Notes on Discrete Choice Models

Javier Fuertes Pina

March 2025

1 Introduction

Discrete choice models have frequently been employed to analyze decisions among mutually exclusive alternatives. Among these, the Multinomial Logit (MNL) model, introduced by [**?**], stands out due to its straightforward closed-form solution. This model is inherently linked to Random Utility Models (RUM), although it relies on a critical identifying assumption known as Independence of Irrelevant Alternatives (IIA). In many practical scenarios, however, this assumption lacks credibility.

In response to the limitations posed by IIA, several alternative models have been developed for discrete choice analysis. A prominent approach involves maintaining the error distribution assumption underlying the MNL model but allowing for correlation among subsets of alternatives by grouping them into nests. This structure facilitates the modeling of unobserved heterogeneity and introduces dependencies within nests while retaining the IIA assumption across nests. This model, known as the Nested Logit (NL) model, was first proposed by [**?**] and has since been extensively utilized to analyze hierarchical discrete choice problems in fields such as migration, spatial economics, and Industrial Organization (IO). In the NL framework, IIA is preserved at each hierarchical level, yet dependencies arise among products within the same nest. Empirical results obtained from this approach are sensitive to both the method and number of nests created

2 Multinomial Logit (MNL)

There are $\mathcal{J}$ single-product firms. In the MNL agents choose among the whole set of alternatives at the same time. They choose the alternative that maximizes

their utility among all the products:

$$U_{ij} = \beta' X_j + \alpha P_j + \xi_j + \nu_{ij}, \tag{1}$$

where the previous equation represents the utility of agent i by buying product j. The vector of observable characteristics of product j is X_j. The variable ξ_j represents the unobserved characteristics. P_j is the price of product j. The error term is represented by ε_j and follows a Type I Generalised Extreme Value (Gumbel distribution) distribution.

Let's define y_i a discrete variable indicating the alternative that agent. In addition, define $\mathbf{W} = (X, P, \xi)$

$$\begin{aligned}
\mathbb{P}(y_i = j|\mathbf{W}) &= \mathbb{P}(U_{ij} > U_{ik}, \forall k \neq j|\mathbf{W}) \\
&= \mathbb{P}(\nu_{ij} - \nu_{ik} > \beta'(X_k - X_j) + \alpha(P_k - P_j) + (\xi_k - \xi_j), \forall k \neq j|\mathbf{W}) \\
&= \mathbb{P}(\nu_{ij} - \nu_{ik} > V_{ik} - V_{ij}, \forall k \neq j|\mathbf{W}).
\end{aligned}$$

Lemma 2.1 *The difference of two Gumbel distributed random variables is distributed as a logistic if they share the same scale parameter,*

$$X \sim Gumbel(\mu_X, \beta); \quad Y \sim Gumbel(\mu_Y, \beta)$$

$$X - Y \sim Logistic(\mu_X - \mu_Y, \beta)$$

Proof:

Let X be a Gumbel distributed random variable with scale parameter μ_X and location β, and Y follows a Gumbel distribution with scale parameter μ_Y and location β. This can be represented as:

$$X = \mu_X + \beta G_1; \quad Y = \mu_Y + \beta G_2,$$

where G_1 and G_2 are two standard Gumbel distributed random variables (i.e. location parameter 0 and scale parameter 1). Then their respectively CDF and pdf are:

$$F_G(x) = exp(-e^{-x}), \quad x \in \mathbb{R} \tag{2}$$

$$f_G(x) = e^{-x} exp(-e^{-x}) \tag{3}$$

First, let's show that the difference between follows a standard Logistic distribution:

$$\begin{aligned}\mathbb{P}(G_1 - G_2 \leq z) = & \quad \text{(Using LIE and iid property of Gumbels)} \\ &= \int \mathbb{P}(G_1 \leq g_2 + z) f_{G_2}(g_2) dg_2 \\ &= \int F_{G_1}(g_2 + z) f_{G_2}(g_2) dg_2\end{aligned}$$

Using the formulas 2 and 3 we have that

$$\begin{aligned}\mathbb{P}(G_1 - G_2 \leq z) &= \int exp(-e^{-(g_2+z)}) \left[e^{-g_2} exp(-e^{-g_2})\right] dg_2 \\ &= \int e^{-g_2} exp(-e^{-g_2}(1 + e^{-z})) dg_2 \\ &= \left[\frac{exp(-e^{-g_2}(1 + e^{-z}))}{1 + e^{-z}}\right]_{g_2=-\infty}^{g_2=\infty} \\ &= \frac{exp(0)}{1 + e^{-z}} - 0 = \frac{1}{1 + e^{-z}}\end{aligned}$$

which is the CDF formula of a standard Logistic Random Variable. Now for $Z = X - Y$:

$$Z = \mu_X - \mu_Y + \beta(G_1 - G_2),$$

since $G_1 - G_2 \sim Logistic(0, 1)$ it is straight forward that:

$$Z \sim Logistic(\mu_X - \mu_Y, \beta) \tag{4}$$

■

Therefore, if we assume that $\nu_i = (\nu_{i1}, \ldots, \nu_{iJ})'$ is a vector of random variables that share common scale parameter, the probability of choosing alternative j is a multivariate logistic CDF, i.e.

$$\mathbb{P}(y_i = j|\mathbf{W}) = \mathbb{P}(\nu_{ij} - \nu_{ik} > V_{ik} - V_{ij}, \forall k \neq j|\mathbf{W}) \tag{5}$$

If $J = 2$ then we obtain the famous closed-form solution probability formulas, using the logistic CDF formulas.

$$\mathbb{P}(j = 1|\mathbf{W}) = \frac{1}{1 + exp(V_{i2} - V_{i1})} = \frac{exp(V_{i1})}{exp(V_{i1}) + exp(V_{i2})} \tag{6}$$

If $J \geq 3$ then expression (5) is a multivariate logistic distribution which does not have closed form solution. To find an expression similar to (6) we need to impose extra assumptions.

McFadden in his seminal paper proposed three assumptions in order to find the famous closed-form formulas. One of this is the **Independence of Irrelevant Alternatives (IIA)**. Many people confuse this with a condition of the parametric assumptions of the model. However, this is an identifying assumption.

Let me now introduce the assumptions in the way McFadden introduced them and explain the implications.

MNL1 (IIA): For all possible alternative sets $\mathcal{B}$, atributes s, and members x and y of $\mathcal{B}$.

$$\mathbb{P}(x|s, \{x, y\})\mathbb{P}(y|s, \mathcal{B}) = \mathbb{P}(y|s, \{x, y\})\mathbb{P}(x|s, \mathcal{B}).$$

MNL2 (Positivity): $\mathbb{P}(x|s, \mathcal{B}) > 0$ so that

$$\frac{\mathbb{P}(y|s, \mathcal{B})}{\mathbb{P}(x|s, \mathcal{B})} = \frac{\mathbb{P}(y|s, \{x, y\})}{\mathbb{P}(x|s, \{x, y\})}.$$

(note $\mathbb{P}(x|s, \mathcal{B}) > 0 \implies \mathbb{P}(x|s, \{x, y\}) > 0$)

This two assumptions imply that the odds between the probabilities of choosing two alternatives does not depend on the choice set that you are considering. **Note this assumption is probably not realistic in many situations**[1]

For simplicity, let me denote $\mathbb{P}(y|s, \{x, y\}) = P_{y,x}$ and symmetrically $\mathbb{P}(x|s, \{x, y\}) = P_{x,y}$, then

$$\mathbb{P}(y|s, \mathcal{B}) = \frac{P_{y,x}}{P_{x,y}}\mathbb{P}(x|s, \mathcal{B}), \tag{7}$$

and by definition of a probability,

$$1 = \sum_{y \in \mathcal{B}} \mathbb{P}(y|s, \mathcal{B}) = \left(\sum_{y \in \mathcal{B}} \frac{P_{y,x}}{P_{x,y}} \right) \mathbb{P}(x|s, \mathcal{B})$$

Therefore, we can use this last expression to give an expression for $\mathbb{P}(x|s, \mathcal{B})$,

$$\mathbb{P}(x|s, \mathcal{B}) = \frac{1}{\sum_{y \in \mathcal{B}} \frac{P_{y,x}}{P_{x,y}}} \tag{8}$$

[1] Check Famous RB-BB-C illustration (Example 2.1) proposed by McFadden to understand why this condition is not realistic in some settings.

Now, using same again the definition of IIA,

$$\frac{P_{y,x}}{P_{x,y}} = \frac{\mathbb{P}(y|s,\mathcal{B})}{\mathbb{P}(x|s,\mathcal{B})} \tag{9}$$

$$= \frac{\frac{P_{y,z}}{P_{z,y}}\mathbb{P}(z|s,\mathcal{B})}{\frac{P_{x,z}}{P_{z,x}}\mathbb{P}(z|s,\mathcal{B})} \tag{10}$$

$$= \frac{\frac{P_{y,z}}{P_{z,y}}}{\frac{P_{x,z}}{P_{z,x}}} \tag{11}$$

where second equatily follows form using representation from equation 7 wrt a different variable z$\in \mathcal{B}$. Now, plugging (11) into (8) we have that

$$\mathbb{P}(x|s,\mathcal{B}) = \frac{1}{\sum_{y\in\mathcal{B}} \frac{\frac{P_{y,z}}{P_{z,y}}}{\frac{P_{x,z}}{P_{z,x}}}} = \frac{\frac{P_{x,z}}{P_{z,x}}}{\sum_{y\in\mathcal{B}} \frac{P_{y,z}}{P_{z,y}}} \tag{12}$$

Let me define $l(s,x,z) = log(\frac{P_{x,z}}{P_{z,x}})$, then equation (12) is transformed into

$$\mathbb{P}(x|s,\mathcal{B}) = \frac{exp(l(s,x,z))}{\sum_{y\in\mathcal{B}} exp(l(s,y,z))} \tag{13}$$

MNL3 (Irrelevance of Alternative set effect): Let me define $l(s,x,z) = log(\frac{\mathbb{P}(x|s,\{x,z\})}{\mathbb{P}(z|s,\{x,z\})})$, where s is the taste effect, x is the choice alternative effect and z the alternative set effect. Then this function must be additive separable.

$$l(s,x,z) = f(s,x) - f(s,z).$$

where f(s,.) represents any function that depends only on variables of alternative x.

Notice This is not an extra assumption since the parametric assumption of the distribution in the error satisfies this assumption.

$$l(s,x,z) = log\left(\frac{\mathbb{P}(x|s,\{x,z\})}{\mathbb{P}(z|s,\{x,z\})}\right) \tag{14}$$

$$= log(exp(V_x)) - log(V_x + V_z) - log(V_z) + log(V_x + V_z) \tag{15}$$

$$= V_x - V_z \tag{16}$$

where second equality comes from expression in equation (6)). Then equation (13) can be reexpressed as

$$\mathbb{P}(x|s,\mathcal{B}) = \frac{exp(l(s,x,z))}{\sum_{y\in\mathcal{B}} exp(l(s,y,z))} = \frac{exp(V_x - V_z)}{\sum_{y\in\mathcal{B}} exp(V_y - V_z)} \tag{17}$$

$$= \frac{exp(V_x)exp(-V_z)}{exp(-V_z)\sum_{y\in\mathcal{B}} exp(V_y)} \tag{18}$$

$$= \frac{exp(V_x)}{\sum_{y\in\mathcal{B}} exp(V_y)}. \tag{19}$$

Notice that equation (19) is the famous expression of the MNL model (without normalizing any variable to be the outside option). V_x represent the deterministic part of the utility obtained from choosing product x in our model this expression would be:

$$\mathbb{P}(y = j|\mathbf{W}) = \frac{exp(\beta' X_j + \alpha P_j + \xi_j)}{\sum_{k=1}^{J} exp(\beta' X_k + \alpha P_k + \xi_k)} \tag{20}$$

Now, in this model the shares of alternatives chosen is the same as the probabilities of choosing that product, i.e.

$$s_j = \mathbb{P}(y = j|\mathbf{W}) = \frac{exp(\beta' X_j + \alpha P_j + \xi_j)}{\sum_{k=1}^{J} exp(\beta' X_k + \alpha P_k + \xi_k)}.$$

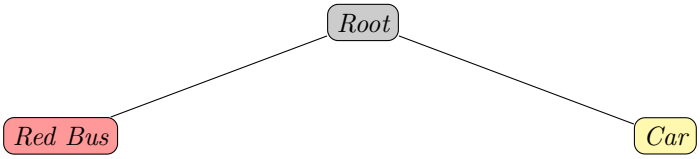
Therefore, from that formula we can compute some interesting statistics such as the price elasticities,

$$\epsilon_{j,k} = \frac{\partial s_j}{\partial P_k}\frac{P_k}{s_j} = \begin{cases} (\alpha s_j - \alpha s_j^2)\frac{P_k}{s_j} & if \quad k = j \\ -\alpha s_j^2 \frac{P_k}{s_j} & if \quad k \neq j \end{cases} \tag{21}$$

$$= \begin{cases} \alpha P_j(1 - s_j) & if \quad k = j \\ -\alpha P_k s_j & if \quad k \neq j \end{cases} \tag{22}$$

Example 2.1 *This example illustrates the unrealistic nature of the Independence of Irrelevant Alternatives (IIA) assumption by examining the classic "Red Bus, Blue Bus, and Car" scenario, as originally discussed by McFadden.*

Suppose an individual must choose a mode of transportation in a city. First, consider the case in which the available options are limited to a car and a red bus. The following diagram represents the decision tree for this initial choice set:



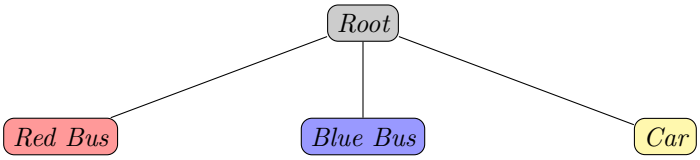
Under the assumption that the individual is indifferent between these two alternatives, we have

$$\mathbb{P}(RB \mid \mathcal{B} = \{RB, C\}) = \mathbb{P}(C \mid \mathcal{B} = \{RB, C\}) = 0.5.$$

Consequently, the ratio of these probabilities is

$$\frac{\mathbb{P}(C \mid \mathcal{B} = \{RB, C\})}{\mathbb{P}(RB \mid \mathcal{B} = \{RB, C\})} = 1.$$

Now, assume that a new alternative—the blue bus—is introduced. The following diagram depicts the decision structure when the blue bus is added:



In this extended choice set $\mathcal{B} = \{RB, BB, C\}$, if individuals remain indifferent regarding the color of the bus, one might expect the following:

$$\mathbb{P}(RB \mid \mathcal{B} = \{RB, BB, C\}) = \mathbb{P}(BB \mid \mathcal{B} = \{RB, BB, C\}) = 0.25,$$

and

$$\mathbb{P}(C \mid \mathcal{B} = \{RB, BB, C\}) = 0.5.$$

However, the IIA property enforces that the ratio of the probabilities between the car and the red bus remains unchanged:

$$\frac{\mathbb{P}(C \mid \mathcal{B} = \{RB, C\})}{\mathbb{P}(RB \mid \mathcal{B} = \{RB, C\})} = \frac{\mathbb{P}(C \mid \mathcal{B} = \{RB, BB, C\})}{\mathbb{P}(RB \mid \mathcal{B} = \{RB, BB, C\})} = 1.$$

This constraint then implies that in the presence of the blue bus, the probabilities would adjust to

$$\mathbb{P}(RB \mid \mathcal{B} = \{RB, BB, C\}) = \mathbb{P}(BB \mid \mathcal{B} = \{RB, BB, C\}) = \mathbb{P}(C \mid \mathcal{B} = \{RB, BB, C\}) = \frac{1}{3},$$

which is counterintuitive; one would not expect the introduction of a similar alternative (another bus) to cause the car's probability to drop from 0.5 to approximately 0.33.

This example demonstrates that the IIA assumption can lead to unintuitive predictions when a new alternative similar to an existing one is introduced.

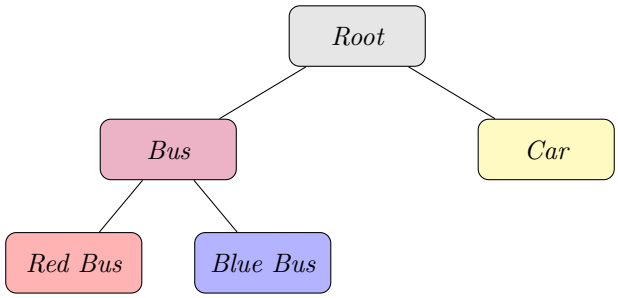
3 Nested Logit (NL)

The results derived from the Multinomial Logit (MNL) model rely heavily on the assumption that the random disturbances are independent and identically distributed (i.i.d.) with constant variance (i.e., homoscedastic). One approach to relaxing this assumption is to classify alternatives into subgroups, thereby allowing for heteroskedasticity across these groups. This leads to the **Nested Logit (NL)** model, where similar alternatives are grouped into "nests" that share common unobserved components in their error terms.

In the NL specification, alternatives within the same nest exhibit correlated unobserved utility components, captured by a nest-specific scale parameter, denoted λ_l. In contrast, alternatives belonging to different nests remain uncorrelated. Unlike the MNL model, the NL model accommodates correlations among alternatives, offering a more flexible framework for modeling choice behavior.

For a comprehensive reference on the Nested Logit model, see [**?**].

Example 3.1 *Going back to the RB-BB-C example, the basic idea is that instead of considering all the alternatives as independent, we group both buses into the same nest.*



This way $\mathbb{P}(C|\mathcal{B} = \{RB, BB, C\}) = \mathbb{P}(BUS|\mathcal{B} = \{BUS, C\}) = 0.5$ *and for BB and RB* $\mathbb{P}(RB|\mathcal{B} = \{RB, BB, C\}) = \mathbb{P}(BB|\mathcal{B} = \{RB, BB, C\}) = 0.25$*, which is more realistic*

3.1 Variance Decomposition of the NL

To model de variance decomposition we follow [**?**] where he decompose the disturbance into two terms. Let me define $\nu_{i,j,l} = \nu_{il} + \lambda_l \nu_{ij}$, where the first term is the taste of agent i to the products in the nest l. The term ν_{il} is the specific taste of agent i for alternative j. Parameter λ_l and the specific form of the error term are explained in the following theorem from Cardell's paper.

Theorem 3.1 *For $0 < \lambda < 1$ and ε, a random variable distributed as Type I extreme value, there exist a unique distribution, denoted as $C(\lambda)$ such that for ν, a random variable, ν and ε independent, then $\nu + \lambda\varepsilon$ is a random variable distributed as Type I extreme value, iff ν is distributed as $C(\lambda)$, where its pdf of this distribution is*

$$f_\lambda(\nu) = \frac{1}{\lambda} \sum_{n=0}^{\infty} \frac{(-1)^n e^{n\nu}}{n!\Gamma(-n\lambda)}. \tag{23}$$

There is no closed form solution for the CDF of $C(\lambda)$

For the $C(\lambda)$ distribution, the single parameter λ determines the mean, the shape, and the scale of the distribution. It is sometimes useful to generalize the $C(\lambda)$ distribution to include a scale parameter. Accordingly, if ν is distributed as $C(\lambda)$ and δ is a fixed scalar, $\delta\nu$ is said to be distributed as $C(\lambda, \delta)$.

Therefore Applying Theorem 3.1 repeatedly, one can create recursively a variance components structure following a Type I Extreme Value distribution.

Example 3.2 *Suppose ν_1, ν_2, ν_3 and ε independent random variables. In addition, $\nu_k \sim C(\lambda_k) \quad \forall k = 1, 2, 3$ and $\varepsilon \sim Gumbel$, then*

$$\begin{aligned}
&\nu_3 + \lambda_3\varepsilon \sim Gumbel \\
&\nu_2 + \lambda_2(\nu_3 + \lambda_3\varepsilon) \sim Gumbel \\
&\nu_1 + \lambda_1(\nu_2 + \lambda_2(\nu_3 + \lambda_3\varepsilon)) \sim Gumbel
\end{aligned}$$

where first line follows directly from Theorem 3.1, second from the fact that $\nu_3 + \lambda_3\varepsilon \sim Gumbel$ and Theorem 3.1 and third line from $\nu_2 + \lambda_2(\nu_3 + \lambda_3\varepsilon) \sim Gumbel$ and Theorem 3.1. This way we can construct a higher order tree, each term in the variance component is one level in the tree. In this example we have three levels.

To construct the general setting for a higher order tree let me present theorem 2.3 from [**?**].

Theorem 3.2 *Let Q be the number of terms in the variance component structure. Given $0 \leq \lambda_k \leq 1$ for $1 \leq k \leq Q$, $\lambda_0 = 1$, $\xi_k = \Pi_{l=0}^{k}\lambda_l$. It follows that:*

1. *If ν_k is independently distributed as $C(\lambda_k)$ for $k = 1, \ldots, Q$, then $\sum_{l=0}^{k} \xi_{k-1}\nu_k \sim C(\xi_K)$*

2. *Alternatively, for η_k independently distributed as $C(\lambda_k, \xi_{k-1})$, then $\sum_{k=1}^{K} \eta_k \sim C(\xi_K)$.*

To proof this theorem let me first state and important lemma derived from the Cardell's distribution.

Lemma 3.1 *For $\nu_1 \sim C(\lambda_1)$ and $\nu_2 \sim C(\lambda_2)$, then ν_1 and ν_2 are independent and $\nu_1 + \nu_2 \sim C(\lambda_1 \lambda_2)$*

Proof (Statetment 1): I am going to prove by induction,

$$k = 1; \quad \nu_1 \sim C(\lambda_1) \quad by \quad Assumption$$
$$k = 2; \quad \nu_1 + \lambda_1 \nu_2 = \xi_0 \nu_1 + \xi_1 \nu_2 \sim C(\lambda_1) =_d C(\xi_2) \quad by \quad Lemma \quad 3.1$$

Induction hypothesis:

$$For \quad k = L, \quad \sum_{k=1}^{L} \xi_{k-1} \nu_k \sim C(\Pi_{i=1}^{L} \lambda_i) =_d C(\xi_L)$$

Let me define $\eta = \sum_{k=1}^{L} \xi_{k-1} \nu_k$, which is distributed as $C(\xi_L)$ by induction hypothesis.

$$For \quad k = L+1, \quad \sum_{k=1}^{L+1} \xi_{k-1} \nu_k = \eta + \xi_L \nu_{L+1} \sim C(\Pi_{i=1}^{L+1} \lambda_i) =_d C(\xi_{L+1}) by \quad Lemma \quad 3.1$$

■

Proof (Statetment 2): The proof for this second statement is almost identical to the previous one using induction. We just need to redefine the problem:

$$\eta_k \sim C(\lambda_k, \xi_{k-1}) \quad \Longrightarrow \quad \frac{\eta_k}{\xi_{k-1}} \sim C(\lambda_k)$$

■

Then $\sum_{k=1}^{Q} \xi_{k-1} \nu_k$ is a Q-term variance component structure for a single Type I extreme-value random variable. Now, consider a vector of J Type I extreme-value random variables denoted by $\zeta_j = \sum_{k=1}^{Q_j} \xi_{k-1,j} \nu_{i,k-1,j}$ for $1 \leq j \leq J$. The variable $\nu_{i,k,j} \sim C(\lambda_{k,j})$ $\forall i, k$ and $j = 1, \ldots, J$. While $\nu_{i,Q_j,j}$ for $1 \leq j \leq J$. Let

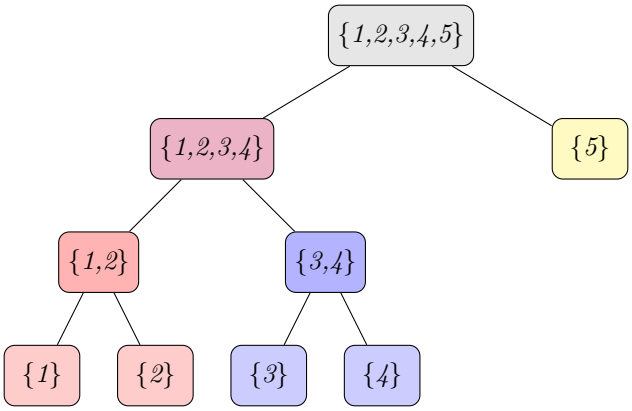
Q_j denote the number of terms in the variance components structure of ζ_j. Let $\lambda_{k,j}$ be fixed constants where $0 \leq \lambda_{k,j} \leq 1$ for $1 \leq j \leq J$ and $1 \leq k \leq Q_j$. Then the utility in a nested logit would be:

$$U_{i,j} = X'_{i,j}\beta + \sum_{k=1}^{Q_j} \xi_{k-1,j}\nu_{i,k,j} = X'_{i,j}\beta + \zeta_j \tag{24}$$

Example 3.3 *In order to understand better how the error term is modeled let's see the following example. Suppose* $\mathcal{J} = \{1, 2, 3, 4, 5\}$*, where*

1. *Car alone*
2. *Carpool*
3. *Bus*
4. *Subway*
5. *Walk*

Let's say $Q_1 = Q_2 = Q_3 = Q_4 = 3$ *and* $Q_5 = 1$*. This implies that alternatives* $1, 2, 3, 4$ *will be in a three level hierarchi grouped in different nests, while* 5 *will be in a single node. Suppose the tree is of the form*



This implies in terms of the error components (let me drop the subindex i for simplicity):

$$
\begin{aligned}
U_1 &= X_1'\beta + \nu_{1,1} + \lambda_{1,1}\nu_{2,1} + \lambda_{1,1}\lambda_{2,1}\nu_{3,1} \\
U_2 &= X_2'\beta + \nu_{1,2} + \lambda_{1,2}\nu_{2,2} + \lambda_{1,2}\lambda_{2,2}\nu_{3,2} \\
U_3 &= X_1'\beta + \nu_{1,3} + \lambda_{1,3}\nu_{2,3} + \lambda_{1,3}\lambda_{2,3}\nu_{3,3} \\
U_4 &= X_1'\beta + \nu_{1,4} + \lambda_{1,4}\nu_{2,4} + \lambda_{1,4}\lambda_{2,4}\nu_{3,4} \\
U_5 &= X_1'\beta + \nu_{1,5}
\end{aligned}
$$

Now we understand the variance decomposition of the error term in the NL model let's review Generalized Extreme Value Models (GEV) theory.

3.2 Generalized Extreme Value Model (GEV)

For a good reference regarding Nested Logit see [**?**] and [**?**].

This model allows for some pattern of dependency among the unobserved attributes if the alternatives and yields an analytically tractable closed form for the probabilities.

Suppose $G(x_1, \ldots, x_J)$ is a non-negative, homogenous of degree one function of $(x_1, \ldots, x_J) \geq 0$. In addition, suppose,

- $G \to \infty$ if $x_i \to \infty$ for each i and,
- for k distinct components

$$\frac{\partial^k G}{\partial x_1 \partial \ldots \partial x_k}$$

  is non-negative if k is odd, and nonpositive if k is even.

Then,

$$\mathbb{P}(y = j) = \frac{exp(V_j)G(exp(V_1), \ldots, exp(V_J))}{G(exp(V_1), \ldots, exp(V_J))},$$

which defines a probabilistic choice model form alternatives $j = 1, \ldots, J$, which is consistent with utility maximization. Further, expected maximum utility, defined by

$$\overline{U} = \int \max_j (V_j + \varepsilon_j) f(\varepsilon) d\varepsilon = log(G(exp(V_1), \ldots, exp(V_J)) + \gamma,$$

where $\gamma = 0.57721$ is Euler's constant.

Proof: Let's start by defining

$$M \;=\; \max_{1\le j\le J}\big(V_j + \varepsilon_j\big).$$

where ε_j follows a Gumbel distribution. By definition of M,

$$\Pr(M \le z) \;=\; \Pr\Big(\varepsilon_j \le z - V_j \;\text{ for all } j\Big).$$

Using the GEV form,

$$\Pr(\varepsilon_1 \le x_1, \ldots, \varepsilon_J \le x_J) \;=\; \exp\Big\{-G\big(\exp(-x_1), \ldots, \exp(-x_J)\big)\Big\},$$

we substitute $x_j = z - V_j$. Then

$$\Pr(M \le z) \;=\; \exp\Big\{-G\big(\exp(-(z-V_1)), \ldots, \exp(-(z-V_J))\big)\Big\}.$$

Factor out $\exp(-z)$ in each argument of G:

$$\exp(-(z - V_j)) \;=\; e^{-z}\, e^{V_j}.$$

By the 1-homogeneity of G, we have

$$G\big(e^{-z}\, e^{V_1}, \ldots, e^{-z}\, e^{V_J}\big) \;=\; e^{-z}\, G\big(e^{V_1}, \ldots, e^{V_J}\big).$$

Let

$$A \;=\; \ln\Big[G\big(e^{V_1}, \ldots, e^{V_J}\big)\Big].$$

Then

$$\Pr(M \le z) \;=\; \exp\Big\{-e^{-z}\, e^{A}\Big\} \;=\; \exp\big[-\exp\big(-(z-A)\big)\big].$$

This is precisely the distribution function of a Gumbel$(A, 1)$ random variable. Hence $M \sim$ Gumbel$(A, 1)$. It is well known that if $X \sim$ Gumbel$(\mu, 1)$, then $\mathbb{E}[X] = \mu + \gamma$, where γ is the Euler–Mascheroni constant. Therefore,

$$\mathbb{E}[M] \;=\; A \;+\; \gamma \;=\; \ln\Big[G\big(e^{V_1}, \ldots, e^{V_J}\big)\Big] \;+\; \gamma,$$

as was to be shown.

■

The expected maximum utility satisfies that the probability of choosing alternative j is obtained from the derivative $\overline{U}$ wrt V_j, i.e.

$$\frac{\partial}{\partial V_j}\mathbb{E}[M] \;=\; \frac{\partial}{\partial V_j}\Big\{\ln\Big[G\big(e^{V_1}, \ldots, e^{V_J}\big)\Big]\Big\} \;=\; \frac{1}{G\big(e^{V_1}, \ldots, e^{V_J}\big)}\,\frac{\partial}{\partial V_j}\,G\big(e^{V_1}, \ldots, e^{V_J}\big).$$

By the chain rule,

$$\frac{\partial}{\partial V_j} G\big(e^{V_1}, \ldots, e^{V_J}\big) \;=\; \frac{\partial G}{\partial x_j}\big(e^{V_1}, \ldots, e^{V_J}\big) \cdot \frac{d}{dV_j}\big(e^{V_j}\big) \;=\; \frac{\partial G}{\partial x_j}(e^{V_1}, \ldots, e^{V_J}) \cdot e^{V_j}.$$

Thus,

$$\frac{\partial}{\partial V_j}\mathbb{E}[M] \;=\; \frac{e^{V_j}\,\left[\frac{\partial G}{\partial x_j}(e^{V_1}, \ldots, e^{V_J})\right]}{G\big(e^{V_1}, \ldots, e^{V_J}\big)}.$$

One fundamental property of GEV models is that the *choice probability* of alternative j can be written (in simplified notation) as

$$P_j \;=\; \frac{e^{V_j}\,\frac{\partial G}{\partial x_j}\big(e^{V_1}, \ldots, e^{V_J}\big)}{G\big(e^{V_1}, \ldots, e^{V_J}\big)}.$$

(For multinomial logit, for example, $G(\mathbf{x}) = \sum_k x_k$, hence $P_j = \frac{e^{V_j}}{\sum_k e^{V_k}}$).

The Generating GEV function for Nested Logit with L nests is

$$G(V_1, \ldots, V_J) = \sum_{l=1}^{L} a_l \left[\sum_{i \in B_l} exp\left(\frac{V_j}{\lambda_l}\right)\right]$$

This function is known as multivariate Extreme Value Distribution. Where as mentioned above λ_l is a measure of correlation among alternatives in nest l. The parameter a_l **READ MORE ABOUT THIS PARAMETER** and both λ_l and a_l can be also estimated among the rest parameters of the models.

Example 3.4 *Suppose,*

$$G(V_1, V_2, V_3) = exp(V_1) + \left[exp(V_1)^{\frac{1}{\lambda}} + exp(V_2)^{\frac{1}{\lambda}}\right]^{\lambda}$$

where 1 is an alternative in one nest and 2 and 3 are in another nest.

$$\mathbb{P}(1|\{1,2,3\}) = \frac{exp(V_1)}{exp(V_1) + (exp(\frac{V_2}{\lambda}) + exp(\frac{V_3}{\lambda}))^{\lambda}} \tag{25}$$

$$\mathbb{P}(k|\{1,2,3\}) = \frac{exp(\frac{V_k}{\lambda})(exp(\frac{V_2}{\lambda}) + exp(\frac{V_3}{\lambda}))^{\lambda-1}}{exp(V_1) + (exp(\frac{V_2}{\lambda}) + exp(\frac{V_3}{\lambda}))^{\lambda}} \quad \forall k = 2,3 \tag{26}$$

Notice that if $\lambda = 1$ (common scale parameter for all nests) we are back to the MNL (this should not suprise to those who have read [?] since when $\lambda = 1$, $\nu_{i,l} = 0$ a.s). When $\lambda = 0$ then,

$$\mathbb{P}(1|\{1,2,3\}) = \frac{exp(V_1)}{exp(V_1) + max\{exp(V_2), exp(V_3))}$$

$$\mathbb{P}(k|\{1,2,3\}) = \begin{cases} \frac{exp(V_k)}{exp(V_1)+exp(V_k)} & if \quad V_k > V_h \\ \frac{0.5exp(V_k)}{exp(V_1)+exp(V_k)} & if \quad V_2 = V_3 \\ 0 & if \quad V_k < V_h \end{cases} \qquad \forall k \neq h \quad \& \quad k, h \in \{2,3\}$$

Which imply that alternative 2 and 3 are perfectly substitutes. Now to understand how IIA affects the nested logit check the following probabilities,

$$\mathbb{P}(m|\{1,2\}) = \frac{exp(V_m)}{exp(V_1) + exp(V_2)} \quad m \in \{1,2\} \tag{27}$$

Notice that compare to the odds that we would have with (25) and (26) the IIA is not satisfied. This shows that in the NL the IIA is not satisfied across products in different nests.

$$\mathbb{P}(k|\{2,3\}) = \frac{exp(\frac{V_k}{\lambda})}{exp(\frac{V_2}{\lambda}) + exp(\frac{V_3}{\lambda})} \quad k \in \{2,3\} \tag{28}$$

While in this case adding the alternative 1 does not affect the odds, i.e

$$\frac{\mathbb{P}(2|\{2,3\})}{\mathbb{P}(3|\{2,3\})} = \frac{\mathbb{P}(2|\{1,2,3\})}{\mathbb{P}(3|\{1,2,3\})} = exp(V_2 - V_3)$$

Therefore, IIA is still applyied for alternatives within the same nest.

3.3 Our model

Going back to our setting, we assume there are $\mathcal{J}$ single-product firms, each of the alternatives belong to one of the $\mathcal{L}$ nests and in each nest there is J_l alternatives. Where $\cup_l J_l = \mathcal{J}$, this means a nested partition of the choice set.

Then the utility of agents is defined as:

$$U_{ijl} = \beta' X_j + \alpha P_j + \gamma' \tilde{X}_l + \xi_j + \nu_l + \lambda_l \nu_{ij}, \tag{29}$$

where ν_l follows a Cardell distribution with parameter $\lambda_l \in (0,1)$ ($C(\lambda_l)$). From Theorem 3.1, $\nu_{i,j,l} = \nu_l + \lambda_l \nu_{ij}$ follows a Gumbel distribution as discussed above. we have also added a vector of observed characteristics of the nest, $\tilde{X}_l$. Rest of the variables are the same as defined in the MNL. With this model we are able to establish some pattern of dependency among unobserved attributes of alternatives (through λ_l) and at the same time yields an analytically tractable closed form for the probabilities.

Using the results from previous subsection, we now that the probability of choosing alternative j (without conditioning on this alternative belonging to any specific nest) is:

$$\begin{aligned}
\mathbb{P}(y = j, j \in l|\mathbf{W}) &= \frac{exp(\frac{\beta' X_j + \alpha P_j + \gamma' \tilde{X}_l + \xi_j}{\lambda_l}) \left[\sum_{m=1}^{J_l} exp(\frac{\beta' X_m + \alpha P_m + \gamma' \tilde{X}_l + \xi_m}{\lambda_l})\right]^{\lambda_l - 1}}{\sum_{b=1}^{L} \left\{ \left(\sum_{m=1}^{J_b} exp(\frac{\beta' X_m + \alpha P_m + \gamma' \tilde{X}_b + \xi_m}{\lambda_b})\right)^{\lambda_b} \right\}} \\
&= \frac{exp(\frac{\beta' X_j + \alpha P_j + \gamma' \tilde{X}_l + \xi_j}{\lambda_l}) exp(\frac{(\lambda_l - 1)}{\lambda_l} \gamma' \tilde{X}_l) \left[\sum_{m=1}^{J_l} exp(\frac{\beta' X_m + \alpha P_m + \xi_m}{\lambda_l})\right]^{\lambda_l - 1}}{\sum_{b=1}^{L} \left\{ \left(\sum_{m=1}^{J_b} exp(\frac{\beta' X_m + \alpha P_m + \gamma' \tilde{X}_b + \xi_m}{\lambda_b})\right)^{\lambda_b} \right\}} \\
&= \frac{exp(\frac{\beta' X_j + \alpha P_j + \xi_j}{\lambda_l} + \gamma' \tilde{X}_l) \left[\sum_{m=1}^{J_l} exp(\frac{\beta' X_m + \alpha P_m + \xi_m}{\lambda_l})\right]^{\lambda_l - 1}}{\sum_{b=1}^{L} exp(\gamma' \tilde{X}_b) \left\{ \left(\sum_{m=1}^{J_b} exp(\frac{\beta' X_m + \alpha P_m + \xi_m}{\lambda_b})\right)^{\lambda_b} \right\}}
\end{aligned}$$

Let me define the Inclusive value of nest l as $I_l = log\left(\sum_{m=1}^{J_l} exp(\frac{\beta' X_m + \alpha P_m + \xi_m}{\lambda_l})\right)$. Finally we obtain that the probability is

$$\mathbb{P}(y = j, j \in l|\mathbf{W}) = \frac{exp(\frac{\beta' X_j + \alpha P_j + \xi_j}{\lambda_l} + \gamma' \tilde{X}_l + (\lambda_l - 1) I_l)}{\sum_{b=1}^{L} \left\{ \exp(\gamma' \tilde{X}_b + \lambda_b I_b) \right\}} \tag{30}$$

Now the probability of choosing alternative j given that we know it is in nest l it a one-to-one mapping with the probabilities from the MNL. This is because once we condition on a given nest the IIA is satisfied for all those alternatives so we are back to a MNL.

$$\mathbb{P}(y=j|j\in l,\mathbf{W}) = \mathbb{P}(U_{ij}\geq U_{ik}|\forall j\neq k,\quad j,k\in l,\mathbf{W})$$
$$= \mathbb{P}(\beta' X_j + \alpha P_j + \gamma'\tilde{X}_l + \xi_j + \nu_l + \lambda_l\nu_{ij} \geq \beta' X_k + \alpha P_k + \gamma'\tilde{X}_l + \xi_k + \nu_l + \lambda_l\nu_{ik}|\forall j\neq k,\quad j,k\in l,\mathbf{W})$$
$$= \mathbb{P}(\lambda_l(\nu_{ij}-\nu_{ik}) \geq \beta'(X_k - X_j) + \alpha(P_k - P_j) + \xi_k - \xi_j|\forall j\neq k,\quad j,k\in l,\mathbf{W})$$
$$= \mathbb{P}\left(\nu_{ij}-\nu_{ik} \geq \frac{\beta'(X_k - X_j) + \alpha(P_k - P_j) + \xi_k - \xi_j}{\lambda_l}|\forall j\neq k,\quad j,k\in l,\mathbf{W}\right),$$

where the LHS of the inequality of last line follows a logistic distribution. Then we have a logistic multivariate distribution, but IIA is applying so the expression of that probability is just

$$\mathbb{P}(y=j|j\in l,\mathbf{W}) = \frac{exp(\frac{\beta' X_j + \alpha P_j + \xi_j}{\lambda_l})}{\sum_{m=1}^{J_l} exp(\frac{\beta' X_m + \alpha P_m + \xi_j}{\lambda_l})} \tag{31}$$

Finally, the probability of choosing nest l is

$$\begin{aligned}\mathbb{P}(l|\mathbf{W}) &= \sum_{j=1}^{J_l}\mathbb{P}(y=j, j\in l|\mathbf{W}) \\ &= \frac{\sum_{j=1}^{J_l} exp\left(\frac{\beta' X_j + \alpha P_j + \xi_j}{\lambda_l} + \gamma'\tilde{X}_l + (\lambda_l - 1)I_l\right)}{\sum_{b=1}^{L}\exp(\gamma'\tilde{X}_b + \lambda_b I_b)\Big\}} \\ &= \frac{exp\left(I_l + \gamma'\tilde{X}_l + (\lambda_l - 1)I_l\right)}{\sum_{b=1}^{L}\exp(\gamma'\tilde{X}_b + \lambda_b I_b)\Big\}}\end{aligned}$$

Then the formula is

$$\mathbb{P}(l|\mathbf{W}) = \frac{exp\left(\gamma'\tilde{X}_l + \lambda_l I_l\right)}{\sum_{b=1}^{L}\exp(\gamma'\tilde{X}_b + \lambda_b I_b)\Big\}} \tag{32}$$

As you can notice equation 32 has a MNL-formula form. This is because NL assumes IIA in each of the levels of the tree across alternatives that share the same root in each level. This is in the graph of example (3.3) we would have

- There would be IIA between $\mathbb{P}(\{1,2,3,4\})$ and $\mathbb{P}(\{5\})$

- IIA among $\mathbb{P}(\{1,2\}|\{1,2,3,4\})$ and $\mathbb{P}(\{3,4\}|\{1,2,3,4\})$
- IIA among $\mathbb{P}(\{1\}|\{1,2\})$ and $\mathbb{P}(\{2\}|\{1,2\})$ and;
- IIA among $\mathbb{P}(\{3\}|\{3,4\})$ and $\mathbb{P}(\{4\}|\{3,4\})$

With equations 30, 31 and 32 we can characterize the whole Nested Logit Model.

Elasticities in this model have an special characteristic. **Cross-Price Elasticities wrt other product's prices among products in the same nest are the same**. This reflects the an homogeneity condition of products from a same nest, as [**?**] explains in a similar setting. Mathematically means that $\forall j, m \in l$ and $\forall k \neq m, k$:

$$\epsilon_{jk} = \frac{\partial \mathbb{P}(y=j, j \in l|\mathbf{W})}{\partial P_k} \frac{P_k}{\mathbb{P}(y=j, m \in l|\mathbf{W})} = \frac{\partial \mathbb{P}(y=m, j \in l|\mathbf{W})}{\partial P_k} \frac{P_k}{\mathbb{P}(y=m, m \in l|\mathbf{W})} = \epsilon_{mk}$$

In this model, notice that shares and probabilities are the same. Now let's develope the algebra of elasticities. From now on I will omit the conditional on observables ($\mathbf{W}$) to reduce notation.

First, for the cross-price elasticity

$$\begin{aligned}
&\frac{\partial exp(\frac{\beta' X_j + \alpha P_j + \xi_j}{\lambda_l} + \gamma' \tilde{X}_l + (\lambda_l - 1) I_l)}{\partial P_k} = \frac{\partial exp(\frac{\beta' X_j + \alpha P_j + \xi_j}{\lambda_l} + \gamma' \tilde{X}_l + (\lambda_l - 1) I_l)}{\partial I_l} \frac{\partial I_l}{\partial P_k} \mathbb{1}\{k \in l\} \\
&= (\lambda_l - 1) exp\left(\frac{\beta' X_j + \alpha P_j + \xi_j}{\lambda_l} + \gamma' \tilde{X}_l + (\lambda_l - 1) I_l\right) \frac{\partial log\left(\sum_{j=1}^{J_l} exp\left(\left(\frac{\beta' X_j + \alpha P_j}{\lambda_l}\right)\right)\right)}{\partial P_k} \mathbb{1}\{k \in l\} \\
&= (\lambda_l - 1) exp\left(\frac{\beta' X_j + \alpha P_j + \xi_j}{\lambda_l} + \gamma' \tilde{X}_l + (\lambda_l - 1) I_l\right) \frac{\alpha}{\lambda_l} \frac{exp\left(\frac{\beta' X_k + \alpha P_k + \xi_k}{\lambda_l}\right)}{\sum_{j=1}^{J_l} exp\left(\frac{\beta' X_j + \alpha P_j}{\lambda_l}\right)} \mathbb{1}\{k \in l\}
\end{aligned}$$

Now,

$$\frac{\partial \sum_{b=1}^{L} exp\left(\tilde{X}_b'\gamma + \lambda_b I_b\right)}{\partial P_k} = \sum_{b=1}^{L} \mathbb{1}\{k \in b\} \frac{\partial exp\left(\tilde{X}_l'\gamma + \lambda_b I_b\right)}{\partial I_b} \frac{\partial I_b}{\partial P_k}$$
$$= \sum_{b=1}^{L} \mathbb{1}\{k \in b\} \lambda_b exp\left(\tilde{X}_b'\gamma + \lambda_b I_b\right) \frac{\partial log\left(\sum_{j=1}^{J_b} exp\left(\left(\frac{\beta' X_j + \alpha P_j}{\lambda_b}\right)\right)\right)}{\partial P_k}$$
$$= \sum_{b=1}^{L} \mathbb{1}\{k \in b\} \lambda_b exp\left(\tilde{X}_b'\gamma + \lambda_b I_b\right) \frac{\alpha}{\lambda_l} \frac{exp\left(\frac{\beta' X_k + \alpha P_k + \xi_k}{\lambda_b}\right)}{\sum_{j=1}^{J_l} exp\left(\frac{\beta' X_j + \alpha P_j}{\lambda_b}\right)}$$

Therefore, combining both expressions and after some algebra we get

$$\frac{\partial \mathbb{P}(y = j, j \in l)}{\partial P_k} = \begin{cases} -\alpha \mathbb{P}(y = j, j \in l) \left(\left(\frac{1-\lambda_l}{\lambda_l}\right) \mathbb{P}(y = k | k \in l) + \mathbb{P}(y = k, k \in l)\right) & for \quad k \neq j \quad \& \quad k \in l \\ -\alpha \mathbb{P}(y = j, j \in l) \mathbb{P}(y = k, k \in l) & for \quad k \neq j \quad \& \quad k \in h \neq l \end{cases}$$

Now we can calculate a general expression for the cross product elasticities:

$$\epsilon_{jk} = \frac{\partial \mathbb{P}(y = j, j \in l)}{\partial P_k} \frac{P_k}{\mathbb{P}(y = j, m \in l)}$$
$$= \alpha P_k \left[\frac{\left(\mathbb{1}\{k \in l\}(1 - \frac{1}{\lambda_l}) \mathbb{P}(y = k | k \in l) \left(\sum_{b=1}^{L} exp\left(\tilde{X}_b'\gamma + \lambda_b I_b\right)\right)\right)}{\sum_{b=1}^{L} exp\left(\tilde{X}_b'\gamma + \lambda_b I_b\right)} \right.$$
$$\left. - \frac{\left(\sum_{b=1}^{L} \mathbb{1}\{k \in b\} exp(\tilde{X}_b'\gamma \lambda_b I_b) \mathbb{P}(y = k | k \in b)\right)}{\sum_{b=1}^{L} exp\left(\tilde{X}_b'\gamma + \lambda_b I_b\right)} \right].$$

This implies if $k \neq j \quad \& \quad k \in l$

$$\epsilon_{jk} = -\alpha P_k \mathbb{P}(y = k | k \in l)(\mathbb{P}(l) + \frac{1 - \lambda_l}{\lambda_l}), \tag{33}$$

if $k \neq j \quad \& \quad k \in h \neq l$

$$\epsilon_{jk} = -\alpha P_k \mathbb{P}(y = k | k \in h) \mathbb{P}(h) = -\alpha P_k \mathbb{P}(y = k, k \in h), \tag{34}$$

and finally the own price elasticity we need to calculate

$$\frac{\partial}{\partial P_j}\exp\left(\frac{\beta' X_j + \alpha P_j + \xi_j}{\lambda_l} + \gamma'\tilde{X}_l + (\lambda_l - 1)I_l\right)$$
$$= \frac{\alpha}{\lambda_l}\left[\exp\left(\frac{\beta' X_j + \alpha P_j + \xi_j}{\lambda_l} + \gamma'\tilde{X}_l + (\lambda_l - 1)I_l\right)\right.$$
$$\left. + (\lambda_l - 1)\exp\left(\frac{\beta' X_j + \alpha P_j + \xi_j}{\lambda_l} + \gamma'\tilde{X}_l + (\lambda_l - 1)I_l\right)\mathbb{P}(y = j \mid j \in l)\right].$$

Then we can compute,

$$\epsilon_{jj} = \alpha P_j\left(\left(\frac{1}{\lambda_l} - 1\right)(1 - \mathbb{P}(y = j|j \in l)) + (1 - \mathbb{P}(y = j, j \in l))\right) \qquad (35)$$

Notice that if we set $\lambda_l = 1 \quad \forall l$ we go back to the MNL formolas.