EFFICIENT ESTIMATION OF NESTED LOGIT MODELS: AN APPLICATION TO TRIP TIMING

Kenneth A. Small
Princeton University
and
David Brownstone
Princeton University
and
Stockholm School of Economics

Econometric Research Program
Research Memorandum No. 296

March 1982

Financial support from the National Science Foundation through grant no. SES-8007010 is gratefully acknowledged. The results and views expressed are the authors' responsibility and should not be attributed to the National Science Foundation. We are grateful to Kenneth Boese and Douglas Holtz-Eakin for highly creative and efficient research assistance, and to Jerry A. Hausman for comments on an earlier draft.

Econometric Research Program
Princeton University
207 Dickinson Hall
Princeton, New Jersey

ABSTRACT

This paper examines the Sequential, Full Information Maximum Likelihood (FIML), and Linearized Maximum Likelihood (LME) estimators for Nested Logit models of time-of-day choice for work trips. All three are consistent, but the first is not efficient. The efficiency gain from using FIML or LME is substantial, and the LME has modest computational costs. The sequential estimator is useful for preliminary specification checks and for getting consistent starting values for computing efficient estimators. However, the uncorrected sequential-estimator standard-error estimates from standard multinomial logit packages can be gross underestimates. We implemented a correction, but it is as difficult to program and nearly as costly to compute as the more efficient LME estimator which gives standard error estimates as a byproduct. Thus, we do not recommend the sequential procedure for final estimation. Although these results are based on a single data set, they are consistent with the few pertinent results reported by others.

We found that the nested logit model fits these data significantly better than multinomial logit, although the qualitative properties of the estimated indirect utility functions are no different.

1. INTRODUCTION

The practical application of qualitative choice models has presented econometricians with many difficult problems. The Multinomial Logit (MNL) model is easy to estimate and work with, but its adoption requires that the researcher accept the Independence from Irrelevant Alternatives (IIA) axiom. In many applications this assumption is clearly erroneous and its use causes serious biases. Until recently the only alternative was Multinomial Probit [see Hausman and Wise (1978)] which, though useful in cases with up to three discrete alternatives, is computationally intractible for larger choice sets.

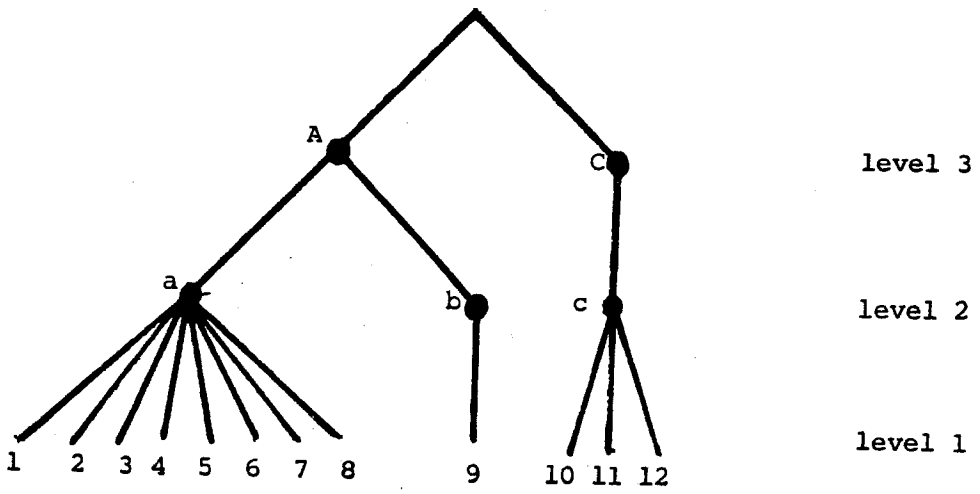
Recently McFadden (1981) has proposed a new class of qualitative choice models, called Generalized Extreme Value (GEV) models, which do not require the IIA assumption and are relatively easy to estimate. GEV models are consistent with random utility maximization in a consumer choice framework, and contain the MNL model as a special case. With the exception of Small (1982), the only practical applications of GEV models have been with a subclass called Nested Logit (NL). NL models have been used in transportation mode choice [Cosslett (1978)], consumer

durable choice [Brownstone (1980)], and household energy demand [Goett (1980)]. There are a number of practical problems including choice of estimation technique and computation of standard error estimates which have not been carefully addressed in this literature. This paper attempts to investigate these problems while using NL to model the choice of time-of-day for work trips.

The general NL model requires complex notation, for which the reader is referred to McFadden (1981) or Cosslett (1978). We use here a simplified notation suitable for the most general model considered in this paper. To describe an NL model, we first consider a hierarchy of groupings of alternatives called a tree structure and represented by a diagram such as Figure 1. At each level of the tree, the alternatives (or groups of alternatives) defined at the next lower level are grouped, indicated diagrammatically by connecting them to a node. Level 1 consists of all the alternatives, denoted j=1, ..., 12 in the figure. Level 2 consists of nodes indicating groups of alternatives; in the figure, the level-2 nodes are denoted r=a,b,c and correspond to groups $B_a = \{1,...,8\}$, $B_b = \{9\}$, and $B_c = \{10,11,12\}$. Level 3 consists of groups of level-2 nodes; the level-3 nodes L=A,C correspond to groups $B_A = \{a,b\}$ and $B_C = \{c\}$. The choice probability for alternative k attached to level-2 node s which in turn connects to level-3 node M is

