Specification and estimation of the nested logit model: alternative normalisations ☆

David A. Hensher [a], William H. Greene [b,*]

[a] *Faculty of Economics and Business, Institute of Transport Studies,
The University of Sydney, Sydney, NSW 2006, Australia*
[b] *Department of Economics, Stern School of Business, New York University,
44 West 4th Street, Mec 7-80, New York, NY 10012, USA*

Abstract

The nested logit model is currently the preferred extension to the simple multinomial logit (MNL) discrete choice model. The appeal of the nested logit model is its ability to accommodate differential degrees of interdependence (i.e., similarity) between subsets of alternatives in a choice set. The received literature displays a frequent lack of attention to the very precise form that a nested logit model must take to ensure that the resulting model is invariant to normalisation of scale and is consistent with utility maximisation. Some recent papers by F.S. Koppelman, C.H. Wen [Transp. Res. B 32 (5) (1998a) 289; Transp. Res. Record 1645 (1998b) 1] and G.L. Hunt [Nested logit models with partial degeneracy, Department of Economics, University of Maine, December 1998 (revised)] have addressed some aspects of this issue, but some important points remain somewhat ambiguous.

When utility function parameters have different implicit scales, imposing equality restrictions on common attributes associated with different alternatives (i.e., making them generic) can distort these differences in scale. Model scale parameters are then 'forced' to take up the real differences that should be handled via the utility function parameters. With many variations in model specification appearing in the literature, comparisons become difficult, if not impossible, without clear statements of the precise form of the nested logit model. There are a number of approaches to achieving this, with some or all of them available as options in commercially available software packages. This article seeks to clarify the issue, and to establish the points of similarity and dissimilarity of the different formulations that appear in the literature. © 2001 Elsevier Science Ltd. All rights reserved.

1. Introduction

The *nested logit* (NL) model is the preferred specification of a discrete choice model when analysts move beyond the multinomial logit (MNL) model. [See Ortuzar (2000) for an historical perspective on nested logit models.] Despite the increasing availability of other less restrictive models (in terms of the way that the random components of the utility expressions for each alternative are handled) such as *heteroscedastic extreme value*, *mixed logit*, *random parameter logit*, *covariance heterogeneity logit* and *multinomial probit* – see Louviere et al. (2000, Chapter 6, Appendix B) for a review – there remain reasons why the NL model will continue to be estimated. For example, the NL model is relatively easy to estimate, and, with its closed-form structure, it is easy to implement in the simulation of market shares before and after a policy change.

Specialists involved in the development of NL models, especially the active set of individuals researching estimation methods and developing software, have recently entered into a dialogue on the model specifications required in using software to ensure that the estimation is consistent with utility maximisation, and how one should handle degenerate branches (i.e., those with only a single alternative). Much of the discussion has taken place by email, however, the sentiment of the dialogue is partially represented in a series of recent papers by Koppelman and Wen (1998a, 1998b) and Hunt (1998). The objective of this article is to gather the presentation into a single transparent notation and to illustrate how one sets up an NL model to obtain outputs consistent with McFadden's NL model for utility maximisation, a derivative of his generalised extreme value (GEV) model (McFadden, 1981).

2. A common notation for NL models

We propose the following notation as a method of unifying the different forms of the NL model. [1] Each observed (or representative) component of the utility expression for an alternative (usually denoted as V_k for the kth alternative) is defined in terms of four parts – the parameters associated with the explanatory variables, β, an alternative-specific constant, α_k, a scale parameter, θ, and the explanatory variables, $\mathbf{x}$. The utility of alternative k for individual t is

$$U_{tk} = g_k(\alpha_k, \beta'\mathbf{x}_{tk}, \varepsilon_{tk}) = g_k(V_{tk}, \varepsilon_{tk}) = \alpha_k + \beta'\mathbf{x}_{tk} + \varepsilon_{tk}, \tag{1}$$

$$\mathrm{Var}[\varepsilon_{tk}] = \sigma^2 = \kappa/\theta^2. \tag{2}$$

The scale parameter, θ, is proportional to the inverse of the standard deviation of the random component in the utility expression, [2] σ, and is a critical input into the set up of the NL model (Ben-Akiva and Lerman, 1985; Louviere et al., 2000). Under the assumptions now well established in the literature, utility maximisation in the presence of random components which have

---

[1] A referee pointed out that the notation used is a standard that already exists. One may wish to disagree with this position (remembering a failed attempt 15 years ago to agree on common nomenclature in the transport research community). It is, however, still necessary to set out this standard herein.

[2] We have used θ here to avoid any confusion with its equivalent in various models below where we use λ, μ and γ.

independent (across choices and individuals) extreme value distributions produces a simple closed form for the probability that choice k is made, known as the MNL model;

$$\text{Prob}[U_{tk} > U_{tj} \quad \forall j \neq k] = \frac{\exp(\alpha_k + \beta' \mathbf{x}_{tk})}{\sum_{l=1}^{K} \exp(\alpha_l + \beta' \mathbf{x}_{tl})}. \tag{3}$$

Under these assumptions, the common variance of the assumed i.i.d. random components is lost. The same observed set of choices emerges regardless of the (common) scaling of the utilities. Hence, the latent variance is normalised at one, not as a restriction, but of necessity for identification.

One justification for moving from the MNL model to an NL model is to recognise (or at least test for) the possibility that the standard deviations (or variances) of the random error components in the utility expressions are different across groups of alternatives in the choice set. This arises because the sources of utility associated with the alternatives are not fully accommodated in V_k. The missing sources of utility may differentially impact on the random components across the alternatives, resulting in different variances. To accommodate the possibility of differential variances, we must explicitly introduce the scale parameters into each of the utility expressions. (If all scale parameters are equal, then the NL model 'collapses' back to a simple MNL model.) Hunt (1998) discusses the underlying conditions that produce the NL model as a result of utility maximisation within a partitioned choice set.

