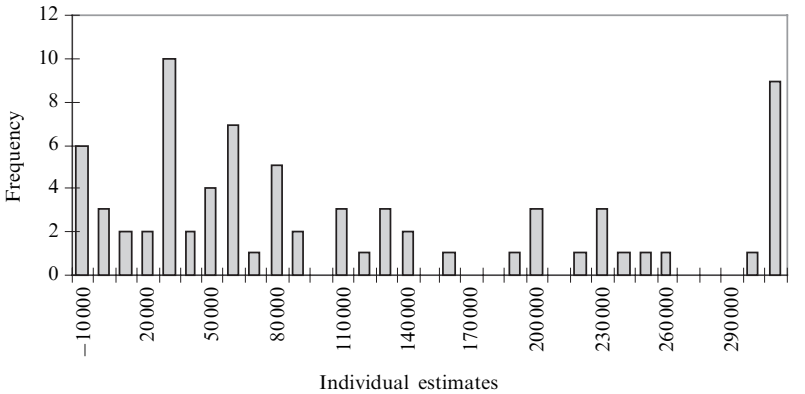

Figure 4. Individual-level of willingness to pay point estimates for reductions in DA.

attribute, we can debate whether those observations do not consider the cost attribute at all, or whether the weight they place on it is negligible in relation to the rest of the attributes. If the last is the case, the interpretation of an extremely large WTP value would be correct. If not, monetary valuations can not be computed for these observations. Further theoretical development is necessary to define criteria to help answer this question, but note that it is case-specific (that is, it depends on the survey design, the underlying microeconomic model, and the characteristics of the valued attributes).

Having cleared the above, we can now derive the real mean value of the WTP distribution over the population. The means and standard deviations of the valued attributes, estimated from the discrete set of individual WTP point estimates, are presented in table 14. Comparison of these values with those presented in tables 2, 8, 9, 11, and 12, illustrates the potential miscalculation of WTP values which may arise from attempting to treat the variability of a finite population as a continuous distribution. The large variances of the WTP values are caused by the extremely large values resulting from the division by close-to-zero RENT coefficients in some cases, as already mentioned.

The estimation of individual-level WTP values is as close as we can get to the correct method of valuation inference from mixed logit models. However, for project evaluation and cost–benefit analysis we usually need data for different groups or strata in the population. The beauty of individual-level data is that an analysis at the level of a given stratification can be performed simply by averaging the WTP values of those individuals present in each strata, along with their cluster variance. In fact, thresholds (or strata boundaries) can even be defined ex-post in order to minimise

Table 14. Mean and standard deviation of individual-level willingness to pay (WTP) values in multinomial (MNL) and mixed logit (ML) models.

Attribute		Individual WTP values	
		MNL	ML1
TTW	mean	36	41.6
(Ch$ per minute)	standard deviation		42 379
TTS	mean	22	18.7
(Ch$ per minute)	standard deviation		11 513
DA	mean	124 362	139 920
(Ch$ per days of alert per year)	standard deviation		1.7×10^{11}

the variance of the WTP values across the group, and hence allow more homogeneous segments to be defined for project evaluation and detailed analysis.

8 Conclusions

We have shown the complexity associated with the use of random-parameter models for estimating willingness to pay. We have also shown that a useful procedure is to estimate individual-level parameters, rather than population-distribution parameters as is normally done. Among other things, this may allow us to find out more accurately if, for example, some individuals have not responded seriously to a stated-preference survey. Also, and perhaps more speculatively, with results at the individual level it may be possible to search for 'representative' individuals of a particular class when sampling in order to collect smaller samples which are richer in individuals with the appropriate features to represent previously defined strata of interest.

The power of the approach suggests that more emphasis should be put on the collection of data of high quality, rather than excessive preoccupation with sample size. Future research can explore the potential of the use of more informative priors to reduce the necessity for larger sample sizes (even though they will always be preferable). As evidence for this, in this paper we used a very small sample size (75 individual households) and still managed to obtain useful results.[13] However, the relation between the classical and Bayesian estimates was not smooth, as may be the case for larger samples [that is, more than 300 individuals (see, for example Huber and Train, 2001)], where the available evidence would suggest that Bayesian estimation does not have clear advantages over classical procedures. Notwithstanding, our results suggest that the possibility of estimating robust models with smaller samples could be an important advantage of this technique.[14]

Finally, it is important to bear in mind that these results are data specific and follow the main purpose of the study—which is to provide evidence of the potential scenarios which a researcher could face when estimating these kind of models and, in particular, willingness-to-pay values. Discussions on the stability and robustness of classically estimated parameters are available in the existing literature (Bhat, 2001; Garrido and Silva, 2004; Walker, 2001). Further research is being carried out to look into robustness issues of Bayesian outputs, but these did not lie within our scope in this paper.