Figure 1

Nested Logit Tree Structure



$$(1) \qquad P_k = P(k|s) \cdot P(s|M) \cdot P(M)$$

where

$$(1a) \qquad P(k|s) = \frac{\exp(V_k/\rho_s)}{\sum_{j \in B_s} \exp(V_j/\rho_s)} \equiv \frac{\exp(V_k/\rho_s)}{\exp(I_s)}$$

$$(1b) \qquad P(s|M) = \frac{\exp(\rho_s I_s/\rho_M)}{\sum_{r \in B_M} \exp(\rho_r I_r/\rho_M)} \equiv \frac{\exp(\rho_s I_s/\rho_M)}{\exp(I_M)}$$

$$(1c) \qquad P(M) = \frac{\exp(\rho_M I_M)}{\sum_L \exp(\rho_L I_L)}$$

$$(1d) \qquad V_j = \beta\, z_j \qquad j=1,\ldots,12$$

$$(1e) \qquad I_r = \log \sum_{j \in B_r} \exp(V_j/\rho_r) \qquad r=a,b,c$$

$$(1f) \qquad I_L = \log \sum_{r \in B_L} \exp(\rho_r I_r/\rho_L) \qquad L=A,C \ .$$

The ρ parameters must be in the unit interval, and must satisfy

$$(2) \qquad 0 < \rho_r/\rho_L \leq 1$$

whenever node r is attached to higher-level node L.

The strict utility V_j is specified to be a function of observable characteristics z_j describing both the alternative j and the individual in question, a function linear in unknown parameter vector β. In this paper, the additional unknown parameters $\{\rho_r\}$ and $\{\rho_L\}$ are referred to generally as "the ρ's", and specifically using the notation just introduced. The quantity I_r or I_L is called the inclusive value of node r or L. A branch of the tree corresponding to a given node consists of that node, the lower-level nodes attached to it, all nodes or alternatives attached to those, etc.; for example, {C,c,10,11,12} constitute one main branch of the tree in Figure 1. Nodes such as b and C which are attached to only one lower-level node are called degenerate. It is easily seen from the above equations that the conditional probability for the node just below a degenerate node is unity, and the ρ parameter for the degenerate node drops out of the choice probability formula. In Figure 1, for example, $\rho_b I_b = V_9$ and $\rho_C I_C = \rho_c I_c$.

The log-likelihood function is formed adding the natural logarithms of equation (1) for the chosen alternative of all members of the sample. The full information maximum likelihood (FIML) estimator of the unknown parameters is that which maximizes this likelihood function. The sequential estimator

takes advantage of the additive separability of the logarithm of (1) by performing a sequence of simpler maximizations. Denote a member of the sample by superscript i; denote the alternative chosen by that member and the nodes above it by k^i, s^i, and M^i. The log-likelihood function is then

$$L = \sum_i \log P^i_{k^i}$$

$$= \sum_i \log P^i(k^i | s^i) + \sum_i \log P^i(s^i | M^i) + \sum_i \log P^i(M^i)$$

$$\equiv L_1 + L_2 + L_3 .$$

The sequential estimator first estimates β/ρ_r by maximizing L_1 (first stage);[1] uses these estimates to compute I_r, then estimates ρ_r/ρ_L by maximizing L_2 (second stage); and so forth. When the ρ's at a given level are constrained equal, that stage involves maximizing a log-likelihood function exactly like that for the MNL model, permitting use of fast existing MNL algorithms.

---

[1] One or more components of β may not be identified at the first stage because the corresponding variables do not vary over alternatives within groups B_r. An example is a dummy variable equal to 1 for $j \in B_r$ and 0 otherwise. Such variables can simply be omitted in calculating inclusive value, and entered separately as additional variables at higher stages. To see this, partition $\beta = (\beta^1, \beta^2)$ and $z_j = (z_j^1, z_j^2)$, where $z_j^2 \equiv z_r^2$ for all $j \in B_r$ so that β^2 is not identified at the first stage. Then from (1 e),

$$I_r = \log \left[\exp (\beta^2 z_r^2) \cdot \sum_{j \in \beta_r} \exp (\beta^1 z_j^1) \right] = \beta^2 z_r^2 + I_r^1,$$

where I_r^1 is the inclusive value computed from (1 e) omitting variables z^2.

The sequential estimator, however, has a number of disadvantages. As is well known, it is not efficient, since information about higher-level choices is not used in estimating the lower-level parameters. Furthermore, the amount of information lost goes up dramatically with the number of distinct parameters describing the tree structure, making it difficult to obtain powerful tests of restrictions. Since the estimates at each level depend on parameters estimated at lower levels, errors may accumulate up the tree. Finally, the standard errors of the ρ's are incorrectly estimated by the MNL algorithm, and the correction factor required is quite complicated (see Appendix A).

This paper attempts to assess the practical significance of these difficulties by comparing the sequential estimator to FIML. One goal is to determine how much improvement can be obtained through efficient estimation. Previous experience is not very illuminating. Of the two estimates by Cosslett reported in McFadden (1981), the only one which noticeably raised the likelihood had an estimated ρ greater than one and hence was not a valid NL model. Brownstone (1980) found some models where FIML produced reasonable estimates, but he was unable to compute the sequential estimator due to his complicated tree structures and small sample size.

Sections 3 and 4 compare alternative estimators of an NL model for a particular empirical example described in the next section. Our results would be more general if we based comparisons on a number of different data sets

in a Monete Carlo framework, but this would involve much greater computer expense. Instead, we chose an example which has already been investigated thoroughly using MNL models, and for which NL is a plausible generalization. Our results support the desirability of efficient estimation of the NL model, and we hope they will provide guidance and encouragement to researchers using other data sets for which nested logit seems appropriate.

2. DATA: TIME-OF-DAY CHOICE

The empirical example is the choice of time-of-day for work trips, previously modelled by McFadden et. al. (1977), Small (1982), and Abkowitz (1980). Because of analytical difficulties with treating the choice as continuous, plus a tendency of respondents to round off replies to the nearest five minutes, all of these authors estimated an MNL model of choice among 12 discrete alternatives, each representing arrival at work within a particular 5-minute interval. The choice set consists of intervals centered from 40 minutes before to 15 minutes after the official work start time for the individual. Data were collected on the actual arrival times, the official work start times, and other characteristics of 527 individuals who commuted by auto to a major city in the San Francisco Bay area in 1972 (see McFadden et. al., 1977). These were supplemented with engineering calculations of the travel times each would have faced at each of the 12 alternative arrival times.

The most well-behaved MNL specification found in Small's work is shown in the first column of Table 1, with variables defined as follows:[1]

SD = Schedule Delay: actual arrival time minus official work start time, in minutes, for a given alternative. Thus its value for alternative j is $SD_j = 5(j-9)$, $j = 1,\ldots,12$.

R15 = $\begin{cases} 1 & \text{if } SD = -30, -15, 0, 15 \\ 0 & \text{otherwise.} \end{cases}$

---

[1] The sample used here is larger than that in Small (1982) because of reconstruction of some previously missing carpool data. The coefficient estimates are nearly identical.

R10 = $\begin{cases} 1 & \text{if SD} = -40, -30, -20, -10, 0, 10 \\ 0 & \text{otherwise.} \end{cases}$

TIM = Travel Time in minutes

SDE = Max. {-SD, 0}.

SDL = Max. {SD, 0}.

FLEX = Answer to question: "How many minutes late can you arrive at work without it mattering very much?".

SDLX = Max. {SD-FLEX, 0}.

D2L = $\begin{cases} 1 & \text{if SD} \geq \text{FLEX} \\ 0 & \text{otherwise.} \end{cases}$

CP = Dummy for car pool.

This model captures the trade-off between the desire to avoid congestion on the one hand, and the desire to avoid arriving too early or late on the other. The estimated marginal rates of substitution imply that the average non-carpooler would incur .53 minute of travel time to avoid arriving an extra minute early; 1.24 minute to avoid arriving an extra minute late; and an additional 1.53 minute to avoid arriving an extra minute beyond the reported employer's flexibility range. The implication for transportation analysis is that significant shifts in the timing and duration of the peak period will occur in response to any factor substantially affecting congestion, and that accurate predictions of traffic conditions must take this scheduling responsiveness into account. The results of the present paper further support this conclusion.

It is clear that the IIA assumption is not strictly appropriate here. At least two correlation patterns other than independence might plausibly be postulated for the unobserved preferences for these 12 alternatives. One,

explored by Small (1981), is induced by the ordering of the alternatives and involves a closer correlation among "nearby" alternatives. The other, explored here, assumes that commuters have unmeasured preferences for arriving early, on-time, or late, thereby inducing correlation within the corresponding groups of alternatives. Three groupings are considered: (1) alternatives 1-8 (early arrival) vs 9-12 (on-time or late); (2) alternatives 1-8 (early) vs. 9 (on-time) vs. 10-12 (late) as three distinct groups; and (3) 1-9 vs 10-12. These are indicated by the corresponding tree diagrams on the tables. In each case, the two nondegenerate level-1 nodes have parameters denoted by ρ_a and ρ_c; when constrained equal they are denoted by ρ. Note that tree structures (2) and (3) are special cases of the three-level tree of figure 1, corresponding to $\rho_A = 1$ and $\rho_A = \rho_a$, respectively.

3. SEQUENTIAL ESTIMATES

Sequential estimates were obtained using QUAIL, a versatile qualitative-choice computer program with a fast MNL algorithm and matrix manipulation capabilities. Because we wished to constrain certain parameters (e.g. the coefficient of travel time) to be identical on all branches of the tree, the first MNL stage was carried out by "stacking" the cases included in each of the two conditional choice problems. For example, in tree (1), the 318 individuals choosing an early alternative were included with possible choices 1-8, and the 209 individuals choosing an on-time or late alternative were given possible choices 9-12. By combining these into a single first-stage MNL estimation of the conditional lower-level choices, we estimated β/ρ with less loss of efficiency than if we had estimated some coefficients separately. To our knowledge, this procedure has not been mentioned in the general descriptions of the NL sequential estimator such as McFadden's (1981), though it has been used in practice [e.g. Train (1980)]. Of course, this limited us to assuming $\rho_a = \rho_c \equiv \rho$.

Of the two-level trees shown in Table 1, structure (1) does not fit the data as well as the other two structures: It achieves a much lower log likelihood, and it greatly underpredicts the fraction choosing alternatives 1-6. Evidently the first-stage coefficient estimates, especially of the lateness dummy D2L, are not sufficiently accurate for the inclusive value to have any explanatory power at the second stage. This suggests that on-time arrival (alternative 9) is viewed as distinctly different from late arrival, so that grouping them together gives a poor fit.[1]

Tree structure (2) involves substantial loss of efficienty in sequential estimation, which is manifested in large standard errors. This is because the 187 individuals choosing alternative 9 are dropped from the first-stage estimation, node b being degenerate. Tree structure (3) achieves the highest log-likelihood and a precision in $\hat{\beta}/\hat{\rho}$ nearly as good as the MNL model. However, the hypothesis that the true model is MNL cannot quite be rejected at a 15% significance level, using a one-tailed asymptotic t-test of the null hypothesis $\rho=1$ against $\rho<1$.