The notation for a three-level NL model covers the majority of applications. The literature suggests that very few analysts estimate models with more than three levels, and two levels are the most common. However, it will be shown below that a two-level model *may* require a third level (in which the lowest level is a set of dummy nodes and links) simply to ensure consistency with utility maximisation (which has nothing to do with a desire to test a three-level NL model). It is also common for a nested structure to have a branch with only one alternative. This is referred to as a *degenerate branch*. This requires careful definition in estimation. We will return to this point below.

It is useful to represent each level in an NL tree by a unique descriptor. For a three-level tree (Fig. 1), the top level will be represented by *limbs*, the middle level by a number of *branches* and the bottom level by a set of *elemental alternatives*, or *twigs*. We have $k = 1, \ldots, K$ elemental alternatives, $j = 1, \ldots, J$ branch composite alternatives and $i = 1, \ldots, I$ limb composite alternatives.

We use the notation $k|ji$ to denote alternative k in branch j of limb i and $j|i$ to denote branch j in limb i.

Define parameter vectors in the utility functions at each level as follows: β for elemental alternatives, γ for branch composite alternatives, and δ for limb composite alternatives. The branch-level composite alternative involves an aggregation of the lower-level alternatives. As discussed below, a branch specific scale parameter $\mu(j|i)$ will be associated with the lowest level of the tree. Each elemental alternative in the jth branch will actually have scale parameter $\mu'(k|ji)$. Since these will, of necessity, be equal for all alternatives in the same branch, the distinction by k is meaningless. As such, we collapse these into $\mu(j|i)$. The parameters $\lambda(j|i)$ will be associated with the branch level. The inclusive value (IV) parameters at the branch level will involve the ratios $\lambda(j|i)/\mu(j|i)$. The IV parameters associated with the IV variable in a branch, calculated from the natural logarithm of the sum of the exponentials of the V_k expressions at the elemental alternative level directly below a branch (Eq. (4)),
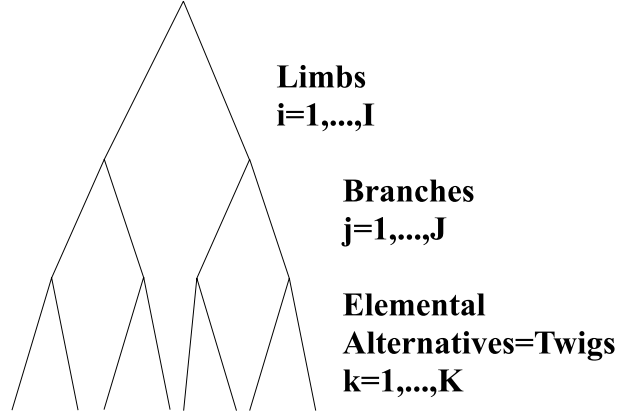
Fig. 1. Descriptors for a three-level NL tree.

$$\text{IV}(j|i) = \log \sum_{l=1}^{K|ji} \exp\left(\alpha_{l|ji} + \beta'\mathbf{x}(l|ji)\right) \tag{4}$$

have associated parameters defined as the $\lambda(j|i)/\mu(h|i)$, but, as noted, some normalisation is required. Normalisation is simply the process of setting one or more scale parameters equal to unity, while allowing the other scale parameters to be estimated. Some analysts do this without acknowledgment of which normalisation they have used, which makes the comparison of reported results between studies difficult. One approach restricts the numerator of $\lambda(j|i)/\mu(j|i)$ to be equal to one and the other so restricts the denominator.

The literature is vague on the implications of choosing the normalisation of $\lambda(j|i) = 1$ vs. $\mu(j|i) = 1$. It is important to note that the notation $\mu'(m|ji)$ used below refers to the scale parameter for each elemental alternative. However, since a NL structure is specified to test for the presence of identical scale within a subset of alternatives, it comes as no surprise that all alternatives partitioned under a common branch have the same scale parameter imposed on them. Thus, $\mu'(k|ji) = \mu(j|i)$ for every $k = 1, \ldots, K|ji$ alternatives in branch j in limb i.

We now set out the probability choice system (PCS) defined for later purposes as a three-level PCS (Eq. (5)),

$$P(k, j, i) = P(k|j, i) \times P(j|i) \times P(i). \tag{5}$$

In introducing alternative normalisations, we emphasise that there is one model normalised in different ways. When we normalise $\mu(j|i)$ to one, we refer to random utility model 1 (RU1), and when we normalise $\lambda(j|i)$ to one, we refer to Random Utility Model 2 (RU2). We ignore the subscripts for an individual.

2.1. Random Utility Model 1 (RU1)

The choice probabilities for the *elemental alternatives* are defined as

$$P(k|j, i) = \frac{\exp\left[\alpha_{k|ji} + \beta'\mathbf{x}(k|ji)\right]}{\sum_{l=1}^{K|j,i} \exp\left[\alpha_{l|ji} + \beta'\mathbf{x}(l|ji)\right]} = \frac{\exp\left[\alpha_{k|ji} + \beta'\mathbf{x}(k|ji)\right]}{\exp\left[\text{IV}(j|i)\right]}, \tag{6}$$

where $k|ji$ = elemental alternative k in branch j of limb i, $K|ji$ = number of elemental alternatives in branch j of limb i, and the inclusive value for branch j in limb i is

$$\text{IV}(j|i) = \log \sum_{l=1}^{K|ji} \exp[\alpha_{l|ji} + \beta'\mathbf{x}(l|ji)]. \tag{7}$$

The *branch-level* probability is

$$P(j|i) = \frac{\exp\{\lambda(j|i)[\gamma'\mathbf{y}(j|i) + \text{IV}(j|i)]\}}{\sum_{m=1}^{J|i} \exp\{\lambda(m|i)[\gamma'\mathbf{y}(m|i) + \text{IV}(m|i)]\}} = \frac{\exp\{\lambda(j|i)[\gamma'\mathbf{y}(j|i) + \text{IV}(j|i)]\}}{\exp[\text{IV}(i)]}, \tag{8}$$

where $j|i =$ branch j in limb i, $J|i$ = number of branches in limb i, and

$$\text{IV}(i) = \log \sum_{j=1}^{J|i} \exp\{\lambda(j|i)[\gamma'\mathbf{y}(j|i) + \text{IV}(j|i)]\}. \tag{9}$$