Acknowledgements. We wish to thank Kenneth Train and Joan Walker for many ideas and comments, and Pilar Iglesias for her help in Bayesian issues and for introducing us to WinBUGS. David Hensher, Sergio Jara-Díaz, Luis I Rizzi, and Huw Williams also deserve our thanks for always being available whenever we consulted them; two anonymous referees provided very useful comments to improve the paper and we thank them also. Finally, we wish to acknowledge the support of the Chilean Fund for Scientific and Technological Development (FONDECYT) for having provided the funds to complete this research through Project 1020981.

References

Algers S, Bergstrom M, Dahlberg M, Dillen J, 1999, "Mixed logit estimation of the value of travel time", working paper, Department of Economics, Uppsala University, Uppsala

Allenby G, Rossi P, 1999, "Marketing models for consumer heterogeneity" *Journal of Econometrics* **89** 57 – 78

Andrews R L, Ansari A, Currim I S, 2002, "Hierarchical Bayes versus finite mixture conjoint analysis models: a comparison of fit, prediction and partworth recovery" *Journal of Marketing Research* **39** 87 – 98

Armstrong P M, Garrido R A, Ortúzar J de D, 2001, "Confidence intervals to bound the value of time" *Transportation Research* **37E** 143 – 161

[13] Huber and Train (2001) label 340 individuals a 'small' sample.

[14] This is being studied at the moment with the aid of simulated data in a separate piece of research (Godoy, 2004).

Arora N, Allenby G, Ginter J, 1998, "A hierarchical Bayes model of primary and secondary demand" *Marketing Science* **17** 29 – 44
Bhat C, 2001, "Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model" *Transportation Research* **35B** 677 – 695
Chib S, Greenberg E, 1995, "Understanding the Metropolis – Hastings algorithm" *The American Statistician* **49** 327 – 335
Cowles M K, Carlin B P, 1996, "Markov chain Monte Carlo convergence diagnostics: a comparative review" *Journal of the American Statistical Association* **91** 883 – 904
Espino R, Ortúzar J de D, Román C, 2004, "Confidence intervals for willingness to pay measures in mode choice models", in *Proceedings XIII Panamerican Congress of Traffic and Transportation Engineering* Albany, USA, http://www.eng.rpi.edu/panam/
Ettema D, Gunn H, De Jong G, Lindveld K, 1997, "A simulation method for determining the confidence interval of a weighted group average value of time", in *Transportation Planning Methods I, Volume P414. Proceedings of the 25th European Transport Forum* (PTRC Education and Research Services, London) pp 101 – 112
Fieller E, 1932, "The distribution of the index in a normal bivariate population" *Biometrika* **24** 428 – 440
Galilea P, Ortúzar J de D, 2004, "Valuing noise level reductions in a residential location context" *Transportation Research* **9D** forthcoming
Garrido R A, Silva M, 2004, "Low discrepancy sequences for the estimation of mixed logit models" *Transportation Research B* forthcoming
Gaudry M J I, Jara-Díaz S R, Ortúzar J de D, 1989, "Value of time sensitivity to model specification" *Transportation Research* **23B** 151 – 158
Godoy G, 2004, "Estimación Bayesiana de modelos flexibles de elección discreta" [Bayesian estimation of flexible discrete choice models], MSc thesis, Department of Transport Engineering, Pontificia Universidad Católica de Chile, Santiago
Hajivassiliou V, Ruud P, 1994, "Classical estimation methods for LDV models using simulation", in *Handbook of Econometrics, Volume IV* Eds R Engle, D McFadden (Elsevier, New York) pp 2383 – 2441
Hensher D A, 2001a, "The sensitivity of the valuation of travel time savings to the specification of unobserved effects" *Transportation Research* **37E** 129 – 142
Hensher D A, 2001b, "The valuation of commuter travel time savings for car drivers: evaluating alternative model specifications" *Transportation* **28** 101 – 118
Hensher D A, 2001c, "Measurement of the valuation of travel time savings" *Journal of Transport Economics and Policy* **35** 71 – 98
Hensher D A, Greene W H, 2003, "The mixed logit model: the state of practice" *Transportation* **30** 133 – 176
Hinkley D, 1969, "On the ratio of two correlated normal random variables" *Biometrika* **56** 635 – 639
Horowitz J L, 1981, "Sampling error, specification and data errors in probabilistic discrete choice models", appendix C of *Applied Discrete Choice Modelling* D A Hensher, L W Johnson (Croom Helm, London) pp 417 – 435
Huber J, Train K E, 2001, "On the similarity of classical and Bayesian estimates of individual mean partworths" *Marketing Letters* **12** 257 – 267
Iragüen P, Ortúzar J de D, 2004, "Willingness-to-pay for reducing fatal accident risk in urban areas: an Internet-based web page stated preference survey" *Accident Analysis and Prevention* **36** 513 – 524
Jara-Díaz S R, 1990, "Income and taste in mode choice models: are they surrogates?" *Transportation Research* **25B** 341 – 350
Kass R, Carlin B, Gelman A, Neal R, 1998, "Markov chain Monte Carlo in practice: a roundtable discussion" *The American Statistician* **52** 93 – 100
Lahiri K, Gao J, 2001, "Bayesian analysis of nested logit model by Markov chain Monte Carlo", working paper, Department of Economics, State University of New York at Albany, NY
Lenk P, De Sarbo W, Green P, Young M, 1999, "Hierarchical Bayes conjoint analysis: recovery of partworth heterogeneity from reduced experimental designs" *Marketing Science* **15** 173 – 191
Louviere J J, Hensher D A, Swait J D, 2000 *Stated Choice Methods: Analysis and Applications* (Cambridge University Press, Cambridge)
McCulloch R, Rossi P, 1994, "An exact likelihood analysis of the multinomial probit model" *Journal of Econometrics* **64** 207 – 240
McFadden D, Train K E, 2000, "Mixed MNL models for discrete response" *Journal of Applied Econometrics* **15** 447 – 470

