A note on the nested Logit model

In this note we present the basic idea of the nested multiple Logit model and how it relates to and generalizes the standard multiple Logit model. At the start, we restate some important implications of the latter model.

Multinomial Logit: Conditional probabilities

Essentially, the standard multinomial Logit model can be viewed as a modelling of relative response probabilities by functions of the form

\[ \frac{P_{ji}}{P_{ji} + P_{ki}} = \frac{G(x_i \beta_j)}{G(x_i \beta_j) + G(x_i \beta_k)} \]

\[ \iff \]

\[ \frac{P_{ji}}{P_{ki}} = \frac{G(x_i \beta_j)}{G(x_i \beta_k)}, \quad j, k = 1, \ldots, J, k \neq j, \]

where \( G(\cdot) \) is a suitable positive function with one argument. Combining this with the condition \( \sum_{j=1}^{J} P_{ji} = 1 \) it follows that

\[ P_{1i} = F_1(x_i, \beta) = \frac{G(x_i \beta_1)}{\sum_{k=1}^{J} G(x_i \beta_k)}, \]

\[ \vdots \]

\[ P_{ji} = F_j(x_i, \beta) = \frac{G(x_i \beta_j)}{\sum_{k=1}^{J} G(x_i \beta_k)}. \]

[1]: In the Logit model we let the common function \( G(\cdot) \) be a simple exponential function: \( G(x_i \beta_j) = e^{x_i \beta_j} \) (\( j = 1, \ldots, J \)).

[2]: A property of this function, on which much of the algebra for Logit models relies, is: \( G(a_1)G(a_2) \ldots G(a_P) \equiv G(a_1 + a_2 + \cdots + a_P) \).

We further let \( \beta_1 = 0 \iff G(x_i \beta_1) = G(0) = 1 \forall i \). This indicates that alternative 1 is a kind of base alternative. We then obtain the system of Logit probabilities

\[ P_{1i} = P(y_{1i} = 1) = \frac{1}{1 + \sum_{k=2}^{J} e^{x_i \beta_k}}, \quad j = 2, \ldots, J \]

\[ P_{ji} = P(y_{ji} = 1) = \frac{e^{x_i \beta_j}}{1 + \sum_{k=2}^{J} e^{x_i \beta_k}}, \quad i = 1, \ldots, n. \]
An important property of the standard Logit model is that all conditional probabilities also have Logit-form i.e., the same form as (2). The following examples illustrate this: The probability of choosing (i) alternative \( j \) given that either \( j \) or \( k \) is chosen and (ii) alternative \( j \), given that one of \((j, k, l)\) is chosen, are, respectively,

\[
\begin{align*}
(i) \quad P_{j|(jk)} &= P(y_{ji} = 1 | y_{ji} = 1 \lor y_{ki} = 1) = \frac{P_{ji}}{P_{ji} + P_{ki}} \\
&= \frac{e^{x_i \beta_j}}{e^{x_i \beta_j} + e^{x_i \beta_k}} = \frac{e^{x_i (\beta_j - \beta_k)}}{1 + e^{x_i (\beta_j - \beta_k)}}; \\
(ii) \quad P_{j|(jkl)} &= P(y_{ji} = 1 | y_{ji} = 1 \lor y_{ki} = 1 \lor y_{li} = 1) = \frac{P_{ji}}{P_{ji} + P_{kl} + P_{li}} \\
&= \frac{e^{x_i \beta_j}}{e^{x_i \beta_j} + e^{x_i \beta_k} + e^{x_i \beta_l}} = \frac{e^{x_i (\beta_j - \beta_l)}}{1 + e^{x_i (\beta_j - \beta_k)} + e^{x_i (\beta_k - \beta_l)}}; \\
\end{align*}
\]

letting \( \lor \) symbolize ‘or’. We see that the level of the coefficients – \((\beta_j, \beta_k)\) in example (i), \((\beta_j, \beta_k, \beta_l)\) in example (ii) – will be indeterminate even if the conditional response probabilities are known. These conditional probabilities are uniquely characterized by coefficient differences: \((\beta_k - \beta_j)\) in example (i), \((\beta_k - \beta_j, \beta_l - \beta_j)\) in example (ii).

