Testing Mixed Logit and Probit Models by Simulation

Marcela A. Munizaga and Ricardo Alvarez-Daziano

Discrete choice models with error structures that are not independent and identically distributed have received enormous attention in the recent literature. A detailed synthetic study tests this type of model in a controlled case. With mixed logit and probit models as the study objects, calibration was implemented with the use of software available on the Internet. The controlled situation was built as a simulation laboratory, which generated databases with known parameters. The effects of various elements were analyzed: number of repetitions of the simulation, number of observations in the database, and how the use of Halton sequences improves the mixed logit calibration. The scale effects on the different models are also discussed. The results obtained in this specific context lead to some recommendations for future users of these powerful modeling tools. In particular, flexible structures require large sample sizes to calibrate the elements of the covariance matrix.

During the past decade, remarkable advances have been made in the calibration of discrete choice models that are not independent and identically distributed (i.e., non-IID). Various model possibilities have become available, such as the multinomial probit (*1*), heteroscedastic extreme value logit (*2*), and mixed logit (ML) (*3*), the most popular. Even though none of these models is really new conceptually, all their calibration is now feasible computationally, such that it is possible to take full advantage of their flexibility. Practitioners and researchers are willing to move from the multinomial logit (MNL) and nested logit (NL) models, which were the standard until only a few years ago, toward these more general models—specifically toward ML.

This paper focuses on the ML model, which is becoming popular, and the probit model, which has been available longer but does not appear as popular. In both models, the likelihood function cannot be evaluated directly, and simulation or other approximation methods must be used. The modeler must weigh various aspects to determine the model to be used in a specific problem, the number of observations to collect, and the number of repetitions of the simulation procedure required. To assess how these key elements influence the estimation procedure, a simulation experiment was conducted according to the methodology proposed by Williams and Ortúzar (*4*) and applied by Munizaga et al. (*5*). Both the number of repetitions of the simulation and the number of observations were varied, and then the model behavior was analyzed. In addition, analysis was performed on the behavior of the simpler models (MNL and NL) in the case of a more complex reality in which non-IID errors are present. Model behavior was analyzed in three dimensions: capability to recover the indirect utility function parameters, behavior of the likelihood function, and prediction capabilities.

The next section describes the differences between and similarities of ML and probit in theoretical terms, with emphasis on those aspects that have practical consequences and are worth exploring with the simulation experiments. Then, a detailed description of the simulation procedure follows, and the simulation results are presented and analyzed. Finally, a synthesis of the conclusions of the whole process is presented.

DIFFERENCES AND SIMILARITIES BETWEEN MIXED LOGIT AND PROBIT

The ML model (also known as error components or kernel logit) is built on the basis of an MNL model by including additional error terms to impose the desired non-IID effects. So the utility function of alternative i for individual n (U_{in}) is defined as

$$U_{in} = V_{in} + \eta_{in} + \epsilon_{in} \quad (1)$$

where

V_{in} = deterministic component of utility,
η_{in} = any density function, and
ϵ_{in} = IID Gumbel error term.

Conditional in η, the choice probabilities are exactly those of the MNL model. But the choice probability of this model, represented by the MNL kernel integral over η, does not have a closed mathematical expression as the MNL or NL models do. Because the choice probability integral cannot be solved analytically, simulation is used to evaluate it. The estimation procedure is well described by Train (*6*) and, in a complementary paper, by Hensher and Greene (*7*).

On the one hand, it can be said that the ML model is built on the assumption of additional error terms that may imply a heteroscedastic and correlated covariance matrix. On the other hand, a multinomial probit is derived on the assumption that, given a utility function $U_{in} = V_{in} + \epsilon_{in}$, the vector $\epsilon_n = (\epsilon_{1n}, \ldots, \epsilon_{in}, \ldots, \epsilon_{Jn})'$ distributes multivariate normal with Σ covariance matrix; only one error term is assumed, but it can have a general covariance matrix. The probit model does not have a closed-form expression of the choice probability either, so it becomes necessary to use approximation or simulation. Presently, the most used estimation method is probably the simulated maximum likelihood with the Geweke–Hajivassiliou–Keane (GHK) simulator (*8*), which recursively reduces the dimension of the integral up to an equivalent problem in which repetitions of a truncated unidimensional normal are required. The simulated probabilities are unbiased, continuous, and differentiable.

The simulations for the probit and ML models have different dimensions analytically (and therefore computationally): the number of alternatives minus one for the probit model (because it is based on

Department of Civil Engineering, University of Chile, Casilla 228-3, Santiago, Chile.

Transportation Research Record: Journal of the Transportation Research Board, No. 1921, Transportation Research Board of the National Academies, Washington, D.C., 2005, pp. 53–62.

utility differences) and the number of random terms plus the basic Gumbel term for the ML model.

Both models are subject to identifiability restrictions that have been studied and are now well known (*9*). Apart from the traditional identifiability restrictions that apply to all discrete choice models because the decisions are determined by utility differences, some special conditions must be imposed in the deviated covariance matrix to ensure identifiability of the additional parameters.

Also related to this subject is the scale effect. All discrete choice models must be scaled to become identifiable. For MNL models, the calibrated parameters are scaled by a factor equal to $\pi/\sqrt{6}\sigma$ that cannot be identified. This effect also appears in the ML and probit case. For the probit model, the exact factor depends on how the covariance matrix is normalized. However, for the ML model, it is unclear. The problem for the specific model presented is discussed in the following section.

DESCRIPTION OF SIMULATION EXPERIMENTS

The simulation experiments were implemented as a realistic case on the basis of transport mode choice in which only the values of the explanatory variables and the chosen option were available in the estimation process, and they were consistent with the random utility maximization theory. The error distribution assumed was consistent with ML in the case of correlation between alternatives.

The data sets were generated by computing the simulated choice for each observation as the alternative that has the largest utility (U_{in}). Those utilities were calculated as the sum of the observable component V_{in} (sum of the taste parameters times the corresponding attributes) and the error terms sampled according to the selected distributions. The attributes, generated by pseudo random sampling, were travel cost, travel time, and access time for each of four modes (car, bus, metro, and taxi) and a binary dummy variable for high income added to the car utility. The time and cost attributes were normally distributed, with mean and variance taken from a real database. The parameters of the utility function, also taken from models fitted to real data, were –0.005 for cost, –0.08 for travel time, and –0.16 for access time. The magnitude of the variances of the error terms was chosen to achieve a reasonable balance between the number of individuals who would change the chosen option due to the error term and those who would not.