Finally, the *limb level* is defined by

$$P(i) = \frac{\exp\{\rho(i)[\delta'\mathbf{z}(i) + \text{IV}(i)]\}}{\sum_{n=1}^{I} \exp\{\rho(n)[\delta'\mathbf{z}(n) + \text{IV}(n)]\}} = \frac{\exp\{\rho(i)[\delta'\mathbf{z}(i) + \text{IV}(i)]\}}{\exp(\text{IV})}, \tag{10}$$

where I = number of limbs in the three-level tree and

$$\text{IV} = \log \sum_{i=1}^{I} \exp\{\rho(i)[\delta'\mathbf{z}(i) + \text{IV}(i)]\}. \tag{11}$$

RU1 has been described e.g., by Koppelman and Wen (1998a) and Bates (1999) as corresponding to a non-normalised nested logit (NNNL) specification, since the parameters are scaled at the lowest level, i.e., for $\mu'(k|j, i) = \mu(j|i) = 1$. Thus, note in this NNNL context, that there is no explicit scaling in (6) and (7) at the lowest level.

2.2. Random utility model 2 (RU2)

Suppose, instead, we normalise the upper-level parameters and allow the lower-level scale parameters to be free. The *elemental alternatives-level* probabilities will be

$$\begin{aligned} P(k|ji) &= \frac{\exp[\mu'(k|ji)[\alpha_{k|ji} + \beta'\mathbf{x}(k|ji)]]}{\sum_{l=1}^{K|j,i} \exp[\mu'(l|ji)[\alpha_{l|ji} + \beta'\mathbf{x}(l|ji)]]} = \frac{\exp[\mu'(k|ji)[\alpha_{k|ji} + \beta'\mathbf{x}(k|ji)]]}{\exp[\text{IV}(j|i)]} \\ &= \frac{\exp\left[\mu(j|i)[\alpha_{k|ji} + \beta'\mathbf{x}(k|ji)]\right]}{\sum_{l=1}^{K|ji} \exp\left[\mu(j|i)[\alpha_{l|ji} + \beta'\mathbf{x}(l|j, i)]\right]} = \frac{\exp[\mu(j|i)[\alpha_{k|ji} + \beta'x(k|ji)]]}{\exp[\text{IV}(j|i)]} \end{aligned} \tag{12}$$

(with the latter equality resulting from the identification restriction $\mu'(k|ji) = \mu'(m|ji) = \mu(j|i)$) and

$$\text{IV}(j|i) = \log \sum_{l=1}^{K|ji} \exp\{\mu(j|i)[\alpha_{l|ji} + \beta'\mathbf{x}(l|ji)]\}. \tag{13}$$

The *branch level* is defined by:

$$p(j|i) = \frac{\exp\left\{\rho(i)\left[\gamma'\mathbf{y}(j|i) + \frac{1}{\mu(j|i)}\mathrm{IV}(j|i)\right]\right\}}{\sum_{m=1}^{J|i}\exp\left\{\rho(i)\left[\gamma'\mathbf{y}(m|i) + \frac{1}{\mu(m|i)}\mathrm{IV}(m|i)\right]\right\}}$$
$$= \frac{\exp\left\{\rho(i)\left[\gamma'\mathbf{y}(j|i) + \frac{1}{\mu(j|i)}\mathrm{IV}(j|i)\right]\right\}}{\exp[\mathrm{IV}(i)]} \tag{14}$$

and

$$\mathrm{IV}(i) = \log\sum_{m=1}^{J|i}\exp\left\{\rho(i)\left[\gamma'\mathbf{y}(m|i) + \frac{1}{\mu(m|i)}\mathrm{IV}(m|i)\right]\right\}. \tag{15}$$

The *limb level* is defined by

$$P(i) = \frac{\exp\left[\delta'\mathbf{z}(i) + \frac{1}{\rho(i)}\mathrm{IV}(i)\right]}{\sum_{n=1}^{I}\exp\left[\delta'\mathbf{z}(n) + \frac{1}{\rho(n)}\mathrm{IV}(n)\right]} = \frac{\exp\left[\delta'\mathbf{z}(i) + \frac{1}{\rho(i)}\mathrm{IV}(i)\right]}{\exp(\mathrm{IV})}, \tag{16}$$

$$\mathrm{IV} = \log\sum_{i=1}^{I}\exp\left[\delta'\mathbf{z}(i) + \frac{1}{\rho(i)}\mathrm{IV}(i)\right]. \tag{17}$$

It is typically assumed that it is arbitrary as to which scale parameter is normalised. (See Hunt, 1998 for a useful discussion.) Most applications normalise the scale parameters associated with the branch-level utility expressions (i.e., $\lambda(j|i)$) at 1 as in RU2 above, then allow the scale parameters associated with the elemental alternatives $[\mu(j|i)]$ and hence the IV parameters in the branch composite alternatives to be unrestricted. It is implicitly assumed that the empirical results are identical to those that would be obtained if RU1 were instead the specification (even though parameter estimates are numerically different). *But, within the context of a two-level partition of a nest estimated as a two-level model, unless all attribute parameters are alternative-specific, this assumption is only true if the non-normalised scale parameters are constrained to be the same across nodes within the same level of a tree (i.e., at the branch level for two levels, and at the branch level and the limb level for three levels).* This latter result actually appears explicitly in some early studies of this model, e.g. Maddala (1983, p. 70) and Quigley (1985), but is frequently ignored in more recent applications. Note that in the common case of estimation of RU2 with two levels (which eliminates $\gamma(i)$) the 'free' IV parameter estimated will typically be $1/\mu(j|i)$. Other interpretations of this result are discussed in Hunt (1998).

2.3. Conditions to ensure consistency with utility maximisation

The previous section set out a uniform notation for a three-level NL model, choosing a different level in the tree for normalisation (i.e., setting scale parameters to an arbitrary value, typically unity). We have chosen levels one and two, respectively, for the RU1 and RU2 models. We are now ready to present a range of alternative empirical specifications for the NL model, some of which satisfy utility maximisation either directly from estimation or by some simple transfor-

mation of the estimated parameters. Compliance with utility maximisation requires that any monotonically increasing transformation of the utility functions of all elemental alternatives leave unaffected the ranking of the choice probabilities of the alternatives (McFadden, 1981). We limit the discussion to a two-level NL model and initially assume that all branches have at least two elemental alternatives. The important case of a degenerate branch (i.e., only one elemental alternative) is treated separately later.