Hence, it follows from (2) and examples (i) and (ii) that

\[
\ln \left[ \frac{P_{ji}}{P_{ki}} \right] = \ln \left[ \frac{P_{j|(jk)}}{P_{k|(jk)}} \right] = \ln \left[ \frac{P_{j|(jkl)}}{P_{k|(jkl)}} \right] = x_i (\beta_j - \beta_k), \quad j, k = 2, \ldots, J, \quad j \neq k \neq l,
\]

Such ratios are, as in the binomial model, denoted as log-odds ratios. The remarkable thing is that the three log-odds-ratios – one marginal and two conditional – are equal and described by only the variable vector and the coefficient vector for alternatives \( j \) and \( k \). Thus it is immaterial whether the choice set we condition on, includes all \( J \) alternatives, only \((j, k, l)\) or only \((j, k)\).

**The General Structure of the Probabilities**

The standard multinomial Logit model has the limitation that all alternatives are considered as ‘being on the same level’, co-ordinated. In the model class now to be considered, we assume that there is two levels in the hierarchy, \( M \) groups (often denoted as nests), indexed by \( m \) \((m = 1, \ldots, M)\), at the upper level, and \( K_m \) in group \( m \). Let alternative \( k = 1, \ldots, K_m \) in group \( m \) be indexed by \((mk)\). Examples will be given in the lectures. Two types of variable vectors occur, one being specific for the group and common to all alternatives in the group – for group \( m \) denoted as \( z_m \) – and one being specific for the alternative within a group – for alternative \((mk)\) denoted as \( x_{mk} \). Further let \( x_m \) be the variable vector which contains all alternative specific variables in group \( m \), that
is $x_m = (x_{m1}, x_{m2}, \ldots, x_{mK_m})$. We assume that no coefficients are alternative specific, some are group specific and some are common to all groups: For group $m$ the vector $\beta_m$ is attached to the alternative specific vector $x_{mk}$, while the vector $\alpha$ is attached to the group specific vectors $z_m$. Let in the following $\sum_{k^*} \sum_{j^*}$ be shorthand notation for $\sum_{k=1}^{K_j} \sum_{l=1}^{K_j}$ and $\sum_{m^*} \sum_{j^*} \sum_{r^*}$ be shorthand notation for $\sum_{m=1}^{M} \sum_{j=1}^{M} \sum_{r=1}^{M}$.

Within each group (nest) it is assumed that the conditional probability to choose $(mk)$ given that $m$ is chosen, is of the form (1):

$$p_{mk|m} = \frac{G(x_{mk}\beta_m^*)}{\sum_l G(x_{ml}\beta_m^*)}, \quad \sum_k p_{mk|m} = 1 \quad \forall m, m = 1, \ldots, M; \quad k = 1, \ldots, K_m,$$

where $G(\cdot)$ is a so far unspecified function, which in nested Logit models, like in the standard ones, is parameterized as an exponential function. A simple specification of the probability of choosing an alternative in group $m$—regardless of which—might be that it has the standard Logit form:

$$p_m = \frac{G(z_m\alpha)}{\sum_j G(z_j\alpha)}, \quad m = 1, \ldots, M.$$

Then, however, no variables in $x_m$ would have affected whether group $m$ is chosen or not: The probability that alternative $(mk)$ is chosen would be the product of two multinomial choice probabilities:

$$p_{mk} = p_m p_{k|m} = \frac{G(z_m\alpha)}{\sum_j G(z_j\alpha)} \cdot \frac{G(x_{mk}\beta_m^*)}{\sum_l G(x_{ml}\beta_m^*)}, \quad m = 1, \ldots, M; \quad k = 1, \ldots, K_m.$$

Why would this be inconvenient when the alternatives form a hierarchic structure?

**Nested Logit as a special case**

In the nested multinomial Logit model a more general parametrization of the choice probabilities is chosen: For group $m$ the group specific choice probability is assumed to be of the form

$$p_m = \frac{G(z_m\alpha)H(x_m, \beta_m^*, \rho_m)}{\sum_r G(z_r\alpha)H(x_r, \beta_r^*, \rho_r)}, \quad m = 1, \ldots, M,$$

where $H(x_m, \beta_m^*, \rho_m)$ is a so far unspecified positive function which in addition to $x_m$ and $\beta_m^*$ also contains the scalar parameter $\rho_m$. The latter serves to weight the different $G(z_m\alpha)_s$. From (4) and (5) it follows that

$$p_{mk} = p_m p_{k|m} = \frac{G(z_m\alpha)H(x_m, \beta_m^*, \rho_m)G(x_{mk}\beta_m^*)}{\sum_r G(z_r\alpha)H(x_r, \beta_r^*, \rho_r)[\sum_l G(x_{ml}\beta_m^*)]}, \quad m = 1, \ldots, M; \quad k = 1, \ldots, K_m.$$
Will now the very restrictive IIA-axiom hold? [See Dagsvik, Section 3.2 (in syllabus), for a precise definition of IIA and discussion of its implications.] First, from (5) it follows that the odds-ratio for groups $m$ and $j$ is

\[
p_m \frac{p_j}{p_j} = \frac{G(z_m \alpha)H(x_m, \beta_m^*, \rho_m)}{G(z_j \alpha)H(x_j, \beta_j^*, \rho_j)}, \quad m, j = 1, \ldots, M.
\]

Since no attributes for groups outside $m$ and $j$ enter this expression, we can conclude that IIA still holds between all pairs of groups. Second, from (6) it follows that the odds-ratio for alternatives $(mk)$ and $(jl)$ is

\[
\frac{p_{mk}}{p_{jl}} = \frac{G(z_m \alpha)H(x_m, \beta_m^*, \rho_m)G(x_{mk} \beta_m^*) \sum_r G(x_{jr} \beta_j^*)}{G(z_j \alpha)H(x_j, \beta_j^*, \rho_j)G(x_{jl} \beta_j^*) \sum_r G(x_{mr} \beta_m^*)},
\]

\[m, j = 1, \ldots, M; \quad k = 1, \ldots, K_m; \quad l = 1, \ldots, K_j.\]

Because attributes for alternatives outside $(mk)$ and $(jl)$ enter this expression, we know that IIA does not in general hold between all pairs of alternatives.

However, in two special cases will IIA be satisfied:

**CASE 1:** The alternatives belong to the same group (nest) ($m = j$). Then (8) is simplified to

\[
\frac{p_{mk}}{p_{ml}} = \frac{p_{klm}}{p_{lkm}} = \frac{G(x_{mk} \beta_m^*)}{G(x_{ml} \beta_m^*)}, \quad m = 1, \ldots, M; \quad k, l = 1, \ldots, K_m.
\]

**CASE 2:** The restrictions $H(x_m, \beta_m^*, \rho_m) = \sum_r G(x_{mr} \beta_m^*)$ ($m = 1, \ldots, M$) apply. Then (8) is simplified to

\[
\frac{p_{mk}}{p_{jl}} = \frac{G(z_m \alpha)G(x_{mk} \beta_m^*)}{G(z_j \alpha)G(x_{jl} \beta_j^*)}, \quad m, j = 1, \ldots, M; \quad k = 1, \ldots, K_m; \quad l = 1, \ldots, K_j.
\]

So far, the functions $G(\cdot)$ and $H(\cdot)$ have been specified as arbitrary functions. In the *Nested Logit* model we, in particular, do the following:

[1]: We parameterize $G(\cdot)$ as

\[
G(w \theta) = e^{w \theta} \implies G(w_1 \theta_1)G(w_2 \theta_2) \equiv G(w_1 \theta_1 + w_2 \theta_2).
\]

[2]: We rewrite $\beta_m^*$ as $\beta_m^* = \beta_m / \rho_m$ and at the same time assume

\[
H(x_m, \beta_m^*, \rho_m) = \left[ \sum_k G(x_{mk} \beta_m^*) \right]^{\rho_m} = \left[ \sum_k e^{x_{mk} (\beta_m / \rho_m)} \right]^{\rho_m}.
\]

The last assumption can be rewritten as

\[
H(x_m, \beta_m^*, \rho_m) = e^{w_m \rho_m}
\]
where

\[ w_m = \ln \left( \sum_l e^{x_{ml}(\beta_m/\rho_m)} \right) \quad \iff \quad e^{w_m} = \sum_l e^{x_{ml}(\beta_m/\rho_m)} \]

If \(\rho_1, \ldots, \rho_m\) are free parameters, it follows from (4)–(6) and (10) that

\[ p_{km} = \frac{e^{x_{mk}(\beta_m/\rho_m)} \sum_l e^{x_{ml}(\beta_m/\rho_m)} \rho_m}{\sum_j e^{z_j \alpha + w_j \rho_m} \sum_l e^{x_{jl}(\beta_j/\rho_j)} \rho_j} \]

\[ p_m = \frac{\sum_j e^{z_j \alpha + w_j \rho_m} \sum_l e^{x_{jl}(\beta_j/\rho_j)} \rho_j}{\sum_j e^{z_j \alpha + w_j \rho_m} \sum_l e^{x_{jl}(\beta_j/\rho_j)} \rho_j} \]

\[ p_{mk} = \frac{e^{z_m \alpha + w_m (\rho_m - 1) + x_{mk}(\beta_m/\rho_m)} \sum_l e^{x_{ml}(\beta_m/\rho_m)}}{e^{z_m \alpha + w_m r} \sum_l e^{x_{ml}(\beta_m/\rho_m)}} \quad m = 1, \ldots, M; \quad k = 1, \ldots, K_m. \]

Note that in (13) \(w_m\) enters as a regressor with coefficient \(\rho_m\). From the last equality it follows that the odds-ratios become:

\[ \frac{p_{mk}}{p_{jl}} = \frac{e^{z_m \alpha + w_m (\rho_m - 1) + x_{mk}(\beta_m/\rho_m)}}{e^{z_j \alpha + w_j (\rho_j - 1) + x_{jl}(\beta_j/\rho_j)}} \quad m, j = 1, \ldots, M; \quad k = 1, \ldots, K_m; \quad l = 1, \ldots, K_j. \]

Here the IIA-axiom is violated since variables and parameters outside alternatives (\(mk\)) and (\(jk\)) occur in the second term of the exponents of the numerator and the denominator, via the transformed variables \(w_m\) and \(w_j\).

Such a model can be utilized to take account of that fact that substitution possibilities or correlation pattern or association within groups (nests) differ from those between groups (nests). These can be represented by \(\kappa_m = 1 - \rho_m\) \((m = 1, \ldots, M)\), such that \(\rho_m = 1, \kappa_m = 0\) \((m = 1, \ldots, M)\) express that there is no correlation. \(1 - \kappa_m\) provides a measure of correlation within group \(m\) [see Train (2003, pp. 83–84)]. Certain authors have denoted \(\kappa_1, \ldots, \kappa_M\) as dissimilarity parameters.

In the special case \(\rho_m = 1 \implies \kappa_m = 0, \beta_m^* = \beta_m\) \((m = 1 \ldots, M)\), which corresponds to case 2 above, since

\[ H(x_m, \beta^*_m, 1) = \sum_k G(x_{mk}\beta_m) = \sum_k e^{x_{mk}\beta_m}, \]
it follows from (12)–(14) that

\[ p_{k|m} = \frac{e^{x_{mk} \beta_m}}{\sum_l e^{x_{ml} \beta_m}}, \]

\[ p_m = \frac{\sum_l e^{z_m \alpha + x_{ml} \beta_m}}{\sum_j \sum_l e^{z_j \alpha + x_{jl} \beta_j}}, \]

\[ p_{mk} = \frac{e^{z_m \alpha + x_{mk} \beta_m}}{\sum_r \sum_l e^{z_r \alpha + x_{rl} \beta_r}}, \]

while \( w_m \) and \( w_j \) drop out of (15), The last equation, (18), is a standard multinomial Logit probability, which confirms that the IIA-axiom holds. Then all \( J = \sum_{m=1}^M K_m \) choice alternatives are in a certain sense co-ordinated, being on the same level, since the substitution and correlation pattern is the same between all pairs of alternatives. The group dimension is uninteresting Nested Logit as a special case.

A REMARK ON ESTIMATION

In the nested logit case with free \( \rho_m \)s estimation can proceed stepwise as follows:

(i) Estimate \( \beta_m^* = \beta_m / \rho_m \) by Maximum Likelihood as in a standard multinomial Logit model for group by using the conditional probabilities (12).

(ii) Compute values of \( w_m \) \((m = 1, \ldots, M)\) from (11) by replacing \( \beta_m^* = \beta_m / \rho_m \) by their estimates from (i).

(iii) Estimate \( \alpha \) and \( \rho_m \) by Maximum Likelihood as in a standard multinomial Logit analysis from the last equality of (13), while replacing \( w_m \) \((m = 1, \ldots, M)\) by their values computed from (ii). Here \( \rho_m \) appear as a standard Logit parameter co-ordinated with the parameter vector \( \alpha \)

(iv) Finally, estimate \( \beta_m \) by multiplying the estimate of \( \beta_m^* = \beta_m / \rho_m \) from (i) by the estimate of \( \rho_m \) from (iii).

[This is a feasible procedure, but from an efficiency point of view not beyond critique; see Train (2003, Section 4.2.4).]

REFERENCE