The focus was on a case in which bus and metro (underground) are considered similar alternatives, which is the classic reason to expect correlation among modes from unobservable effects. To build the stochastic part of the utility function, the nested ML specification proposed by Brownstone and Train (*3*) was used; it includes an error ϵ_i IID Gumbel (0, λ) and an error μ_n distributed normal (0, σ_μ^2) that captures the potential correlation in nest *n*. This specification leads to a correlated and heteroscedastic covariance matrix:

$$U_{\text{car},n} = V_{\text{car},n} + \epsilon_{\text{car},n}$$

$$U_{\text{bus},n} = V_{\text{bus},n} + \mu_n + \epsilon_{\text{bus},n}$$

$$U_{\text{metro},n} = V_{\text{metro},n} + \mu_n + \epsilon_{\text{metro},n}$$

$$U_{\text{taxi},n} = V_{\text{taxi},n} + \epsilon_{\text{taxi},n}$$

$$\Sigma = \begin{bmatrix} \sigma_\epsilon^2 & 0 & 0 & 0 \\ 0 & \sigma_\mu^2 + \sigma_\epsilon^2 & \sigma_\mu^2 & 0 \\ 0 & \sigma_\mu^2 & \sigma_\mu^2 + \sigma_\epsilon^2 & 0 \\ 0 & 0 & 0 & \sigma_\epsilon^2 \end{bmatrix} \quad (2)$$

This matrix is heteroscedastic because the variance of μ_n is added to that of the IID Gumbel term of the MNL kernel. Also implemented was the homoscedastic case by inclusion of an additional IID normal (0, σ_μ^2) term for the nonnested alternatives to make the covariance structure equivalent to that of the NL model. (This additional term is difficult to justify because it does not have a direct theoretical interpretation.)

$$U_{\text{car},n} = V_{\text{car},n} + \mu_{\text{car}} + \epsilon_{\text{car},n}$$

$$U_{\text{bus},n} = V_{\text{bus},n} + \mu_n + \epsilon_{\text{bus},n}$$

$$U_{\text{metro},n} = V_{\text{metro},n} + \mu_n + \epsilon_{\text{metro},n}$$

$$U_{\text{taxi},n} = V_{\text{taxi},n} + \mu_{\text{taxi}} + \epsilon_{\text{taxi},n}$$

$$\Sigma = \begin{bmatrix} \sigma_\mu^2 + \sigma_\epsilon^2 & 0 & 0 & 0 \\ 0 & \sigma_\mu^2 + \sigma_\epsilon^2 & \sigma_\mu^2 & 0 \\ 0 & \sigma_\mu^2 & \sigma_\mu^2 + \sigma_\epsilon^2 & 0 \\ 0 & 0 & 0 & \sigma_\mu^2 + \sigma_\epsilon^2 \end{bmatrix} \quad (3)$$

Model performance was tested in terms of ability to recover the known taste parameters and correlation and in terms of prediction capabilities. Recovery of the taste parameters was evaluated by using the classical statistical indicators of *t*-test and confidence interval and by considering the scale effects. In some cases, the scale effect is clear and can easily be incorporated to allow direct comparisons between the calibrated parameters and the values used to generate the data. However, in other cases, the scale effect cannot be isolated, which makes comparisons more difficult. Those cases are highlighted.

A response analysis was carried out by implementation of some changes in the variables for level of service that represent policy changes and evaluation of the model predictions in those modified scenarios. The reference used to make comparisons is the simulated behavior—obtained as the predictions of the simulator—in the same modified scenario. For this case, the simulator used the known taste parameters and the modified variables for level of service. The predictions of two models calibrated with the same database also can be compared. The adequate tool for these comparisons is the χ^2 test (*10*), calculated as $\chi^2 = \sum_i \left(\hat{N}_i - N_i\right)^2 / N_i$, where $\hat{N}_i$ is the number of individuals that choose alternative *i* according to the prediction made by the model and N_i is the number of individuals choosing alternative *i* according to the simulator.

MODEL ESTIMATION AND ANALYSIS OF RESULTS

With the data sets generated as described and with observations of the chosen option and the attribute values for the complete choice set, the choice models were estimated through maximum likelihood using GAUSS software (*11*). An algorithm was implemented to estimate probit models in GAUSS using the simulated maximum likelihood (*12*) approach with the GHK simulator (*1, 8*) for the choice probabilities. The ML code (*13*) developed by Train was downloaded from his web page (*14*). The two available procedures for generating random numbers were used: pseudo random numbers and quasi-random numbers (Halton sequences). The use of pseudo random numbers is the usual procedure to obtain random draws. Quasi-random numbers are a deterministic series that covers the integration domain

in a more efficient way. One of these series is a Halton sequence. It is proposed as the way to make draws when estimating an ML model; Williams and Ortúzar provide an informative discussion on the use of Halton sequences (*4*). The MNL and NL models calibrated were also implemented in GAUSS with the use of a custom code that is very easy to program.

The specification of each model was as similar as possible to the specification used to generate the data for each case. The deterministic component of the utility function always was the same, including mode constants, travel time, access time, and cost parameters as well as an income dummy variable for the car alternative. The error structures used for generating the data are presented in Equations 2 and 3. It is important to make clear how this error structure was specified for calibration purposes, because it affects the scale of the model parameters. For the MNL model, the covariance matrix is as shown in Equation 4, in which neither correlation nor heteroscedasticity is allowed. The scale parameter λ_{MNL} is not identifiable, and the taste parameters calibrated will have that factor included. Scaling and identifiability issues can be discussed only because the data are synthetic; it would not be feasible when using real data. The normalization strategy that warranties identifiability is trivial for the MNL model but more difficult in complex error structures.

$$\Sigma_{MNL} = \frac{\pi^2}{6\lambda_{MNL}^2}\begin{bmatrix}1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\end{bmatrix} \quad (4)$$

Because $\lambda = \pi/\sqrt{6}\sigma$, if the variance is $\sigma^2 = \pi^2/6$, then λ will be equal to unity.

In the case of NL, the covariance matrix is as described in Equation 5. This model is homoscedastic but allows capturing correlation by the structural parameter φ that can be estimated. As in the MNL case, the scale parameter λ_{NL} is nonidentifiable and will be incorporated into the taste parameters calibrated.

$$\Sigma_{NL} = \frac{\pi^2}{6\lambda_{NL}^2}\begin{bmatrix}1 & 0 & 0 & 0\\ 0 & 1 & (1-\varphi^2) & 0\\ 0 & (1-\varphi^2) & 1 & 0\\ 0 & 0 & 0 & 1\end{bmatrix} \quad (5)$$

In the case of ML, both cases used to generate the data can be implemented: homoscedastic and heteroscedastic. In the heteroscedastic case, the covariance matrix of Equation 2 also can be written as in Equation 6, in which the variance of the Gumbel error is written in terms of the scale parameter λ_{ML}. The variance of the common error component (σ_μ^2) is introduced to capture correlation; it can be estimated within the model calibration process but will be subject to the scale effect $\hat{\sigma}_\mu = \lambda_{ML}\sigma_\mu$. That variance is associated with only one additional error term (μ_n), so the dimension of the simulation required to calibrate the model is equal to one. The homoscedastic case described by Equation 3 has three additional error terms (μ_{car}, μ_n, and μ_{taxi}), so the dimension of the simulation required is three. But all three error terms have the same variance (to achieve homoscedasticity), so in this case also, only σ_μ^2 is calibrated.

$$\Sigma = \begin{bmatrix}\frac{\pi^2}{6\lambda_{ML}^2} & 0 & 0 & 0\\ 0 & \sigma_\mu^2 + \frac{\pi^2}{6\lambda_{ML}^2} & \sigma_\mu^2 & 0\\ 0 & \sigma_\mu^2 & \sigma_\mu^2 + \frac{\pi^2}{6\lambda_{ML}^2} & 0\\ 0 & 0 & 0 & \frac{\pi^2}{6\lambda_{ML}^2}\end{bmatrix} \quad (6)$$

In the probit case, the heteroscedastic and homoscedastic structures can also be accommodated. The heteroscedastic matrix is shown in Equation 7. It is written in terms of the independent alternatives error variance σ_b^2 and the covariance σ_μ^2. To be able to calibrate the model, the basic scale must be fixed. This was accomplished by setting σ_b^2 to its known value. In doing this, the scale is forced to be equal to unity, which simplifies comparisons. In the homoscedastic case, the only difference is that the entire diagonal will be equal to the terms at the center of the matrix.

$$\Sigma = \frac{1}{\sigma_b^2}\begin{bmatrix}1 & 0 & 0 & 0\\ 0 & \frac{\sigma_b^2 + \sigma_\mu^2}{\sigma_b^2} & \frac{\sigma_\mu^2}{\sigma_b^2} & 0\\ 0 & \frac{\sigma_\mu^2}{\sigma_b^2} & \frac{\sigma_b^2 + \sigma_\mu^2}{\sigma_b^2} & 0\\ 0 & 0 & 0 & 1\end{bmatrix} \quad (7)$$

Number of Repetitions

Because ML and probit models require simulation to be estimated, the authors wanted to examine the behavior of the estimates when varying the number of repetitions of the simulation procedure and also when using Halton sequences in the ML model. Here, a database of 4,000 hypothetical individuals was used and a choice made among four alternatives. The error structure is that shown in Equation 3 (homoscedastic), with a correlation of 0.5 for Alternatives 2 and 3. The number of simulation repetitions for the estimation procedure was varied from five to 1,000. In this case, the dimension of the simulation for the probit model is three (four modes), equal to the dimension of the ML (three additional error terms).

The more relevant parameters of the calibration results are reported in Table 1 for the probit model and for the ML model calibrated with pseudo random numbers (MLR) and with quasi-random numbers (MLH). A "target value" reference is included for comparison—that is, the original parameter (the one used to generate the data), scaled appropriately. Because the covariance structure is known, the scale parameter can be calculated in this case, as demonstrated in the previous section. The subjective value of time (SVT) is calculated as the ratio between time and cost parameters; the scale cancels out. The *t*-value for the SVT is calculated with a formula derived from a Taylor expansion (*15*).

The results for the probit model estimation indicate that the parameters stay stable, even for a low number of repetitions. SVT is systematically overestimated and the correlation systematically underestimated. The taste parameters—and consequently the SVT—are quite stable, whereas correlation is detected with increasing accuracy as the number of repetitions of the simulation increases.

At first glance, the point estimates of the ML parameters seem to vary more than the probit parameters with the number of repetitions of the simulation. However, in this case, the scale effect is present in a curious manner. The scale factor depends on only the Gumbel part of the error variance. This database was built in such a way that the total variance is equal to $\pi^2/6$ such that the scale parameter λ should be equal to unity if an MNL model were calibrated; however, in the ML case, λ will depend on how much of that variance can be associated to the Gumbel term. Because in this case $\rho = 0.5$, the scale factor is equal to $\sqrt{2}$ and the target values are the original taste parameters times that factor. However, the calibrated parameters are affected by a scale effect that depends on the magnitude of the correlation actually calibrated by the model. When the number of repetitions is too small, the model cannot capture correlation effects, the Gumbel term explains almost all the variance, and the scale factor is almost one (not $\sqrt{2}$, as it should be). The ML parameters reported in Table 1 indicate that the empirical ratio between the calibrated and true parameters varies from an average of 0.7 to an average of 1.0, whereas the corresponding correlation parameters detected vary from less than 0.03 to more than 1.2. So the scale effect is variable, related to the balance between the Gumbel (independent) and the normal (common to induce correlation) components of the error term.

As Table 1 shows, in the case of probit, the smaller values of computing time were not obtained for the lower number of repetitions. An unstable behavior makes the process converge in more iterations, and even though each of those iterations take less time, the implementation with five repetitions took more total time to converge than that with 10 and even 25 repetitions.

As for the implementations with pseudo random numbers (MLR) and Halton random numbers (MLH), the values obtained from both are very close, including the statistics. However, in terms of correlation coefficient recovery, 50-repetition MLH has a behavior equivalent to 500-repetition MLR in terms of the confidence interval and the *t*-statistic against the target value.

In terms of processing time, in this implementation the use of Halton sequences appeared to improve efficiency for a moderate number of repetitions (25 to 250). As this number increased, the pseudo random number implementation was faster; the reason is that, when the number of repetitions is too high, the Halton sequences implementation requires a huge amount of memory to store the series, which could cause a lack of efficiency in the process. The processing time for probit was notoriously longer than that for either ML implementation. Of course, that is a result of this particular implementation, and more efficient codes probably are available to estimate both ML and probit; it is not known what the relation would be with processing time in those implementations. However, these values correspond to easily accessible codes.

As for model capacity to detect correlation, Figure 1 illustrates that the probit model can obtain biased punctual parameters. Even when the number of repetitions was increased, the parameters stabilized in a value different from the real parameter. However, the confidence intervals appear appropriate (the real parameter is contained in the interval) starting from 50 repetitions. With an MLR, the σ_μ estimation is appropriate from 100 repetitions. In general, the confidence interval for this parameter is adequate, but it presents a peculiar behavior for 50 repetitions (a check showed that it was not an error) in which the *t*-values are particularly high. When MLH is used, the parameter that captures correlation is unbiased. When the number of repetitions is increased, the parameter becomes stable, taking a value very close to the real parameter. In terms of number of repetitions, this behavior is achieved earlier than in the MLR implementation (25 repetitions versus 100).

Because the objective function in this optimization process is the (simulated) log likelihood, examination of its behavior was considered important. (Figure 2). A curious situation is observed for the probit model, because the highest value in the average log likelihood was obtained for 50 repetitions (–1.04472); it decreases with additional repetitions and stabilizes at a value lower than the maximum (–1.04519 for 1,000 repetitions). The MLH achieved log likelihood values larger than –1.045 for 25 repetitions, nearing –1.044 as they increased. In contrast, the MLR reached values greater than –1.045 starting from 250 repetitions.

To take this comparison a step further, the prediction capabilities of the models are evaluated by means of the response analysis; results are reported in Figure 3. The probit model achieved values under the critical value ($\chi^2_{95\%,3} = 7.815$) starting from 10 repetitions and quickly stabilized at very low values, near 3.5. In contrast, the MLH achieved values under the critical value for at least 25 repetitions, whereas MLR did it from 200 repetitions. The MLH stabilized at 5.4 (100 repetitions) and the MLR at 5.8 (500 repetitions). In that sense, probit and MLH behave better than MLR.

Number of Observations

This section reviews how NL, ML, and probit models behave when the sample size is varied. The databases were generated assuming homoscedasticity and correlation (as in Equation 3). In this particular case, NL is as appropriate as NML or probit to represent the correlation structure. To estimate the probit model, the GHK simulator with 10 repetitions was used and for the ML, 200 Halton repetitions (following recommendations from previous studies, even though it was found here that it would be possible to work with a lower number). Estimation results are reported in Table 2. The target values for the taste parameters reported include scaling for ML (calculated as described in the previous section). For both probit and NL models, the scale factor is equal to unity, so the target values are directly the taste parameters used for generating the database. The target value for the structural parameter of the NL model (φ) was calculated as $\sqrt{1-\rho}$ with ρ equal to 0.5 (a direct result of the covariance matrix presented in Equation 5).

The NL model recovers the structural parameter well, even for small sample sizes; a sample size under 8,000 NL has some difficulty reproducing certain parameters (e.g., travel cost). The results for the probit model show that, peculiarly, quite good results are obtained for the smallest sample size. Excluding this special case, the estimations improved when the sample size was increased, especially in regard to correlation. But the estimate of the standard deviation of the common stochastic term that causes correlation remained below the target value, even for a rather high sample size (16,000 observations). In Figure 4, correlation is underestimated. In the case of ML, the standard deviation of the additional stochastic term appears to be well estimated and significantly different from zero, independent of the number of observations. An important effect of the sample size on the confidence interval of the parameters was observed. Apart from the confidence intervals for the parameters associated with correlation (Figure 4), the confidence intervals for the SVT can be derived from values in Table 2. They include the target value in all cases, but it is acceptable in terms of wideness only in the cases of 4,000 or more observations.

TABLE 1 Calibration Results for Variable Number of Repetitions

			Number of Repetitions of the Simulation Procedure									
	Parameter	Target	5	10	25	50	100	200	250	500	750	1,000
Probit	Travel cost	−0.005	−0.0033 (−5.0)	−0.0033 (−4.9)	−0.0036 (−5.0)	−0.0037 (−4.9)	−0.0037 (−4.9)	−0.0037 (−4.9)	−0.0037 (−5.0)	−0.0038 (−5.0)	−0.0037 (−5.0)	−0.0038 (−5.0)
	Travel time	−0.08	−0.0584 (−15.0)	−0.0616 (−15.5)	−0.0638 (−18.4)	−0.0655 (−16.0)	−0.0647 (−15.9)	−0.0652 (−16.3)	−0.0656 (−16.3)	−0.0656 (−16.1)	−0.0654 (−16.9)	−0.0656 (−16.6)
	σ_μ	0.91	0.5263 (3.6)	0.6254 (4.7)	0.7382 (7.2)	0.8053 (6.9)	0.7746 (6.3)	0.7920 (6.6)	0.8024 (6.9)	0.8030 (6.8)	0.7947 (7.0)	0.8049 (7.0)
	SVT travel	16	17.7 (5.2)	18.7 (5.0)	17.7 (5.1)	17.7 (5.1)	17.5 (5.1)	17.6 (5.1)	17.7 (5.1)	17.3 (5.1)	17.7 (5.1)	17.3 (5.1)
	SVT access	32	37.3 (5.3)	38.9 (5.1)	37.7 (5.2)	37.7 (5.1)	37.2 (5.2)	37.5 (5.1)	37.7 (5.2)	36.7 (5.2)	37.6 (5.2)	36.8 (5.2)
	Iterations		10	7	6	6	6	6	6	6	6	6
	Average log likelihood		−1.0545	−1.0492	−1.0456	−1.0447	−1.0459	−1.0460	−1.0453	−1.0452	−1.0453	−1.0452
	CPU time (min)		25.1	15.5	24.0	41.5	72.7	130.5	136.3	325.0	438.8	684.4
MLR Mixed Logit Using Random Numbers	Travel cost	−0.007	−0.0041 (−5.3)	−0.0041 (−5.3)	−0.0041 (−5.2)	−0.0044 (−4.9)	−0.0050 (−4.9)	−0.0054 (−4.9)	−0.0054 (−4.9)	−0.0055 (−4.9)	−0.0056 (−5.0)	−0.0056 (−5.0)
	Travel time	−0.113	−0.0820 (−22.1)	−0.0821 (−22.0)	−0.0824 (−21.4)	−0.0877 (−15.3)	−0.0955 (−16.4)	−0.1000 (−16.7)	−0.1005 (−16.8)	−0.1011 (−16.7)	−0.1014 (−16.7)	−0.1014 (−16.7)
	σ_μ	1.28	0.0185 (0.4)	0.0486 (0.6)	0.1379 (0.8)	0.5799 (0.2)	0.9624 (5.5)	1.1639 (6.9)	1.1889 (7.1)	1.2124 (7.2)	1.2251 (7.2)	1.2235 (7.2)
	SVT travel	16	20.0 (5.2)	20.0 (5.2)	20.1 (5.2)	19.9 (5.1)	19.1 (5.0)	18.5 (5.0)	18.6 (5.0)	18.4 (5.0)	18.1 (5.1)	18.1 (5.1)
	SVT access	32	40.5 (5.3)	40.5 (5.2)	40.7 (5.2)	40.6 (5.1)	39.6 (5.1)	38.8 (5.1)	39.0 (5.1)	38.6 (5.1)	38.1 (5.1)	38.1 (5.1)
	Iterations		5	5	12	13	6	3	3	3	3	3
	Average log likelihood		−1.0479	−1.0478	−1.0478	−1.0477	−1.0465	−1.0451	−1.0449	−1.0449	−1.0448	−1.0449
	CPU time (min)		0.4	0.8	8.9	24.7	24.7	32.0	32.2	55.1	94.5	113.1
MLH Mixed Logit Using Halton Series	Travel cost	−0.007	−0.0041 (−5.3)	−0.0043 (−5.1)	−0.0054 (−5.0)	−0.0055 (−4.9)	−0.0057 (-4.9)	−0.0056 (−5.0)	−0.0057 (−5.0)	−0.0057 (−5.0)	−0.0057 (−5.0)	−0.0057 (−5.0)
	Travel time	−0.113	−0.0820 (−22.1)	−0.0858 (−18.1)	−0.0991 (−17.1)	−0.1009 (−16.7)	−0.1022 (−16.8)	−0.1021 (−17.0)	−0.1023 (−16.9)	−0.1023 (−16.9)	−0.1024 (−16.9)	−0.102 (−16.9)
	σ_μ	1.28	0.0299 (0.3)	0.4666	1.1105 (7.2)	1.2121 (7.1)	1.2632 (7.4)	1.2514 (7.6)	1.2621 (7.6)	1.2651 (7.7)	1.2686 (7.7)	1.2683 (7.7)
	SVT travel	16	20.0 (5.2)	20.0 (5.1)	18.4 (5.1)	18.3 (5.0)	17.9 (5.1)	18.2 (5.1)	17.9 (5.1)	17.9 (5.1)	18.0 (5.1)	18.0 (5.1)
	SVT access	32	40.5 (5.2)	40.6 (5.2)	38.3 (5.2)	38.6 (5.1)	37.8 (5.1)	38.4 (5.1)	37.8 (5.1)	37.9 (5.1)	37.9 (5.1)	37.9 (5.1)
	Iterations	6	14	3	3	3	3	3	3	3	3	
	Average log likelihood		−1.0479	−1.0474	−1.0449	−1.0450	−1.0445	−1.0444	−1.0444	−1.0443	−1.0443	−1.0443
	CPU time (min)		0.5	2.1	1.0	2.0	14.5	24.0	28.7	91.9	141.4	168.4

NOTE:
4,000 observations with correlated alternatives $\rho = 0.5$.
Average log likelihood = (log likelihood)/(number of observations).
Estimated parameters and (*t*-values against zero).

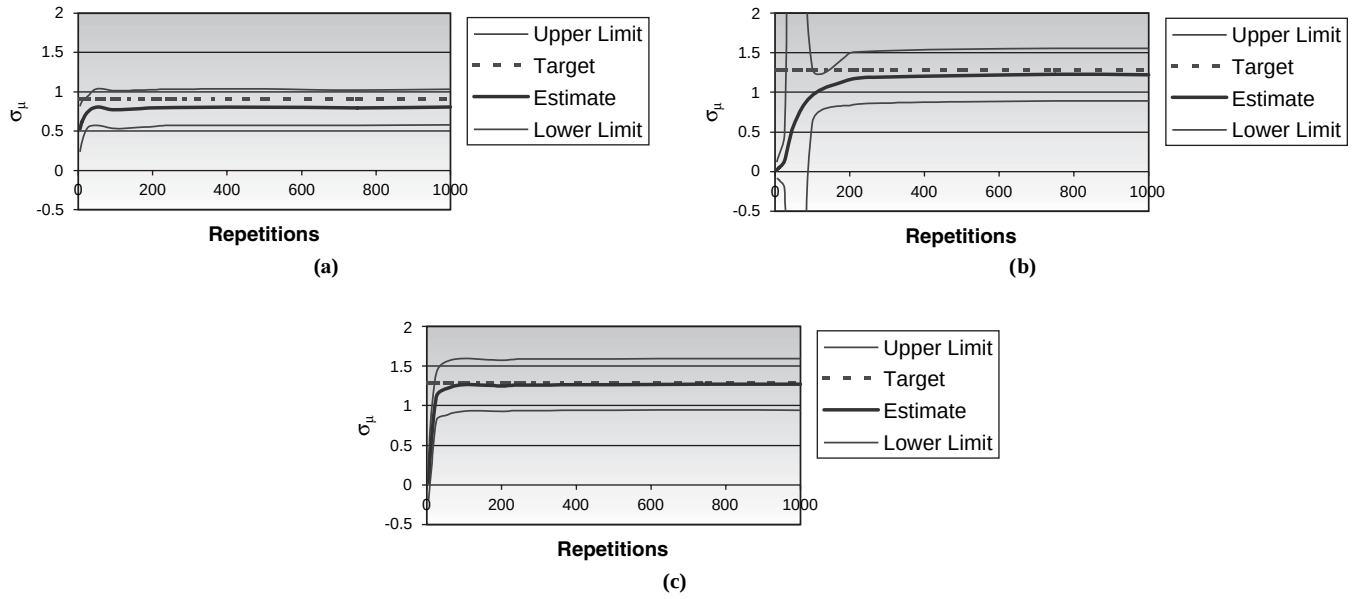

FIGURE 1 Confidence interval for correlation parameter versus number of repetitions.

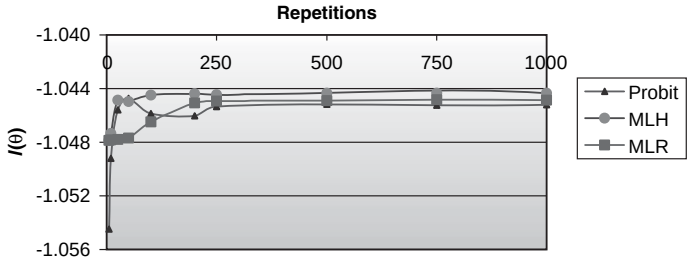

FIGURE 2 Average log likelihood versus repetitions.

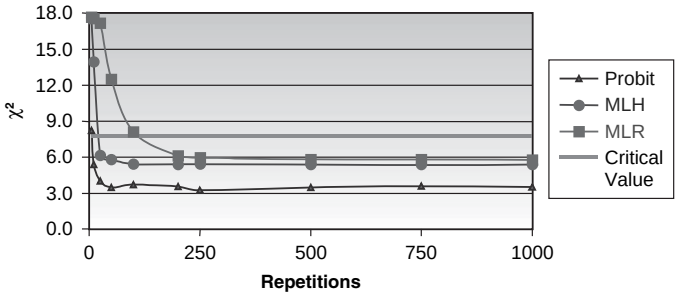

FIGURE 3 χ^2 index of difference between predicted and observed (simulation) values.

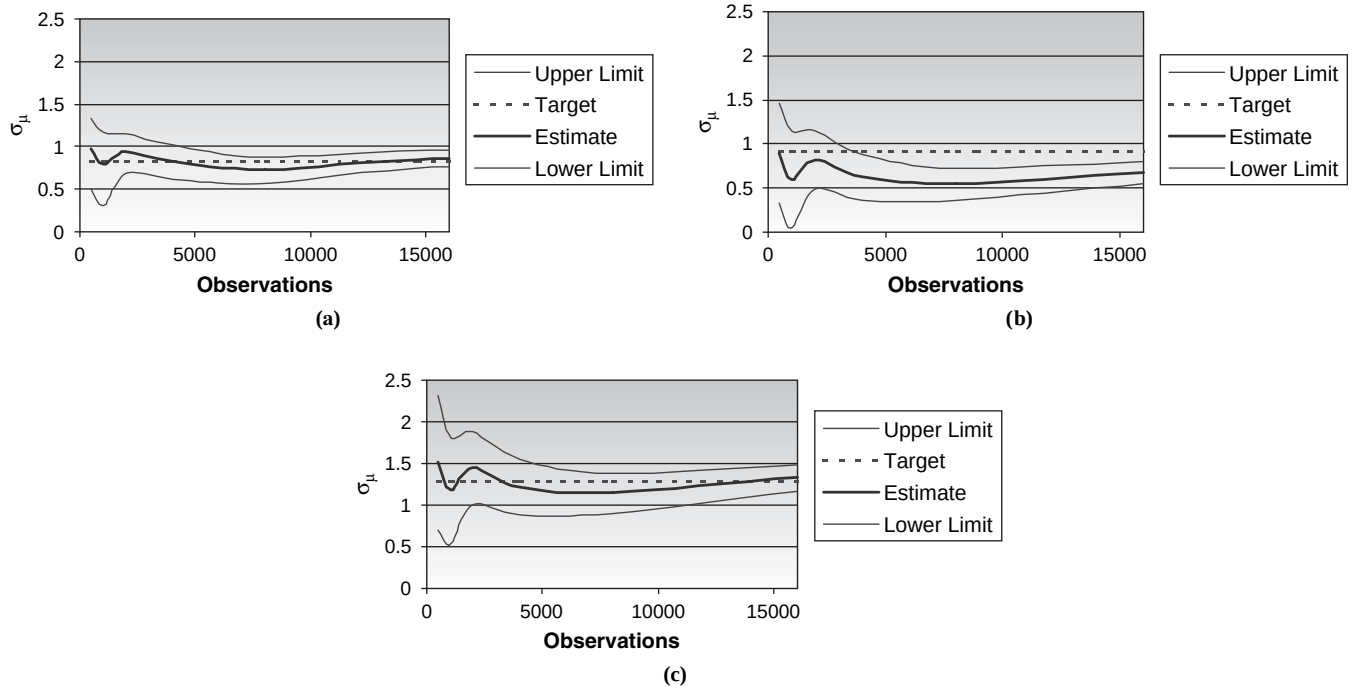

FIGURE 4 Confidence interval for the correlation parameter.

TABLE 2 Calibration Results for Variable Sample Size

			Sample Size					
	Parameter	Target	500	1,000	2,000	4,000	8,000	16,000
Nested Logit	Travel cost	–0.005	–0.0047 (–1.9)	–0.0029 (–1.7)	–0.0039 (–3.3)	–0.0040 (–5.2)	–0.0052 (–8.5)	–0.0054 (–12.6)
	Travel time	–0.08	–0.0859 (–7.4)	–0.0743 (–9.4)	–0.0736 (–14.1)	–0.0722 (–18.8)	–0.0760 (–27.9)	–0.0765 (–40.1)
	Access time	–0.16	–0.2080 (–11.4)	–0.1589 (–13.5)	–0.1579 (–19.0)	–0.1533 (–27.1)	–0.1643 (–38.8)	–0.1580 (–54.5)
	ϕ	0.71	0.6374 (6.4)	0.7216 (7.8)	0.6564 (11.7)	0.7048 (15.7)	0.7458 (22.6)	0.6873 (31.9)
	Iterations		8	6	5	4	5	3
	Average log likelihood		–0.94544	–1.04863	–1.04216	–1.04922	–1.03138	–1.04885
	CPU time (min)		0.1	0.1	0.2	0.2	0.8	1.3
Probit	Travel cost	–0.005	–0.0045 (–1.8)	–0.0023 (–1.6)	–0.0036 (–3.1)	–0.0033 (–4.9)	–0.0041 (–8.1)	–0.0048 (–11.8)
	Travel time	–0.08	–0.0830 (–6.6)	–0.0592 (–7.6)	–0.0693 (–11.4)	–0.0616 (–15.5)	–0.0614 (–21.9)	–0.0666 (–31.9)
	Access time	–0.16	–0.2008 (–7.1)	–0.1289 (–7.6)	–0.1454 (–12.4)	–0.1283 (–15.5)	–0.1323 (–23.7)	–0.1369 (–33.8)
	σ_μ	0.91	0.8985 (3.1)	0.5995 (2.1)	0.8172 (4.8)	0.6254 (4.8)	0.5441 (5.9)	0.6788 (10.8)
	Iterations		7	9	6	7	7	6
	Average log likelihood		–0.94544	–1.04863	–1.04216	–1.04922	–1.03138	–1.04885
	CPU time (min)		1.2	7.5	10.5	15.5	35.2	82.5
Mixed Logit	Travel cost	–0.007	–0.0072 (–1.9)	–0.0037 (–1.6)	–0.0060 (–3.2)	–0.0056 (–5.0)	–0.0071 (–9.7)	–0.0079 (–11.8)
	Travel time	–0.113	–0.1345 (–6.6)	–0.1023 (–8.5)	–0.1122 (–12.3)	–0.1014 (–16.7)	–0.1026 (–30.8)	–0.1109 (–33.8)
	Access time	–0.226	–0.3230 (–7.1)	–0.2176 (–8.9)	–0.2393 (–13.6)	–0.2131 (–17.5)	–0.2208 (–36.7)	–0.2285 (–36.9)
	σ_μ	1.28	1.5128 (3.6)	1.1779 (3.6)	1.4484 (6.4)	1.2235 (7.2)	1.1426 (9.4)	1.3325 (16.4)
	Iterations		6	4	4	3	2	3
	Average log likelihood		–0.9349	–1.0357	–1.0340	–1.0449	–1.0287	–1.0340
	CPU time (min)		14.3	27.5	32.6	24.0	238.3	632.4

NOTE:
Correlated alternatives $\rho = 0.5$.
Average log likelihood = (log likelihood)/(number of observations).
Estimated parameters and (*t*-values against zero).

For synthesis, sample size is an important variable in the model capacity to recover the parameters, especially those associated with correlation. This finding corroborates the results of Munizaga and Ortúzar (*16*), who recommend the use of 8,000 observations to obtain an interesting combination between statistical significance of the parameters and a good recovery within the confidence interval.

In general, for convergence analysis with a variable sample size, the use of flexible models that allow correlation does not present great difficulties or a particularly excessive use of resources for samples of a moderate size. In that sense, the probit model behaved better here than the ML (probit with 10 repetitions, MLH with 200 repetitions), but as in the previous section, the processing times are presented and commented only for completeness because they depend on the particular implementation.

Homoscedasticity and Heteroscedasticity

The nested ML model is naturally heteroscedastic but can be forced to homoscedasticity, whereas the traditional NL model is homoscedastic by construction. Probit models can accommodate both cases easily. To illustrate the differences, a case was implemented with 8,000 observations with a correlation coefficient equal to 0.5 between the bus and metro alternatives, where $\sigma_\mu = \sigma_\epsilon$. The heteroscedastic (Equation 2) and homoscedastic (Equation 3) databases were used.

The MNL, NL, probit, and ML estimation results are listed in Table 3. The probit and ML model parameters can be compared directly with the target values, because the probit covariance matrix was normalized in a way that the scale parameter is equal to unity, and the ML parameters were properly scaled to allow that comparison. They were divided by the known scale parameter, and enough repetitions and observations were used to make sure that correlation was well detected by the model, so the variable scale effect mentioned earlier was not present.

The MNL and NL parameters also can be compared directly with the targets in the homoscedastic case because they are affected by a factor of 1.0. However, the total variance is different for the different alternatives in the heteroscedastic case, so it is not known how the NL and MNL parameters will be affected by scale effects. Table 3 lists the parameter estimates for each model, the *t*-statistic against zero, and the *t*-test for the reference value of the parameter for the ML model. For the NL, the reference value of φ is calculated from the simulated correlation.

The ML model allows all the taste parameters used to generate the database to recover properly, as expected and indicated by the *t*-statistic, which is less than 1.96 in all cases (Table 3, *t*-value against target). In these results, the relationship between the NL estimates and the ML estimates in the heteroscedastic case are highlighted. The ratio between both parameters in each database is relatively constant (among implementations with variable magnitude of correlation, this ratio is larger in cases of more correlation). It can be explained by the scale effect when heteroscedasticity is present, which seems to affect all the parameters. In the ML model, the common error component (μ) is fixed to a certain value on each repetition of the simulation. Therefore, the scale factor of the Gumbel distribution is associated to the ϵ random term only: $\lambda = \pi\sqrt{6}\sigma_\epsilon$. However, for the NL model, even dismissing heteroscedasticity, it is the sum of both error components

TABLE 3 Calibration Results for Heteroscedastic and Homoscedastic Databases

		Heteroscedastic Database				Homoscedastic Database			
Parameter	Target	MNL	NL	Probit 10 Rep	ML 200 Rep	MNL	NL	Probit 10 Rep	ML 200 Rep
Travel cost	–0.005	–0.0070 (–11.4)	–0.0070 (–11.3)	–0.0049 (–10.9)	–0.0055 (–11.0) [–1.0]	–0.0053 (–8.9)	–0.0052 (–8.5)	–0.0041 (–8.1)	–0.0049 (–9.7) [0.2]
Travel time	–0.08	–0.1044 (–36.6)	–0.1005 (–32.0)	–0.0702 (–30.8)	–0.0804 (–31.2) [–0.1]	–0.0835 (–31.5)	–0.0760 (–27.9)	–0.0614 (–21.9)	–0.0791 (–30.8) [0.3]
Access time	–0.16	–0.2012 (–47.0)	–0.1954 (–41.3)	–0.1379 (–35.5)	–0.1563 (–36.1) [0.9]	–0.1765 (–44.9)	–0.1643 (–38.8)	–0.1323 (–23.7)	–0.1596 (–36.7) [0.1]
Income dummy	1.2	1.4928 (24.1)	1.4755 (24.0)	1.0686 (21.3)	1.1776 (21.5) [–0.4]	1.2454 (21.1)	1.2174 (20.8)	0.9998 (15.2)	1.1866 (21.8) [–0.2]
ϕ	0.71		0.8945 (24.0)				0.7458 (22.6)		
σ_μ	0.91			0.5100 (4.6)	0.7601 (8.4) [–1.6]			0.5441 (5.9)	0.8472 (9.4) [–0.7]
SVT travel	16	14.9 (10.9)	15.1 (10.8)	14.3 (10.3)	14.6 (10.4)	15.8 (8.6)	14.6 (8.2)	15.0 (7.6)	16.1 (9.3)
SVT access	32	28.7 (11.1)	27.9 (11.0)	28.1 (10.5)	28.4 (10.6)	33.3 (8.8)	31.6 (8.3)	32.3 (7.7)	32.6 (9.4)
Iterations		5	5	6	3	5	5	7	2
Average log likelihood		0.9347	–0.9343	–0.9369	–0.9329	–1.0318	–1.0292	–1.0314	–1.0287
CPU time (min)	0.6	0.8	35.5	42.5		0.7	0.8	35.2	152.5

NOTE:
8,000 observations.
Average log likelihood = (log likelihood)/(number of observations).
Estimated parameters, (*t*-values against zero), and [*t*-values against target].

that is supposed to be Gumbel distributed, so the scale is smaller. If all the alternatives had had the same variance as the error term, then the NL scale factor would have been $\lambda = \pi \Big/ \sqrt{6(\sigma_\epsilon^2 + \sigma_\mu^2)}$.

The prediction capabilities of the different models are evaluated with the use of the policy scenario reported in Table 4. In several cases, the model predictions are significantly different from the virtual reality. It is the authors' position that a model fails to predict the market shares when the χ^2 index is larger than the critical value ($\chi^2_{95\%,3} = 7.815$). The MNL failed in seven of 12 cases (excluding the base case, in which the market shares always are reproduced exactly). ML reproduces well the behavior of these virtual individuals that behave exactly according to the model assumptions, as expected, but it did fail twice. The behavior of the probit model, which was specified with the correct covariance matrix in each case, is similar to that of the ML, failing three times. Also as expected, the NL model behaves better in the homoscedastic case, in which the database was built with an error structure similar to that of the NL, the only difference being in some of the probability density functions.

The conclusion of this part of the analysis is that all models fail in some cases, but those models whose error structures are more similar to the real error structures of the data fail less often. Next, from the numbers just below the previous ones, the model predictions can be compared, just as could be done in a real-data case (when the underlying reality is not known). The ML was used as a basis, so these numbers reveal how different the MNL, NL, and probit predictions are from the ML predictions. Surprisingly, most predictions are not significantly different.

SYNTHESIS AND CONCLUSIONS

This paper has considered the most flexible and powerful models of the discrete choice family and, to analyze their empirical behavior, reviewed tests of them that were conducted in several ways. Even though the numerical results reported here come from observing the implementation in a particular case (synthetic data, small choice set, and parsimonious specification), the authors believe that they have varied the relevant parameters enough to make the resulting information a piece of empirical evidence valuable to users. The more relevant findings are synthesized below.

The number of repetitions generally used in practice for implementing ML and probit models by simulated maximum likelihood seems adequate, but many more observations than usually are available seem to be required to be able to recover a correlated error structure adequately. In a context like the one implemented in this study, it is suggested that 8,000 observations are enough to recover correlation properly. This warning is important for the use of flexible models with small sample sizes because erroneous conclusions could be obtained about the covariance structure if too much information is demanded from the data.

The use of Halton sequences to generate quasi-random numbers improves the efficiency of the ML model calibration process. Compared with the traditional method, the ML model process needed fewer repetitions to obtain the same quality of estimations and prediction capabilities, and the behavior of the log likelihood was more stable.