One of the bizarre features of the NL sequential estimator is that it is possible to obtain a lower log-likelihood than the MNL model which is a special case of NL. This occurs in tree structures (1) and (2). Structure (1) is especially misleading because a one-tailed t-test would reject the MNL model at a 5% significance level. A suitably modified version of one of the tests discussed in Hausman and McFadden (1981) might give better results, but this was not attempted here.

Table 1 shows both the correct standard errors on $\hat{\beta}$, computed as described in Appendix A, and the uncorrected standard errors on the coefficient of inclusive value as computed by the MNL algorithm in the second stage. Our results corroborate Cosslett's (1978) finding that the uncorrected standard errors are serious underestimates. This is especially true in the better-fitting models.

---

[1] This view is strengthened by the FIML estimate $\hat{\rho} = 1.95$ for this tree structure (Table 3) which indicates strong within-group dissimilarity.

Several more general models were estimated with poor results. We tried computing separate first-stage estimates on the two nondegenerate branches, in order to allow for distinct ρ_a and ρ_c; this led to very large standard errors at the second stage. Using a separate program written in the APL language, we computed sequential estimates with β/ρ_a and β/ρ_c constrained at the first stage to be equal up to a proportionality factor; this turned out to be as expensive as FIML but much less precise. Finally, two three-level trees were tried, the more promising of which is shown in Figure 1 and Table 2; once again, the sequential estimator managed to achieve a lower log-likelihood than was obtained for a special case, tree structure (3) of Table 1; and the estimated ρ's were not all in the unit interval. An attempt to improve the specification by adding explicit occupational variables (in addition to those implicit in D2L) deserves mention, since it illustrates in extreme form how imprecision in first-stage estimates leads to poor results at higher stages. For tree structure (2), as already noted, more than a third of the sample is lost from the first stage estimation. The variable D2L is always zero on alternatives 1-8; furthermore, there is only one individual for whom D2L varies within group $B_c = \{10, 11, 12\}$ and who chooses an alternative with the higher value of D2L. This individual happens to be one of the five with missing occupational data, so with those five dropped from the sample the dummy variable D2L becomes a perfect discriminator at the first stage: L_1 is maximized with coefficient of D2L equal to $-\infty$. This means that inclusive value is $-\infty$ for any node connected to an alternative for which D2L = 1. However, at the second stage many individuals choose such a node despite the availability of node a for which D2L=0; this forces the coefficient $\hat{\rho}$ on inclusive value to be zero. Thus, the loss of information at the first stage leads to absurd results.

4. EFFICIENT ESTIMATES

Efficient estimates for two- and three-level trees were obtained using a modified Newton-Raphson algorithm written in the APL computer language. Two efficient estimates are reported here: Full Information Maximum Likelihhod (FIML), and the "Linearized Maximum-likelihood Estimator" (LME) resulting from one Newton-Raphson step.[1] The program was not designed primarily for speed, and was run on a slower operating system. As a result the estimates were quite expensive: The typical two-level tree required 6 iterations and 150 seconds of central processing time on an IBM 3033 computer, at about 6 times the cost of the sequential estimator's / converge. A single standard error calculations, to iteration, in contrast, cost about as much as the sequential estimator's standard errors

The FIML results[2] in Tables 2 and 3 are quite encouraging. Of the two-level trees, structure (3) has the highest likelihood, corroborating the sequential estimator. However, the precision is much better than with the sequential estimator, and it is now possible to reject fairly confidently the MNL model. The chi-squared statistic for testing constrained model (3) against the MNL is 3.49 with one degree of freedom, significant at the 10% level. Even this is overly conservative since we require the ρ's to be in the unit interval, and in fact reject better-fitting models such as the unconstrained version of (3) on this ground. A more appropriate test is a one-tailed test of $\rho=1$ against $\rho<1$ based on the asymptotic t-ratio $(1-\hat{\rho})/SE(\hat{\rho}) = 1.87$, which rejects the MNL hypothesis at a 3.1% level of significance.[3]

---

[1] As shown by Rothenberg (1973), one such step starting from consistent estimates is asymptotically efficient.

[2] The convergence criterion was that no coefficient changed by more than one percent in the last iteration. For case (1) of Table 1 we also checked the weighted gradient recommended by Belsley (1980), which was .056, indicating a quadratic approximation to logL would have a maximum .014 above the value achieved by the last iteration. Spot checks were made to insure against saddle points and multiple maxima. Additional details about the program are given in Appendix B.

[3] The strictly correct test for this hypothesis would require estimating the model subject to the constraint $0 < \rho \leq 1$ and generalizing Gourieroux and Monfort's (1979) method of testing on parameter-set boundaries.

Furthermore, the FIML estimator allows reasonably powerful tests of constraints on the ρ parameters. For model (2), we can clearly accept the constraint $\rho_a = \rho_c$. For model (3), the same constraint might be rejected based on a chi-squared test, but this would be dubious because the unconstrained model yields unreasonable values of the ρ's (it is possible, however, that maximizing the likelihood while constraining the ρ's to be in the unit interval would yield an acceptable model and still reject the equality constraint). Both models (2) and (3) can be tested against the more general tree structure of Figure 1 and Table 2 based on the local maximum found for that tree structure involving ρ's which are positive (though still not in the unit interval). Chi-squared tests at a 5% significance level reject model (2) (unconstrained version) but accept model (3) (constrained version). This is additional support for the validity of model (3).

The FIML procedure gives estimates of β/ρ qualitatively similar to those from the MNL and from the sequentially estimated NL. In only one case did a coefficient change by more than one standard deviation from the cruder estimates. Furthermore, the sequential estimator correctly selected the best tree structure. Nevertheless, as shown in the last four rows of Tables 1 and 3, some key marginal rates of substitution differ by up to 33% from the MNL and 39% from the sequentially estimated NL. These results lead us to recommend a preliminary screening of models using the relatively cheap sequential estimator, then applying FIML (or the cheaper variant discussed below) to the more promising models for more precise estimation and hypothesis testing.

If the (consistent) sequential estimator is used as starting values, then each iteration towards the FIML estimate is asymptotically efficient. The first iteration may be a useful estimator in its own right due to its low computation costs. This estimator, called Linearized Maximum Likelihood (LME), is compared with the others in Table 4. Only the two parameters which were most volatile, along with their standard errors and the log likelihood, are shown.[1] The LME correctly chose the best tree structure, and for the most part produced coefficient estimates considerably closer to the FIML estimates than did the sequential procedure. In all three cases the log likelihood was raised in the first iteration by at least 58 percent of the difference between the FIML and the sequential values. For all but one coefficient the LME produced standard error estimates smaller than the sequential procedure and close to those of FIML. Finally, for the two poorer-fitting tree structures for which the LME estimates differed by more than a few percent from FIML, the fact that the log likelihood was still lower than that achieved by MNL would serve as a warning that more iterations might substantially change things. In short, performing one iteration using the sequential estimates as starting values provides much of the benefit of FIML at considerably less cost.

5. CONCLUSIONS

The lessons learned from this study are primarily based on the empirical example discussed in this paper. However, since most of our conclusions are corroborated by the work of Brownstone (1980) and Cameron (1982) they are probably valid in other situations as well.

---

[1] The results shown in Table 4 are from a full Newton-Raphson step, without the cubic polynomial interpolation employed in the FIML algorithm (see Appendix B). Thus the programming needed to obtain them would be relatively easy. The use of interpolation probably would have reduced the tendency toward oscillation in parameter estimates characterizing Table 4.

The sequential estimator appears to be useful for identifying promising models, since it is easy to compute. In this study it successfully screened tree structures for the most promising one, and gave coefficient estimates qualitatively similar to FIML for that tree structure. More importantly, the sequential estimator provides unique consistent starting values for computing efficient estimators.

Unfortunately, sequential estimators also have a number of serious disadvantages. The efficiency losses relative to FIML were quite large in this study. In some examples the sequential estimator could not distinguish the NL structure from a simple MNL model, but the FIML estimator clearly rejected the MNL model. Furthermore it is clear that these efficiency losses will get worse as more levels are added to the NL tree structure. We also found situations where the efficiency losses make it impossible to use the sequential estimator (e.g. loss of observations at the first stage causes unbounded coefficients). Although sequential estimates can be computed quite easily using standard MNL computer packages, the standard error estimates produced by these packages seriously underestimate the true standard errors for all

parameters not identified at the first stage. Therefore, if the sequential estimator is going to be used for hypothesis testing, the standard errors must be corrected. Appendix A shows that these correction formulas are complicated even for simple two-level NL models, and the need to compute them greatly reduces the sequential estimator's computational advantages. Of course, it is not necessary to compute the correct standard errors if the sequential estimates are just being used as starting values for computing efficient estimators.

Compared with the sequential estimator, FIML estimators are much more precise and therefore permit more powerful hypothesis testing. Their main disadvantage is computation cost, but this can probably be reduced considerably by more careful programming. Also, if consistent sequential estimators are used as starting values for the FIML calculations, then it is not necessary to iterate to convergence to get asymptotically efficient estimators. In particular, the LME appears to offer a large gain in efficiency with small computational costs. With careful programming, computing the LME should cost no more than computing the correct standard errors for the sequential estimator. This study, based on one data set, suggests that the LME is the best estimator for NL models; further confirmation awaits a Monte Carlo study of the true sampling distributions of the estimators considered here.

One of the difficulties with using NL models in applied work is that the tree structure must be largely specified a priori. It appears from this study and work by Cameron (1982) that the constraints on the ρ's (i.e. that they lie in the unit interval) are very useful for sorting out possible tree structures, since many models can be discarded if the estimates violate these constraints. This fact increases the importance of estimating these parameters efficiently.

Together with other work mentioned in the introduction, this study has shown that nested logit models are attractive, practical qualitative choice models for situations where the Independence from Irrelevant Alternatives axiom is not justified. It is entirely feasible to construct fast algorithms which compute the sequential estimators and, using these for starting values, the asymptotically efficient one-step linearized maximum likelihood estimator. Poor results often indicate an inappropriate model, but can be further explored if desired by performing more iterations. Hypothesis tests of various special cases, including multinomial logit, can be performed to determine a structure which is general enough to be consistent with the data yet estimable with reasonable precision. In a case such as ours in which nested logit is a priori plausible, one can hope to obtain reasonable results with a model that fits demonstrably better than the popular but restrictive multinomial logit.