3. An empirical illustration

To investigate the implications of alternative model specifications, we have estimated nine two-level models, using data collected in 1986 on non-business interurban trips between Sydney, Canberra and Melbourne. A total of 210 travellers chose a mode of transport from four alternatives – plane, car, train and bus. Details of the data are provided in Econometric Software (1998) and Louviere et al. (2000). The utility functions for the four alternatives are specified as follows:

$$U_{\text{Train}} = \alpha_{\text{Train}} + \beta_{\text{GC}}\text{GC}_{\text{Train}} + \beta_H\text{Hinc} + \beta_T\text{Ttime}_{\text{Train}} + \varepsilon_{\text{Train}},$$
$$U_{\text{Bus}} = \alpha_{\text{Bus}} + \beta_{\text{GC}}\text{GC}_{\text{Bus}} + \beta_H\text{Hinc} + \beta_T\text{Ttime}_{\text{Bus}} + \varepsilon_{\text{Bus}},$$
$$U_{\text{Plane}} = \alpha_{\text{Plane}} + \beta_{\text{GC}}\text{GC}_{\text{Plane}} + \beta_H\text{Hinc} + \beta_T\text{Ttime}_{\text{Plane}} + \varepsilon_{\text{Plane}},$$
$$U_{\text{Car}} = \beta_{\text{GC}}\text{GC}_{\text{Car}} + \varepsilon_{\text{Car}}.$$

The variables in the utility functions in addition to the alternative-specific constants are GC = Generalised cost (in dollars) out-of-pocket fuel cost for car or fare for plane, train and bus + time cost (the latter defined for main mode plus access and egress times excluding transfer time); Hinc = household income per annum (in $000's); Ttime = Transfer time (in minutes) the time spent waiting for and transferring to plane, train, bus.

Table 1 presents full information maximum likelihood (FIML) estimates of a two-level NNNL model. The tree structure for Table 1 has two branches, PUBLIC = (Train, Bus) and OTHER = (Car, Plane). In the PCS for this model, household income enters the probability of the branch choice directly in the utility for OTHER. Inclusive values from the lowest level enter both utility functions at the branch level. Table 2 presents FIML estimates of a two-level partially degenerate NL model. The tree structure for the models in Table 2, save for Model 7 which has an artificial third level, is FLY (Plane) and GROUND (Train, Bus, Car).

Estimates for both the NNNL model and the utility maximising (GEV-NL) parameterisations are presented. In the case of the GEV model parameterisation, estimates under each of the two normalisations (RU1 : $\mu = 1$ and RU2 : $\lambda = 1$) are provided as are estimates with the IV parameters restricted to equality within a level of the tree and unrestricted.

Eight models are summarised in Table 1 and six models in Table 2. Since there is only one limb, we drop the limb indicator from $\lambda(j|i)$ and denote it simply as $\lambda(j)$.

Model 1: RU1 with scale parameters equal within a level $[\lambda(1) = \lambda(2)]$.

Model 2: RU1 with scale parameters unrestricted within a level $[\lambda(1) \neq \lambda(2)]$.

Model 3: RU2 with scale parameters equal within a level (not applicable for a degenerate branch) $[\mu(1) = \mu(2)]$.

Table 1
Summary of alternative model specifications for a non-degenerate NL model[a]

Variables	Alternative	Model 1: RU1	Model 2: RU1	Model 3: RU2	Model 4: RU2	Model 5: NNNL[c]	Model 6: NNNL[c]	Model 8: RU1[b]	Model 9: RU2
Train constant	Train	4.542	3.757	5.873	6.159	3.6842	3.757	17.396 (2.2)	2.577
		(6.64)	(5.8)	(5.8)	(5.7)	(2.34)	(5.8)		(3.8)
Bus constant	Bus	3.924	2.977	5.075	5.380	3.218	2.977	19.523 (2.4)	2.892
		(5.83)	(4.4)	(6.0)	(5.8)	(2.18)	(4.4)		(3.73)
Plane constant	Plane	5.0307	4.980	6.507	6.154	3.681	4.980	4.165	4.165
		(6.81)	(6.7)	(5.7)	(5.2)	(2.34)	(6.7)	(3.3)	(3.5)
Generalised cost ($)	All	−0.01088	−0.0148	−0.01407	−0.01955	−0.01169	−0.0148		
		(−2.6)	(−3.5)	(−2.6)	(−3.2)	(−1.8)	(−3.5)		
Transfer time (mins.)	All excl car	−0.0859	−0.0861	−0.1111	−0.1064	−0.0637	−0.0861		
		(−7.4)	(−7.3)	(−5.5)	(−5.2)	(−2.5)	(−7.3)		
Hhld income ($000s)	other	0.03456	0.0172	0.0447	0.0426	0.0426	0.0416	0.04269	0.04269
		(3.2)	(2.9)	(4.0)	(3.8)	(3.8)	(3.4)	(4.24)	(3.8)
Generalised cost ($)	Air							0.00492	0.00492
								(0.56)	(0.60)
Generalised cost ($)	Train							−0.0943	−0.0139
								(−3.0)	(−2.8)
Generalised cost ($)	Bus							−0.1065	−0.0158
								(−2.9)	(−2.9)
Generalised cost ($)	Car							−0.0143	−0.0143
								(−2.5)	(−3.0)

Transfer time (min)	Air							−0.1048 (−6.2)	−0.1048 (−7.1)
Transfer time (min)	Train							−0.0787 (−2.5)	−0.0116 (−1.9)
Transfer time (min)	Bus							−0.1531 (−3.5)	−0.0227 (−1.9)
Inclusive value Parameters:									
IV other	Other	1.293 (5.3)	2.42 (4.6)	0.773 (3.8)	0.579 (3.3)	0.969 (3.2)c	2.42 (4.6)	1.00 (fixed)	1.00 (fixed)
IV public transport d	Public transport	1.293 (5.3)	1.28 (5.1)	0.773 (3.8)	1.03 (3.2)	1.724 (3.3)c	1.28 (5.1)	0.148 (2.0)	6.75 (1.9)
Log-likelihood		−190.178	−184.31	−190.178	−188.43	−188.43	−184.31	−177.82	−177.82
Direct elasticities:									
	Plane	−0.544	−0.797	−0.544	−0.666	−0.666	−0.797	0.228	0.228
	Car	−0.651	−1.081	−0.651	−0.762	−0.762	−1.081	−0.799	−0.799
	Train	−0.629	−0.854	−0.629	−0.910	−0.910	−0.854	−0.188	−1.27
	Bus	−0.759	−1.014	−0.759	−1.174	−1.174	−1.014	−0.343	−2.32

[a] Tree structure: other {plane, car} vs. public transport {train, bus} except for Model 5 which is OTHER {planem (plane), carm(car)} vs. public transport {trainm (train), busm (bus)}. Note: there is no Model 7 in order to keep equivalent model numbering in Tables 1 and 2.

[b] Model 8 with all alternative-specific attributes produces the exact parameter estimates, overall goodness-of-fit and elasticities as the NNNL model (and hence it is not reported).

[c] Standard errors are uncorrected.

[d] = IV parameters in Model 5 based on imposing equality of IV for (other, trainm, busm) and for (public transport, planem,carm).

Table 2
Summary of alternative model specifications for a partial degeneracy NL model[a]

Variables	Alternatives	Model 1: RU1	Model 2: RU1	Model 4: RU2	Model 5: NNNL[b]	Model 6: NNNL[b]	Model 7: RU2
Train constant	Train	5.070 (7.6)	5.065 (7.7)	2.622 (5.9)	9.805 (3.4)	5.065 (7.7)	4.584 (2.7)
Bus constant	Bus	4.145 (6.7)	4.096 (6.7)	2.143 (5.5)	8.015 (3.3)	4.096 (6.7)	3.849 (2.7)
Plane constant	Plane	5.167 (4.3)	6.042 (5.04)	2.672 (3.0)	9.993 (3.8)	6.042 (5.04)	5.196 (2.7)
Generalised cost ($)	All	−0.0291 (−3.6)	−0.0316 (−3.9)	−0.0151 (4.32)	−0.0564 (2.8)	−0.0316 (−3.9)	−0.0187 (−2.4)
Transfer time (min)	All excl car	−0.1156 (8.15)	−0.1127 (−7.9)	−0.0598 (−5.9)	−0.2236 (3.7)	−0.1127 (−7.9)	−0.0867 (−2.7)
Hhld income ($000s)	Fly	0.02837 (1.4)	0.0262 (1.5)	0.0143 (1.35)	0.01467 (1.4)	0.01533 (1.6)	0.02108 (1.4)
Party size	Auto						0.4330 (2.0)
Inclusive value Parameters:							
IV	Fly	0.517 (4.1)	0.586 (4.2)	$1.00(0.67 \times 10^{15})$[c]	$1.00(0.67 \times 10^{15})$[c]	0.586 (4.2)	Not applicable
IV	Ground	0.517 (4.1)	0.389 (3.1)	1.934 (5.0)[c]	0.517 (5.0)	0.389 (3.1)	Not applicable
IV	Auto						
IV	Public transport						1.26 (2.3)
IV	Other						0.844 (2.1)
IV	Land PT						Not applicable
Log-likelihood		−194.94	−193.66	−194.94	−194.94	−193.66	−194.27
Direct elasticities:							
	Plane	−0.864	−1.033	−0.864	−0.864	−1.033	−0.859
	Car	−1.332	−1.353	−1.332	−1.332	−1.353	−0.946
	Train	−1.317	−1.419	−1.317	−1.317	−1.419	−1.076
	Bus	−1.650	−1.878	−1.650	−1.650	−1.878	−1.378

[a] Tree structure: fly {plane} vs. ground {train, bus, car}. *Note:* Model 3 is not defined for a degenerate branch model when the IV parameters are forced to equality. Forcing a constraint on Model 4 (i.e., equal IV parameters) to obtain Model 3 produced exactly the same results for all the parameters. This is exactly what should happen. Since the IV parameter is not identified, no linear constraint that is imposed that involves this parameter is binding. Model 7 tree is other{fly (plane) vs. auto (car)} vs. Land PT{ public transport (train, bus}).
[b] Standard errors are uncorrected.
[c] = IV parameters in Model 5 based on imposing equality of IV for (ground, airm). It is not applicable for (fly, trainm, busm, carm) since fly is degenerate.

Model 4: RU2 with scale parameters unrestricted within a level $[\mu(1) \neq \mu(2)]$.

Model 5: NNNL model with dummy nodes and links to allow unrestricted scale parameters in the presence of generic attributes to recover parameter estimates that are consistent with utility maximisation. This is equivalent up to scale with RU2 (model 4).

Model 6: NNNL model with no dummy nodes/links and different scale parameters within a level. This is a typical NL model implemented by many practitioners (and is equivalent to RU1 (Model 2)).

Model 7: RU2 with unrestricted scale parameters and dummy nodes and links to comply with utility maximisation (for partial degeneracy). Since Model 7 is identical to Model 8 in Table 1, it is presented only in Table 2.

Models 8 and 9: For the non-degenerate NL model (Table 1), these are RU1 and RU2 in which all parameters are alternative-specific and scale parameters are unrestricted across branches.