An unstable behavior of the log likelihood function was observed that makes application of the likelihood ratio test misleading. The average log likelihood did not increase monotonically with the number of repetitions, and it stabilized at different levels for probit, MLR, and MLH.

In all the results reported here, the probit model underestimated the correlation. It might be a coincidence; it might be something to do with the probability distribution assumptions. This subject should be investigated further.

ML models are subject to a scale effect that depends on how the model recovers correlation and do not depend only on the natural correlation present in the database. As a consequence, comparing the calibrated parameters with parameters calibrated with other models is difficult. Something similar happens with the MNL and NL when the database is not homoscedastic. The models somehow manage to estimate parameters that include a scale, but the authors cannot associate that scale with a particular variance. Fortunately, the ratio between parameters does not have a scale included, so SVTs can be compared directly.

TABLE 4 Difference Between Predicted and Simulated Values

	Heteroscedastic Database				Homoscedastic Database			
Policy Scenario	MNL χ^2 (χ^2)	NL χ^2 (χ^2)	Probit χ^2 (χ^2)	ML χ^2	MNL χ^2 (χ^2)	NL χ^2 (χ^2)	Probit χ^2 (χ^2)	ML χ^2
Base: no change	0.0 (0.0)	0.0 (0.0)	0.4 (0.3)	0.0	0.0 (0.0)	0.0 (0.0)	0.5 (0.4)	0.0
Car: cost ↑100%	**10.2** (1.7)	**10.6** (1.7)	**9.0** (0.9)	6.0	1.8 (0.1)	1.8 (0.1)	0.7 (1.6)	2.5
Car: cost ↑100% / access time ↑150%	**9.7** (1.8)	**9.8** (1.9)	7.5 (0.7)	5.7	2.5 (0.8)	0.9 (0.1)	0.8 (0.5)	1.4
Bus: cost ↑100% / access time ↓ 50%	4.3 (3.7)	3.4 (1.0)	1.2 (0.6)	1.4	**34.3** (**7.9**)	**11.8** (0.1)	**11.9** (0.8)	**11.4**
Bus: cost ↓ 50% / travel time ↑100% **Metro:** cost ↑ 50% / travel time ↓ 70%	**10.4** (1.0)	6.6 (0.2)	3.4 (2.2)	**8.0**	**11.6** (4.4)	5.4 (0.0)	**13.5** (2.3)	5.4
Car: access time ↑50%	**9.0** (3.1)	**8.2** (2.5)	4.0 (1.9)	5.2	**9.2** (1.3)	**8.3** (0.7)	1.5 (2.4)	6.0
Car: cost ↑100% / travel time ↓ 50% **Bus:** access ↑ 100% / travel time ↑50% **Metro:** cost ↓ 50%	1.9 (3.7)	3.4 (0.5)	3.3 (1.0)	2.6	4.2 (**13.0**)	5.0 (0.2)	4.9 (3.5)	6.6

NOTE:
χ^2 index of difference between the model predictions and the simulated market shares.
(χ^2) index of difference between the model predictions and the ML model predictions.
Bold typeface numbers mean the model fails to predict the correct market shares.

The results of calibrating a simple model (like the MNL) to a more complex reality indicate that even though the model fails to predict, it does not fail dramatically; and in some cases, its predictions might not be significantly different from those obtained with the correct model (in this case, ML).

ACKNOWLEDGMENTS

This study was partially funded by Fondecyt and Milenium Nucleus Complex Engineering Systems. The authors appreciate comments from Sergio Jara-Díaz and Juan de Dios Ortúzar on early versions of this paper.

REFERENCES

1. Hajivassiliou, V. A., and P. Ruud. Classical Estimation Methods for LDV Models Using Simulation. In *Handbook of Econometrics*, Vol. IV (R. Engle and D. McFadden, eds.). Elsevier, New York, 1994.
2. Bhat, C. R. A Heteroscedastic Extreme Value Model of Intercity Travel Mode Choice. *Transportation Research B*, Vol. 29, No. 6, 1995, pp. 471–483.
3. Brownstone, D., and K. Train. Forecasting New Product Penetration With Flexible Substitution Patterns. *Journal of Econometrics*, Vol. 89, 1999, pp. 109–129.
4. Williams, H. C. W. L., and J. de D. Ortúzar. Behavioural Theories of Dispersion and the Mis-specification of Travel Demand Models. *Transportation Research B*, Vol. 16, 1982, pp. 167–219.
5. Munizaga, M. A., B. G. Heydecker, and J. de D. Ortúzar. Representation of Heteroskedasticity in Discrete Choice Models. *Transportation Research B*, Vol. 34, 2000, pp. 219–240.
6. Train, K. *Discrete Choice Methods with Simulation.* Cambridge University Press, Cambridge, United Kingdom, 2002.
7. Hensher, D. A., and W. H. Greene. The Mixed Logit Model: The State of Practice. *Transportation*, Vol. 30, 2003, pp. 133–176.
8. Börsch-Supan, A., and V. A. Hajivassiliou. Smooth Unbiased Multivariate Probability Simulators for Maximum Likelihood Estimation of Limited Dependent Variable Models. *Journal of Econometrics*, Vol. 58, 1993, pp. 347–368.
9. Ben-Akiva, M., D. Bolduc, and J. Walker. *Specification, Identification and Estimation of the Logit Kernel (or Continuous Mixed Logit) Model.* Working Paper. Massachusetts Institute of Technology, Cambridge, 2001.
10. Gunn, H. F., and J. J. Bates. Statistical Aspects of Travel Demand Modelling. *Transportation Research A*, Vol. 16, 1982, pp. 371–382.
11. *GAUSS User's Manual.* Aptech Systems, Maple Valley, Calif., 1994.
12. Lerman, S. R., and C. F. Manski. On the Use of Simulated Frequencies to Approximate Choice Probabilities. *Structural Analysis of Discrete Data With Econometric Applications* (C. F. Manski and D. McFadden, eds.). MIT Press, Cambridge, Mass., 1981.
13. Bhat, C. Simulation Estimation of Mixed Discrete Choice Models Using Randomized and Scrambled Halton Sequences. *Transportation Research B*, Vol. 37, 2003, pp. 837–855.
14. Train, K. Free Software. elsa.berkeley.edu/~train/software.html.
15. Jara-Díaz, S. R., J. de D. Ortúzar, and R. Parra. Valor Subjetivo del Tiempo Considerando Efecto Ingreso en la Partición Modal (in Spanish). In *Actas del V Congreso Panamericano de Ingeniería de Tránsito y Transporte.* Universidad de Puerto Rico, Mayagüez, 1988.
16. Munizaga, M. A., and J. de D. Ortúzar. On the Applicability of the Multinomial Probit Model. *Proc., 25th European Transport Forum P415*, Association for European Transport, London, United Kingdom, 1997.

The Transportation Demand Forecasting Committee sponsored publication of this paper.