Table 1

Two-Level Trees:

Sequential Estimates

Tree Structure:	MNL	Nested Logit (1) 1-8 9-12	Nested Logit (2) 1-8 9 10-12	Nested Logit (3) 1-9 10-12
$\hat{\beta}/\rho$ (S.E.)				
R15	1.106 (.101)	1.076 (.117)	1.133 (.129)	1.142 (.104)
R10	.398 (.102)	.368 (.119)	.416 (.128)	.427 (.104)
TIM	-.141 (.053)	-.169 (.070)	-.195 (.072)	-.166 (.056)
TIM·CP	.105 (.076)	.124 (.101)	.163 (.104)	.137 (.078)
SDE	-.075 (.006)	-.067 (.008)	-.069 (.009)	-.076 (.006)
SDE·CP	.023 (.009)	.002 (.013)	.004 (.013)	.024 (.009)
SDL	-.175 (.029)	-.191 (.034)	-.192 (.081)	-.189 (.073)
SDLX	-.216 (.081)	-.216 (.083)	-.310 (.209)	-.312 (.208)
D2L	-1.057 (.170)	-0.018 (.666)	-1.015 (1.197)	-1.134 (.174)
$\hat{\rho}$ (S.E.)		.342 (.377)	.882 (.419)	.843 (.152)
[uncorr.S.E.]		[.328]	[.075]	[.067]
Log Likelihood	-994.90	-1030.95	-998.12	-994.03
$\bar{P}_{1-6}$(actual=.378)	.385	.311	.374	.385
$-(\partial SDE/\partial TIM)_V$				
Noncarpoolers	1.88	2.52	2.83	2.18
Carpoolers	0.69	0.69	0.49	0.56
$-(\partial SDL/\partial TIM)_V$				
Noncarpoolers	0.81	0.88	1.02	0.88
Carpoolers	0.21	0.24	0.17	0.15

Notes:

Dependent variable is choice among 12 time-of-day alternatives, each a 5-minute arrival interval. Alternative 9 is on-time arrival.

No. cases = 527.

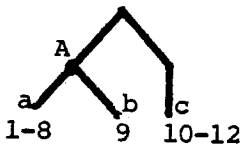
Asymptotic standard errors are in parentheses.

Log likelihood is the sum of the log likelihoods at each of the two stages, $L = L_1 + L_2$, as in equation (A1), Appendix A.

Table 2

Three-Level Tree [a]

Tree Structure:



	Sequential	FIML	
		Local Max with ρ's > 0	Global Max
Parameter Estimate (S.E.):			
$\hat{\rho}/\rho_A$	1.045 (1.301)		
$\hat{\rho}_A$	.821 (.263)		
$\hat{\rho}_a$		.708 (.506)	-1.1 (1.0)
$\hat{\rho}_c$		1.449 (1.388)	8.8 (7.4)
$\hat{\rho}_A$		.278 (.257)	-.87 (.69)
Log Likelihood	-997.19	-992.00	-991.25
χ^2 stat. for $\rho_A=\rho_a=\rho_c$ (deg. freedom)		2.90 (2)	
χ^2 stat. for $\rho_A=1$ (deg. freedom)		4.84 (1)	

[a] See notes to Table 1

Table 3
Two-level Tree:
FIML Estimates [a]

Tree Structure	(1) 1-8 9-12	(2) 1-8 9 10-12		(3) 1-9 10-12	
β/ρ (S.E.)	con-strained	con-strained	uncon-strained [b]	con-strained	uncon-strained
R15	1.067 (.100)	1.134 (.110)	1.137	1.145 (.104)	1.160
R10	.365 (.099)	.419 (.108)	.419	.429 (.104)	.394
TIM	-.077 (.050)	-.163 (.060)	-.165	-.148 (.054)	-.168
TIM-CP	.050 (.057)	.129 (.084)	.131	.115 (.077)	.151
SDE	-.069 (.007)	-.075 (.007)	-.076	-.075 (.006)	-.076
SDE-CP	.017 (.009)	.023 (.010)	.029	.023 (.009)	.024
SDL	-.188 (.030)	-.207 (.050)	-.223	-.237 (.057)	1.674
SDLX	-.210 (.081)	-.281 (.128)	-.291	-.338 (.148)	.426
D2L	-.529 (.356	-1.314 (.362)	-1.343	-1.109 (.174)	-1.134
$\hat{\rho}$	1.953 (1.095)	.807 (.178)		.761 (.128)	
$\hat{\rho}_a$			.789 (.223)		-.936 (.711)
$\hat{\rho}_c$			.865 (.568)		8.4 (6.9)
Log Likelihood	-993.68	-994.43	-994.42	-993.45	-991.41
$\bar{P}_{1-6}$ (actual=.378)	.385	.386	.386	.386	.388
$-(\partial SDE/\partial TIM)_V$:					
noncarpoolers	1.12	2.17	2.17	1.97	2.21
carpoolers	0.52	0.65	0.72	0.63	0.33
$-(\partial SDL/\partial TIM)_V$:					
noncarpoolers	0.41	0.79	0.74	0.62	-0.10
carpoolers	0.14	0.16	0.15	0.14	-0.01

[a] See notes for Table 1

[b] Standard errors were computed for $\hat{\beta}$ but not $\hat{\beta}/\hat{\rho}_a$

Table 4

Two-Level Trees:

Abbreviated Comparison of Sequential, Linearized Maximum Likelihood, and Full Information Maximum Likelihood Estimators[a]

Tree Structure:	(1) 1-8 9-12			(2) 1-8 9-12			(3) 1-9 10-12		
β/ρ (S.E.):	Seq.	LME	FIML	Seq.	LME	FIML	Seq.	LME	FIML
D2L	7.018	-2.218	-.529	-1.015	-1.464	-1.314	-1.134	-1.109	-1.109
	(.666)	(1.045)	(.356)	(1.197)	(.394)	(.362)	(.174)	(.174)	(.174)
$\hat{\rho}$	.342	.689	1.953	.882	.650	.807	.843	.735	.761
(S.E.)	(.377)	(.318)	(1.095)	(.419)	(.134)	(.178)	(.152)	(.121)	(.128)
Log L	-1030.95	-1009.26	-993.68	-998.12	-995.48	-994.43	-994.03	-993.53	-993.45

[a]See notes for Table 1.

Appendix A: Standard Error Formulas for the Sequential Nested Logit Estimator.

The sequential estimator described in the first part of this paper is calculated using MNL estimation programs at each level of the tree structure. Amemiya (1978) first pointed out that the standard error estimates produced by the MNL packages are downward biased except for those parameters identified at the lowest level of the tree structure. McFadden (1981) gives formulas for the correct asymptotic standard arrors, but the notation used is difficult to translate into computational formulas. This appendix specializes McFadden's formulas to the 2-level Nested Logit model with a single ρ parameter.

Following the notation used in the paper, we have:

$$(A1) \qquad L = \sum_i \log P^i(k^i|s^i) + \sum_i \log P^i(s^i) \equiv L_1 + L_2.$$

Partition β and z into two groups $\beta = (\beta^1, \beta^2)$ and $z = (z^1, z^2)$ such that β^2 is not identified at the first stage (i.e. via maximization of L_1). Let z^1 and z^2 be the quantities actually entered as variables at stages 1 and 2, respectively; and γ^1 and γ^2 their coefficients. Thus $z^1=z^1$, $\gamma^1 = \beta^1/\rho$, $z^2 = (z^2, I^1)$, and $\gamma^2 = (\beta^2, \rho)$, where I^1 is the inclusive value from the first stage. z^1 takes on distinct values $z^1_{k,s,n}$ for alternative k attached to nade s for individual n; whereas z^2, which by construction does not vary among alternatives attached to a given node, takes on values $z^2_{s,n}$.

Let

$$(A2) \qquad M_{ij} = E\left(\frac{\partial L_i}{\partial \gamma^i} \cdot \frac{\partial L_i'}{\partial \gamma^j}\right) .$$

McFadden (1981) shows that the asymptotic covariance matrix of the sequential estimator of γ is consistently estimated by

$$(A3) \qquad V = \begin{pmatrix} M_{11}^{-1} & -M_{11}^{-1}M_{21}'M_{22}^{-1} \\ -M_{22}^{-1}M_{21}M_{11}^{-1} & M_{22}^{-1}+M_{22}^{-1}M_{21}M_{11}^{-1}M_{21}'M_{22}^{-1} \end{pmatrix}$$

M_{ii} is just the Information Matrix for the MNL likelihood function at the i:th stage, so the MNL computer packages produce standard error estimates asymptotically equal to M_{ii}^{-1}. It is clear from formula (A3) that these "uncorrected" estimates are correct only for γ^1 and are downward biased for γ^2.

Define the random variables $S_{k,s,n}$ to equal 1 if individual n chooses alternative k attached to mode s. Then $E\, S_{k,s,n} = P^n(k|s)$, and if $S_{s,n} \equiv \sum_k S_{k,s,n}$, $E\, S_{s,n} = P^n(s)$.

Using this notation we have

$$(A4) \qquad L_1 = \sum_n \sum_s \sum_k S_{k,s,n} \log P^n(k|s)$$

and

$$(A5) \qquad L_2 = \sum_n \sum_s S_{s,n} \log P^n(s)$$

Differentiating (A4) and (A5) yields

$$(A6) \qquad \frac{\partial L_1}{\partial \gamma^1} = \sum_n \sum_s \sum_k S_{k,s,n}(Z^1_{k,s,n} - \bar{Z}^1_{s,n}) \quad \text{where} \quad \bar{Z}^1_{s,n} = \sum_k P^n(k|s) Z^1_{k,s,n}$$

$$(A7) \qquad \frac{\partial L_2}{\partial \gamma^1} = \sum_n \sum_s \rho\, S_{s,n}\left(\bar{Z}^1_{s,n} - \left(\sum_t P^n(t)\bar{Z}^1_{t,n}\right)\right)$$

$$(A8) \qquad \frac{\partial L_2}{\partial \gamma^2} = \sum_n \sum_s S_{s,n}(Z^2_{s,n} - \bar{Z}^2_n) \quad \text{where} \quad \bar{Z}^2_n = \sum_s P^n(s) Z^2_{s,n}$$

Note that $\frac{\partial L_1}{\partial \gamma^2} = 0$, and therefore $M_{12} = 0$. Taking expectations of (A6-A8) we have:

$$\text{(A9)} \qquad M_{11} = E\left[\frac{\partial L_1}{\partial \gamma^1} \quad \frac{\partial L_1}{\partial \gamma^1}'\right]$$

$$= \sum_n \sum_s \sum_k (Z^1_{k,s,n} - \bar{Z}^1_{s,n}) P^n(k|s) \; (Z^1_{k,s,n} - \bar{Z}^1_{s,n})'$$

$$\text{(A10)} \qquad M_{21} = \sum_n \sum_s (Z^2_{s,n} - \bar{Z}^2_n) \; \rho \; P^n(s) \left[\bar{Z}^1_{s,n} - \sum_t \bar{Z}^1_{t,n} P^n(t)\right]'$$

$$\text{(A11)} \qquad M_{22} = \sum_n \sum_s (Z^2_{s,n} - \bar{Z}^2_n) \; P^n(s) \; (Z^2_{s,n} - \bar{Z}^2_n)'$$

Formulas (A9)-(A11), with probabilities evaluated at the sequential estimates of γ, were used in formula (A3) to produce the correct standard errors in Table 1. The "uncorrected standard errors" are simply the square roots of the diagonal elements of M^{-1}_{ii}.

Similar formulas were used for the 3-level trees used in this paper. For these and more general formulas the reader is referred to McFadden (1981).

Appendix B: Estimation Programs

The sequential estimates were calculated using the QUAIL qualitative choice analysis computer package (see Berkman and Brownstone (1979)), and a QUAIL program for computing the estimates and the correct standard errors is available from the authors. The QUAIL package can be obtained from Cambridge Systematics, Inc.

The FIML estimates were calculated using some APL programs described in this appendix. These programs are also available from the authors, but they are written specifically for the particular tree structures used in this study and would have to be modified for other tree structures. APL is a high-level interactive language that is easy to use but expensive to execute. Therefore the costs for doing FIML would probably drop considerably if the programs were translated into a language like FORTRAN, although this translation would involve a lot of programming time. More computer savings could be found by reprogramming the likelihood function evaluation routines to reduce evaluation of the exponential function by more careful use of temporaries, but this would also involve a lot of programming time. Finally, APL uses double precision for all calculations on IBM computers. Although double precision is probably needed in some places, its use everywhere undoubtedly increases the computation costs.

The algorithm used to calculate the FIML estimates is a modification of the method of scoring which converges very quickly (i.e. 5-6 iterations for most of the models in this study) but requires computation of analytic derivatives of the log likelihood function, L. The iteration step is:

$$\hat{\beta} = \tilde{\beta} + d \left[E\left(\frac{\partial L(\tilde{\beta})}{\partial \beta} \frac{\partial L(\tilde{\beta})}{\partial \beta}' \right) \right]^{-1} \frac{\partial L(\tilde{\beta})}{\partial \beta}$$

where $\tilde{\beta}$ is the initial guess or value from the previous iteration and d is a scalar calculated to maximize L along the direction vector given by the product of the last two terms. Specifically, d is chosen to maximize a cubic polynomial fitted to L along the direction vector. In addition, d is decreased if necessary to insure that the algorithm always moves uphill. This algorithm also provides a consistent estimate of the covariance matrix for β since the matrix inside the square brackets is just the Information matrix for β. A version of the Berndt, Hall, Hall and Hausman (1974) algorithm was also tried where the analytic covariance matrix inside the brackets was replaced by its method-of-moments sample estimator. This algorithm did not reduce computation costs very much, and the resulting covariance estimator for β seemed to be more unstable.

It is quite possible that some other non-linear maximization algorithm will be more efficient for this problem. Unfortunately, time and budget constraints prevented us from trying other algorithms.

Appendix C

Acronyms used in this paper:

FIML	Full Information Maximum Likelihood
GEV	Generalized Extreme Value
IIA	Independence from Irrelevant Alternatives
LME	Linearized Maximum-Likelihood Estimator
MNL	Multinomial Logit
NL	Nested Logit

REFERENCES

Abkowitz, Mark D.: The Impact of Service Reliability on Work Travel Behavior, Ph.D. Dissertation, M.I.T., 1980.

Amemiya, Takeshi: "On a Two-step Estimation of a Multivariate Logit Model," Journal of Econometrics, 8 (1978), 13-21.

Belsley, David A.: "On the Efficient Computation of the Nonlinear Full-Information Maximum-Likelihood Estimator," Journal of Econometrics, 14 (1980), pp. 203-225.

Berkman, J., D. Brownstone, et. al.: QUAIL 4.0 User's Manual, Computer Center, University of California, Berkeley (1979).

Berndt, E. K., B. H. Hall, R. E. Hall, and J.A. Hausman, "Estimation and Inference in Nonlinear Structural Models," Annals of Economic and Social Measurement, 3 (1974), pp. 653-665.

Brownstone, David: An Econometric Model of Consumer Durable Choice and Utilization Rate, Ph.D. Dissertation, University of California, Berkeley, 1980.

Cameron, Trudy A.: Qualitative Choice Modeling of Energy Conservation Decisions . . ., unpublished Ph.D. Dissertation, Department of Economics, Princeton University (1982).

Cosslett, S.: Efficient Estimation of Discrete-Choice Models from Choice-Based Samples, Ph.D. Dissertation, Department of Economics, University of California, Berkeley (1978).

Goett, A.: "A Structured Logit Model of Appliance Investment and Fuel Choice," mimeo, Cambridge Systematics, Inc./West, Berkeley, Calif. (1979).

Gourieroux, C. and A. Montfort, "Testing for a Null Hypothesis on the Boundary of the Parameter Set," Document de Travail, E.N.S.A.E.-I.N.S.E.E. No. 7908 (1979).

Hausman, J. A. and D. McFadden, "Specification Tests for the Multinomial Logit Model," Dept. of Economics Working Paper Number 292, Massachusetts Institute of Technology, October 1981.

Hausman, J.A. and D. A. Wise: "A Conditional Probit Model for Qualitative Choice: Discrete Decisions Recognizing Interdependence and Heterogeneous Preferences," Econometrica, 78 (1978), pp. 403-426.

McFadden, Daniel: "Econometric Models of Probabilistic Choice," in C. F. Manski and D. McFadden (Eds.), Structural Analysis of Discrete Data with Econometric Applications (Cambridge, Massachusetts: M.I.T. Press, 1981).

McFadden, Daniel, Antti P. Talvitie, and Associates: Demand Model Estimation and Validation, Vol. V of The Urban Travel Demand Forecasting Project Phase I Final Report Series, Institute of Transportation Studies, University of California, Berkeley, June 1977.

Rothenberg, Thomas J.: Efficient Estimation with A Priori Information (New Haven: Yale University Press, 1973).

Small, Kenneth A.: "Ordered Logit: A Discrete Choice Model with Proximate Covariance Among Alternatives," working paper, Princeton University, September 1981.

Small, Kenneth A.: "The Scheduling of Consumer Activities: Work Trips," American Economic Review, 72 (1982), forthcoming.

Train, Kenneth: "A Structured Logit Model of Auto Ownership and Mode Choice," Review of Economic Studies, 47 (1980), 357-370.