All results reported in Tables 1 and 2 are obtained using LIMDEP Version 7 (Revised, December 1998) (Econometric Software, 1998). The IV parameters for RU1 and RU2 that LIMDEP reports are the μ's and the λ's that are shown in the equations above. These μ's and λ's are proportional to the reciprocal of the standard deviation of the random component. The t-values in parenthesis for the NNNL model require correction to compare with RU1 and RU2. Koppelman and Wen (1998b) provide the procedure to adjust the t-values. For a two-level model, the corrected variance and hence standard error of estimate for the NNNL model is

$$\text{Var}[\hat{\beta}_{\text{RU}}] = \beta_{\text{NN}}^2 \text{Var}[\hat{\mu}_{\text{NN}}] + \mu_{\text{NN}}^2 \text{Var}[\hat{\beta}_{\text{NN}}] + 2\mu_{\text{NN}}\beta_{\text{NN}}\text{Cov}[\hat{\mu}_{\text{NN}}, \hat{\beta}_{\text{NN}}]. \tag{18}$$

3.1. The case of generic attribute parameters

Beginning with the non-degenerate case, it can be seen in Table 1 that the GEV parameterisation estimates with IV parameters unrestricted (Models 2 and 4) are *not* invariant to the normalisation chosen. Not only is there no obvious relationship between the two sets of parameter estimates, the log-likelihood function values at convergence are not equal (−184.31 vs. −188.43) illustrating the fact that the normalisation has not been handled properly. When the GEV parameterisation is estimated subject to the restriction that the IV parameters be equal (Models 1 and 3), invariance is achieved across normalisation after accounting for the difference in scaling. The log-likelihood function values at convergence are equal (−190.178), and the IV parameter estimates are inverses of one another ($1/0.773 = 1.293$, within rounding error). Multiplying the utility function parameter estimates at the elemental alternatives level (i.e., $\alpha_{\text{Plane}}, \alpha_{\text{Train}}, \alpha_{\text{Bus}}$, GC, Ttime) by the corresponding IV parameter estimate in one normalisation (e.g., Model 1) yields the utility function parameter estimates in another normalisation (e.g., Model 3). For example, in Model 3, $(1/1.293) \times 5.873$ for Train constant $= 4.542$ in Model 1.

The points made above about invariance, or the lack of it, scaling, and the equivalence of GEV and NNNL under the appropriate set of parametric restrictions are also illustrated in Table 2 for the case of a partially degenerate NL model structure. However, an additional and important result emerges for the partial degeneracy case. If the IV parameters are unrestricted, the GEV model "estimate" of the parameter on the degenerate partition IV is unity under the $\lambda = 1$ normalisation. This will always be the case because of the cancellation of the IV parameter and the lower-level scaling parameter in the GEV model in the degenerate partition. The results will

be invariant to whatever value this parameter is set to. To see this, consider the results for the unrestricted GEV model presented as Model 4 in Table 2. The IV parameter is "estimated" to be 1.934, and if we were to report Model 3, all of the other estimates would be the same as in Model 4 and the log-likelihood function values at convergence are identical (-194.94). In a degenerate branch, whatever the value of $(1/\mu)$ is, it will cancel with the lower-level scaling parameter, μ, in the degenerate partition marginal probability. If we select $\mu = 1$ for normalisation (in contrast to λ) in the presence of a degenerate branch, the results will produce restricted (Model 1) or unrestricted (Model 2) estimates of λ which, unlike μ, do not cancel out in the degenerate branch. (Hunt (1998) pursues this issue at length.)

To illustrate the equivalence of behavioral outputs for RU1 and RU2, Tables 1 and 2 present the weighted aggregate direct elasticities for the relationship between the generalised cost of alternative kji and the probability of choosing alternative $k|ji$. As expected the results are identical for RU1 (Model 1) and RU2 (Model 3) when the IV parameters are equal across all branches at a level in the GEV model. The elasticities are significantly different from those obtained from Models 2 and 4, although Models 4 and 5 produce the same results (see below). Model 6 (equivalent to Model 2) is a common model specification in which parameters of attributes are generic and scale parameters are unrestricted within a level of the NL model with no constraints imposed to recover the utility maximisation estimates.

3.2. Allowing different scale parameters across nodes in a level in the presence of generic and/or alternative-specific attribute parameters between partitions

When we allow the IV parameters to be unrestricted in the RU1 and RU2 GEV models and in the NNNL model we fail to comply with normalisation invariance, and for Models 2 and 6 we also fail to produce consistency with utility maximisation. RU1 (Model 2) fails to comply with utility maximisation because of the absence of explicit scaling in the utility expressions for elemental alternatives. We obtain different results on overall goodness-of-fit and the range of behavioural outputs such as elasticities.

For a given nested structure and set of attributes there can only be one utility maximising solution. This presents a dilemma, since we often want the scale parameters to vary between branches and/or limbs or at least test for non-equivalence. This is after all, the main reason why we seek out alternative nested structures. Fortunately there is a solution, depending on whether one opts for a specification in which some or all of the parameters are generic or whether they are all alternative-specific. Models 5 to 9 are alternative specifications.

If all attributes between partitions are unrestricted (i.e., alternative-specific), unrestricted scale parameters are compliant with utility maximisation under all specifications (i.e., RU2 = 1, RU2 and NNNL). Intuitively, the fully alternative-specific specification avoids any artificial 'transfer' of information from the attribute parameters to the scale parameters that occurs when restrictions are imposed on parameter estimates. Models 8 and 9 in Table 1 are totally alternative-specific. The scale parameters for Models 8 and 9 are the inverse of each other. That is, for the unrestricted IV, 0.148 in Model 8 equals (1/6.75) in Model 9. The alternative-specific parameter estimates associated with attributes in the public transport branch for Model 8 can be recovered from Model 9 by a scale transformation. For example, 0.148×17.396 for the train constant equals 2.577 in Model 9. The estimated parameters are identical in Models 8 and 9 for the 'other'

modes since their IV parameter is restricted to equal unity in both models. This demonstrates the equivalence up to scale of RU1 and RU2 when all attribute parameters (including IV) are unrestricted.

When we impose the generic condition on an attribute associated with alternatives in different partitions of the nest, Koppelman and Wen (1998a,b) (and Daly in advice to ALOGIT (Hague Consulting Group, 1995) subscribers) have shown how one can recover compliance with utility maximisation in an NNNL model under the unrestricted scale condition (within a level of the NL model) by adding dummy nodes and links below the bottom level and imposing cross-branch equality constraints as illustrated in Fig. 2. Intuitively, what we are doing is allowing for differences in scale parameters at each branch but preserving the (constant) ratio of the IV parameters between two levels through the introduction of the scale parameters at the elemental level; the latter requires the additional lower level in an NNNL specification. The NNNL specification does not allow unrestricted values of scale at the elemental level, unlike RU2, for example. Preserving a constant ratio through cross-over equality constraints between levels in the nest satisfies the necessary condition of choice probability invariance to the addition of a constant in the utility expression of all elemental alternatives.

Adding an extra level is not designed to investigate the behavioural implications of a three-level model; rather it is a 'procedure' to reveal the scale parameters at upper levels where they have not been identified. This procedure is fairly straightforward for two branches (see Model 5 in Tables 1 and 2). With more than two branches, one has to specify additional levels for each branch. The number of levels grows quite dramatically. However, there is one way of simplifying this procedure, if we recognise that the ratio of the scale parameters between adjacent levels must be constant. Thus, for any number of branches, consistency with utility maximisation requires that the product of all the ratios of scale parameters between levels must be identical from the root to all elemental alternatives. To facilitate this, one can add a single link below each real alternative with the scale of that link set equal to the product of the scales of all scale parameters not included in the path to that alternative. For example, in the case of three branches with scales equal to
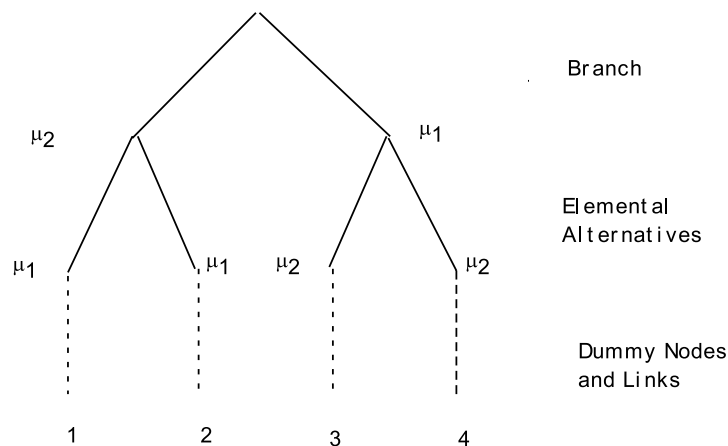

Fig. 2. Estimating a two-level model to allow for unrestricted scale parameters within a level.

λ_1, λ_2 and λ_3; the scale below the first branch would be $(\lambda_2 \times \lambda_3)$, below the second branch it would be $(\lambda_1 \times \lambda_3)$ and below the third branch it would be $(\lambda_1 \times \lambda_2)$.

Model 5 is estimated as an NNNL model with the addition of a lower level of nodes and links with cross-branch equality constraints on the scale parameters. For example, in Table 1, the tree structure is as follows: {Other [planem (plane), carm (car)], Public Transport [trainm (train), busm (bus)]}. The cross-over constraint for two branches sets the scale parameters to equality for {*Other*, trainm, busm} and {*Public Transport*, planem, carm}. The Model 5 (Table 1) produces results which are identical to RU2 (model 4) in respect of goodness-of-fit and elasticities; with all parameter estimates equivalent up to scale. Since we have two scale parameters in Model 5, the ratio of each branches IV parameters to their equivalent in Model 4 provides the adjustment factor to translate Model 5 parameters into Model 4 parameters (or vice versa). For example, the ratio of $0.579/0.969 = 1.03/1.724 = 0.597$. If we multiply the train-specific constant in Model 4 of 6.159 by 0.597, we obtain 3.6842, the train-specific constant in Model 5. This is an important finding, because it indicates that

> the application of the RU2 specification with unrestricted scale parameters in the presence of generic parameters across branches for the attributes is identical to the results obtained by estimating the NNNL model with an extra level of nodes and links.

RU2 thus avoids the need to introduce the extra level. [3] The equivalent findings are shown in Table 2 where the scale ratio is 3.74. Intuitively one might expect such a result given that RU2 allows the scale parameters to be freely estimated at the lower level (in contrast to RU1 where they are normalised to 1.0). One can implement this procedure under an exact RU2 model specification to facilitate situations where one wishes to allow scale parameters at a level in the nest to be different across branches in the presence or absence of a generic specification of attribute parameters. *The estimation results in Model 4 are exactly correct and require no further adjustments.* The procedure can also be implemented under an NNNL specification (with an extra level of nodes and links) (Model 5). The elasticities, marginal rates of substitution, goodness-of-fit are identical in Models 4 and 5. The parameter estimates are identical up to the ratio of scales.

4. Conclusions

This paper has addressed the issue of normalisation of NL models which is a common source of confusion for practitioners and researchers. The paper emphasises two critical points:
1. What may be perceived as different models (as in Koppelman and Wen 1998a,b) are instead, appropriately defined as different normalisation of the same general model. Thus, it is seen that our (6)–(11) and (12)–(17) are not different models at all, but merely two formulations of the one model built around (4) and the surrounding discussion.
2. The impact of normalisation on the scales of some parameters may produce internal inconsistency of the model if not handled properly. Typically, if the same parameter appears in several

---

[3] From a practical perspective, this enables programs like NLOGIT that limit the number of levels which can be jointly estimated to use all levels for real behavioural analysis.

nests, normalisation from the bottom (RU1) may cause problems, as the parameter will be scaled differently in each nest.

The empirical applications and discussion herein have identified the model specification required to ensure compliance with the necessary conditions for utility maximisation.

This can be achieved for a *GEV-NL model* by either

- setting the IV parameters to be the same at a level in the nest in the presence of generic parameters, or
- implementing the RU2 specification and allowing the IV parameters to be free in the presence of generic attribute parameters between partitions of a nest, or
- setting all attribute parameters to be alternative-specific between partitions, allowing IV parameters to be unrestricted.

This can be achieved for a NNNL model by either

- setting the scale parameters to be the same at a level in the nest (for the non-normalised scale parameters) and rescaling all estimated parameters associated with elemental alternatives by the estimated IV parameter, or
- allowing the IV parameters to be free, and adding an additional level at the bottom of the tree through dummy nodes and links, and constraining the scale parameters at the elemental alternatives level to equal those of the dummy nodes of all other branches in the total NL model, or
- setting all attribute parameters to be alternative-specific between partitions, allowing IV parameters to be unrestricted.

The statement is made in Koppelman and Wen (1998a,b) (attributed to Daly (1987)), that the non-normalised form is not consistent with RUM. In view of the preceding, we see that this is not necessarily correct; at best, the statement is imprecise. With the proper normalisation of the model, we see that the NNNL model is, indeed, consistent with RUM.

Appendix A. Random utility model 3 (RU3)

John Bates, in correspondence to the authors questions the usefulness of RU2 (our preferred specification) as an appropriate model of utility maximisation (Bates 1999). He states:

> I am altogether less clear as to the purpose of RU2. The presence of $\mu \neq 1$ at the bottom of the structure signifies that the coefficients are not being scaled at the lowest level. In moving up one level, the logsums are deflated by μ and then a new IV parameter γ is applied. At the top level, the scaling factor is set to 1 and the logsum is deflated by γ. While this appears to scale the parameters at the top of the structure, it does not seem to me to correspond directly with UMNL (utility maximising nested logit). The reason is that the "final" alternatives are still constrained to be located at the bottom of the structure.

Bates' point is well taken. He proposes an alternative specification of the model which he argues is more appropriate. We lay out his model and show that what he has proposed is identical to RU2 with a minor transformation of the parameters. This explains the finding in his example, that though the parameter estimates he presents for a model differ from LIMDEP's RU2 counterparts, the log-likelihood functions are the same.

Bates' proposed alternative specification of the model is as follows:

$$P(k|j,i) = \frac{\exp\left[\frac{1}{\lambda(j|i)\omega(i)}\left[\alpha_{k|ji} + \beta'\mathbf{x}(k|ji)\right]\right]}{\sum_{m=1}^{K|ji} \exp\left[\frac{1}{\lambda(j|i)\omega(i)}\left[\alpha_{m|ji} + \beta'\mathbf{x}(m|ji)\right]\right]},$$

$$\mathrm{IV}(j|i) = \lambda(j|i)\omega(i)\log\sum_{m=1}^{K|ji}\exp\left[\frac{1}{\lambda(j|i)\omega(i)}\left[\alpha_{m|j,i} + \beta'\mathbf{x}(m|ji)\right]\right],$$

$$P(j|i) = \frac{\exp\left[\frac{1}{\omega(i)}\left[\gamma'\mathbf{y}(j|i) + \mathrm{IV}(j|i)\right]\right]}{\sum_{m=1}^{J|i}\exp\left[\frac{1}{\omega(i)}\left[\gamma'\mathbf{y}(m|i) + \mathrm{IV}(m|i)\right]\right]},$$

$$P(i) = \frac{\exp\left[\delta'\mathbf{z}(i) + \mathrm{IV}(i)\right]}{\sum_{n=1}^{J}\exp\left[\delta'\mathbf{z}(n) + \mathrm{IV}(n)\right]},$$

$$\mathrm{IV}(i) = \rho(i)\log\sum_{m=1}^{J|i}\exp\left[\frac{1}{\omega(i)}\left[\delta'\mathbf{y}(m|i) + \mathrm{IV}(m|i)\right]\right].$$

This appears to be a different formulation, but it is not. To see this, note first that since $\lambda(j|i)$ is a free parameter in the model that appears only in the scaling parameter for the elemental utility functions, we may write the model in terms of the compound parameter $\phi(j|i) = \lambda(j|i)\omega(i)$ and, moreover, since MLEs are invariant to transformation, write this as simply $\mu(j|i) = 1/\phi(j|i) = 1/[\lambda(j|i)\omega(i)]$. Inserting this into Bates' $P(k|j,i)$ produces the counterpart for RU2, so the probabilities are identical and, moreover, the slope parameters, β, are the same in the two models. The inclusive values appear to be different, but this is misleading. Now, let $\rho(i) = 1/\omega(i)$ and make the substitutions for $\omega(i)$ and the necessary changes in $\mathrm{IV}(j|i)$ in $P(j|i)$. What emerges, once again, is RU2, so $P(j|i)$ is also identical. No new parameters are introduced by $P(i)$, so the direct substitution of $\rho(i) = 1/\omega(i)$ produces the now expected result.

References

Bates, J., 1999. More thoughts on nested logit, John Bates Services, Oxford (mimeo), January.

Ben-Akiva, M., Lerman, S.R., 1985. Discrete Choice Analysis: Theory and Application to Travel Demand. MIT Press, Cambridge.

Daly, A., 1987. Estimating 'tree' logit models. Transportation Research B 21 (4), 251–267.

Econometric Software, 1998. Limdep Version 7 for Windows, Econometric Software, New York and Sydney, December revision.

Hague Consulting Group, 1995. ALOGIT: User's Guide, Hague Consulting Group, Den Haag.

Hunt, G.L., 1998. Nested logit models with partial degeneracy. Department of Economics, University of Maine, December (revised).

Koppelman, F.S., Wen, C.H., 1998a. Alternative nested logit models: structure, properties and estimation. Transportation Research B 32 (5), 289–298.

Koppelman, F.S., Wen, C.H., 1998b. Nested logit models: which are you using? Transportation Research Record 1645, 1–7.

Louviere, J.J., Hensher, D.A., Swait, J., 2000. Stated Choice Methods: Analysis and Applications in Marketing, Transportation and Environmental Valuation. Cambridge University Press, Cambridge, in press.

Maddala, J.S., 1983. Limited Dependent and Qualitative Variables in Econometrics. Cambridge University Press, New York.

McFadden, D.L., 1981. Econometric models of probabilistic choice. In: Manski, C.F., McFadden, D.L. (Eds.), Structural Analysis of Discrete Data, MIT Press, Cambridge, MA, pp. 198–271.

Ortuzar, Juan de Dios, 2000. A short note on the history of nested logit models. Transportation Research, in press.

Quigley, J., 1985. Consumer Choice of Dwelling, Neighborhood, and Public Services. Regional Science and Urban Economics 15, 41–63.