Meijer E, Rouwendal J, 2000, "Measuring welfare effects in models with random coefficients", research report 00F25, SOM Research School, University of Groningen, Groningen
Morey E, Rossman K G, 2002, "Using stated-preference questions to investigate variation in willingness to pay for preserving marble monuments: classical heterogeneity and random parameters", WP, Economics Department, University of Colorado at Boulder, CO
Ortúzar J de D, Rodríguez G, 2002, "Valuing reductions in environmental pollution in a residential location context" *Transportation Research* **7D** 407 – 427
Ortúzar J de D, Willumsen L G, 2001 *Modelling Transport* 3rd edition (John Wiley, Chichester, Sussex)
Ortúzar J de D, Martínez F J, Varela F J, 2000, "Stated preferences in modelling accessibility" *International Planning Studies* **5** 65 – 85
Ortúzar J de D, Rodríguez G, Sillano M, 2002, "Willingness-to-pay for reducing atmospheric pollution" *Proceedings of the European Transport Conference*, Homerton College, Cambridge, CD, PTRC Education and Research Services Ltd, London, http://www.ptrc-training.co.uk
Pérez P E, Martínez F J, Ortúzar J de D, 2003, "Microeconomic formulation and estimation of a residential location model: implications for the value of time" *Journal of Regional Science* **43** 771 – 789
Raftery A, Lewis S, 1992, "How many iterations in the Gibbs sampler?", in Bayesian Statistics 4 Eds J M Bernardo, A F M Smith, A P David, J O Berger (Oxford University Press, New York) pp 763 – 773
Revelt D, Train K E, 1998, "Mixed logit with repeated choices: households' choice of appliance efficiency level" *Review of Economics and Statistics* **80** 647 – 657
Revelt D, Train K E, 2000, "Customer-specific taste parameters and mixed logit", WP E00-274, Department of Economics, University of California at Berkeley, CA
Rizzi L I, Ortúzar J de D, 2003, "Stated preference in the valuation of interurban road safety" *Accident Analysis and Prevention* **35** 9 – 22
Rizzi L I, Ortúzar J de D, 2004, "Road safety valuation under a stated choice framework" *Journal of Transport Economics and Policy* in press
Ruud P, 1996, "Simulation of the multinomial probit model: an analysis of covariance matrix estimation", working paper, Department of Economics, University of California at Berkeley, CA
Sawtooth Software, 1999 *The CBC/HB Module for Hierarchical Bayes Estimation* http://www.sawtoothsoftware.com/cbc.shtml
Small K E, Rosen H, 1981, "Applied welfare economics with discrete choice models" *Econometrica* **49** 105 – 130
Spiegelhalter D J, Thomas A, Best N G, 2001 *WinBUGS Beta Version 1.4 User Manual* MRC Biostatistics Unit, Institute of Public Health, University of Cambridge, Cambridge
Train K E, 1998, "Recreational demand models with taste differences over people" *Land Economics* **74** 230 – 239
Train K E, 2001, "A comparison of hierarchical Bayes and maximum simulated likelihood for mixed logit", working paper, Department of Economics, University of California at Berkeley, CA
Train K E, 2003 *Discrete Choice Methods with Simulation* (Cambridge University Press, Cambridge)
Train K E, Sonnier G, 2003, "Mixed logit with bounded distributions of partworths", working paper, Department of Economics, University of California at Berkeley, CA
Walker J, 2001 *Extended Discrete Choice Models: Integrated Framework, Flexible Error Structures and Latent Variables* PhD dissertation, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA