Discrete Choice Model - Eurobarometer Survey

Derivation of a Multichoice Logit model

Thomas Monk, 15 July 2020

This derivation adapts Ophem, Stam et al. (1999), and Ben-Akiva, Lerman et al. (1985, pp 104-107).

1 Additive random utility model

The agent is presented with a normalised set of J options. She is instructed to pick the j_1 most important options from that set. In what follows, we consider the case $j_1 = 2$, but the derivation can be extended to higher j_1.
Let there be J alternatives, indexed by $j \in \mathcal{J} = \{1, \ldots, J\}$. The utility generated by alternative j for a given individual (individual subscripts are suppressed for simplicity) is

$$U_j = V_j + \epsilon_j, \tag{1}$$

where V_j is the deterministic component of utility, depending on individual and alternative specific characteristics, and ϵ_j is a stochastic component.
We assume that the disturbances ϵ_j are independently and identically distributed across j and across individuals, following a type-I extreme value (Gumbel) distribution with CDF

$$F_\epsilon(\epsilon) = \exp\bigl(-\exp(-\epsilon)\bigr). \tag{2}$$

Under this assumption, the CDF of U_j is

$$F_{U_j}(u) = \Pr(U_j \le u) = \Pr(\epsilon_j \le u - V_j) = \exp\bigl(-\exp(-(u - V_j))\bigr), \tag{3}$$

that is, U_j is Gumbel with location parameter V_j and unit scale.
We are interested in the probability that two specified alternatives, say s and t, are jointly among the two most preferred alternatives: that is, the event that both U_s and U_t exceed the utilities of all remaining alternatives.
Define

$$U^* \equiv \max_{j \notin \{s,t\}} U_j.$$

The event of interest can be written as

$$E_{s,t} = \{U_s > U^*, \, U_t > U^*\}. \tag{4}$$

Equivalently,

$$E_{s,t} = \bigl\{\min(U_s, U_t) > U^*\bigr\}.$$

To use inclusion–exclusion arguments, define

$$A = \{U_s \le U^*\}, \qquad B = \{U_t \le U^*\}.$$

Then $E_{s,t} = A^c \cap B^c$ and therefore

$$E_{s,t}^c = A \cup B, \tag{5}$$

$$P(E_{s,t}) = 1 - P(A \cup B) = 1 - \bigl[P(A) + P(B) - P(A \cap B)\bigr], \tag{6}$$

where

$$P(A) = P(U_s \le U^*), \quad P(B) = P(U_t \le U^*), \quad P(A \cap B) = P\bigl(\max\{U_s, U_t\} \le U^*\bigr).$$

1.1 Single choice made

In the case where only one option is chosen (e.g. "None" or "Don't know"), we need only the simple single-choice probability for alternative 1:

$$P\Bigl(U_1 > \max_{j=2,\ldots,J} U_j\Bigr), \tag{7}$$

which will be derived explicitly in Section 1.3.3 below.

1.2 Properties of the Gumbel distribution

If the random variable ϵ is distributed under a general Gumbel distribution, then its CDF can be written as

$$F(\epsilon) = \exp\big(-\exp(-\mu(\epsilon - \eta))\big), \tag{8}$$

where η is a location parameter and $\mu > 0$ is a scale parameter. The distribution has the following properties:

1. The mode is η.
2. The mean is $\eta + \gamma/\mu$, where γ is Euler's constant.
3. The variance is $\pi^2/(6\mu^2)$.
4. *Affine transformation:* If ϵ is Gumbel distributed with parameters (η, μ) and $\alpha > 0$, V are scalar constants, then
$$\alpha\epsilon + V \sim \text{Gumbel}(\alpha\eta + V, \mu/\alpha).$$
5. *Difference of two Gumbels:* If ϵ_1 and ϵ_2 are independent Gumbel variables with parameters (η_1, μ) and (η_2, μ) respectively, then
$$\epsilon^\star = \epsilon_1 - \epsilon_2$$
is logistically distributed with location $\eta_1 - \eta_2$ and scale $1/\mu$. Its CDF is
$$F_{\epsilon^\star}(z) = \frac{1}{1 + \exp\big(-\mu(z - (\eta_1 - \eta_2))\big)}. \tag{9}$$
Equivalently,
$$F_{\epsilon^\star}(z) = \frac{1}{1 + \exp\big(\mu(\eta_2 - \eta_1 - z)\big)}.$$
6. *Max of Gumbels:* If $(\epsilon_1, \ldots, \epsilon_J)$ are independent Gumbel random variables with parameters (η_j, μ) respectively, then
$$\max(\epsilon_1, \ldots, \epsilon_J)$$
is Gumbel distributed with parameters
$$\left(\frac{1}{\mu} \ln \sum_{j=1}^{J} \exp(\mu\eta_j),\ \mu \right). \tag{10}$$

The mean of ϵ_j is not identified if V_j contains an intercept. We can therefore, without loss of generality, impose that $\eta_j = 0$ for all j.
More generally, the overall *scale* of utility is not identified: if $\{U_j\}$ and $\{cU_j\}$, for any $c > 0$, are both feasible, they generate the same choice probabilities. Consequently, only the ratios of coefficients to the error scale are identified. A common normalisation is to fix the Gumbel scale at $\mu = 1$ and set the location parameters to zero, $\eta_j = 0$ for all j. Under this normalisation,

$$\epsilon_j \sim \text{Gumbel}(0, 1), \qquad U_j = V_j + \epsilon_j.$$

With this normalisation, the difference of two independent Gumbels simplifies to a standard logistic distribution. In particular, equation (9) reduces to

$$F_{\epsilon^\star}(z) = \frac{1}{1 + \exp(-z)}, \tag{9a}$$

i.e. $\epsilon^\star$ is logistic with unit scale and zero location.

1.3 Deriving the multichoice logit model

Under the normalisation above, we have $\epsilon_j \sim \text{Gumbel}(0, 1)$ and therefore

$$U_j = V_j + \epsilon_j \sim \text{Gumbel}(V_j, 1).$$

1.3.1 Distributions of maxima and differences

Fix two distinct alternatives s and t. Define

$$U^* = \max_{j \notin \{s,t\}} U_j.$$

Using property 6, U^* is also Gumbel distributed:

$$U^* \sim \text{Gumbel}\Big(\ln \sum_{j \notin \{s,t\}} e^{V_j},\, 1\Big). \tag{11}$$

Similarly, the maximum of U_s and U_t,

$$U_{st}^{\max} \equiv \max\{U_s, U_t\},$$

is Gumbel distributed with parameters

$$U_{st}^{\max} \sim \text{Gumbel}\Big(\ln\big(e^{V_s} + e^{V_t}\big),\, 1\Big). \tag{12}$$

Now define the differences

$$Z_s = U^* - U_s, \qquad Z_t = U^* - U_t, \qquad Z_{st} = U^* - U_{st}^{\max}. \tag{13}$$

By property 5, each of these is logistic, with:

- Z_s having location
$$\delta_s = \ln \sum_{j \notin \{s,t\}} e^{V_j} - V_s,$$
- Z_t having location
$$\delta_t = \ln \sum_{j \notin \{s,t\}} e^{V_j} - V_t,$$
- Z_{st} having location
$$\delta_{st} = \ln \sum_{j \notin \{s,t\}} e^{V_j} - \ln\big(e^{V_s} + e^{V_t}\big).$$

In each case, the scale is 1, so the CDF is

$$F_Z(z) = \frac{1}{1 + \exp\big(-(z - \delta)\big)}.$$

1.3.2 Computing the pairwise probability

Introduce the shorthand

$$a_j = e^{V_j}, \qquad R \equiv \sum_{j \notin \{s,t\}} a_j.$$

From the previous subsection,

$$\delta_s = \ln R - V_s \quad \Rightarrow \quad \exp(\delta_s) = \frac{R}{a_s}.$$

Therefore

$$F_{Z_s}(0) = \frac{1}{1 + \exp(\delta_s)} = \frac{1}{1 + R/a_s} = \frac{a_s}{a_s + R}.$$

Since $Z_s = U^* - U_s$, we have

$$P(U^* \le U_s) = P(Z_s \le 0) = \frac{a_s}{a_s + R}. \tag{14}$$

Hence

$$P(U_s \le U^*) = 1 - P(U^* \le U_s) = \frac{R}{a_s + R}. \tag{15}$$

By symmetry,

$$P(U_t \le U^*) = \frac{R}{a_t + R}. \tag{16}$$

For the joint event in (6), note that

$$\max\{U_s, U_t\} \le U^* \iff U_{st}^{\max} \le U^*.$$

Using $Z_{st} = U^* - U_{st}^{\max}$ and the same logistic argument,

$$F_{Z_{st}}(0) = \frac{1}{1+\exp(\delta_{st})} = \frac{1}{1+R/(a_s+a_t)} = \frac{a_s+a_t}{a_s+a_t+R},$$

so

$$P\big(\max\{U_s, U_t\} \le U^*\big) = P(U_{st}^{\max} \le U^*) = 1 - F_{Z_{st}}(0) = \frac{R}{a_s+a_t+R}. \tag{17}$$

Substituting (15), (16), and (17) into the inclusion–exclusion identity (6), we obtain

$$\begin{aligned} P(E_{s,t}) &= 1 - \left[\frac{R}{a_s+R} + \frac{R}{a_t+R} - \frac{R}{a_s+a_t+R}\right] \\ &= 1 - \frac{R}{a_s+R} - \frac{R}{a_t+R} + \frac{R}{a_s+a_t+R}. \end{aligned} \tag{18}$$

Equivalently, using $a_s/(a_s+R) = 1 - R/(a_s+R)$ and similarly for t and (s,t), we can write the same probability as

$$P(E_{s,t}) = \frac{a_s}{a_s+R} + \frac{a_t}{a_t+R} - \frac{a_s+a_t}{a_s+a_t+R}. \tag{19}$$

Both forms (18) and (19) are algebraically equivalent; (18) is close to the inclusion–exclusion expression, while (19) is more directly interpretable in logit terms.

1.3.3 Single-choice probability

Returning to the single-choice case (7), define

$$U_1^* = \max_{j=2,\ldots,J} U_j,$$

and consider

$$E_1 = \{U_1 > U_1^*\}.$$

As before, define $Z_1 = U_1^* - U_1$. Using the same difference-of-Gumbels logic, Z_1 is logistic with location

$$\delta_1 = \ln \sum_{j=2}^{J} e^{V_j} - V_1.$$

Thus

$$P(U_1^* \le U_1) = P(Z_1 \le 0) = \frac{1}{1+\exp(\delta_1)} = \frac{e^{V_1}}{e^{V_1} + \sum_{j=2}^{J} e^{V_j}}.$$

Since $U_1^* = \max_{j=2,\ldots,J} U_j$, this is exactly

$$P\Big(U_1 > \max_{j=2,\ldots,J} U_j\Big).$$

Therefore, the correct single-choice logit probability is

$$P\Big(U_1 > \max_{j=2,\ldots,J} U_j\Big) = \frac{e^{V_1}}{e^{V_1} + \sum_{j=2}^{J} e^{V_j}} = \frac{e^{V_1}}{\sum_{j=1}^{J} e^{V_j}}. \tag{7a}$$

The complementary probability, that some other alternative has utility at least as large as U_1, is

$$P\Big(\max_{j=2,\ldots,J} U_j \ge U_1\Big) = \frac{\sum_{j=2}^{J} e^{V_j}}{\sum_{j=1}^{J} e^{V_j}}.$$

In the single-choice part of the likelihood we must use the former (with e^{V_s} in the numerator), not the latter.

1.4 The likelihood function

Assume the deterministic part of utility for alternative j is a linear function of parameters,

$$V_{ij} = X_i' \beta_j, \tag{20}$$

where X_i is a k-vector of attributes for individual i, and β_j is a k-vector of alternative-specific parameters. We now construct the likelihood for the case where respondents may either choose a *pair* of alternatives (dual choice) or a *single* alternative.

1.4.1 Probabilities

Define

$$a_{ij} = \exp(V_{ij}), \qquad R_{i,s,t} = \sum_{j \notin \{s,t\}} a_{ij}.$$

For an observation i that chooses the unordered pair (s, t), the probability is

$$P_{i,s,t} = 1 - \frac{R_{i,s,t}}{a_{is} + R_{i,s,t}} - \frac{R_{i,s,t}}{a_{it} + R_{i,s,t}} + \frac{R_{i,s,t}}{a_{is} + a_{it} + R_{i,s,t}}. \tag{21}$$

Equivalently,

$$P_{i,s,t} = \frac{a_{is}}{a_{is} + R_{i,s,t}} + \frac{a_{it}}{a_{it} + R_{i,s,t}} - \frac{a_{is} + a_{it}}{a_{is} + a_{it} + R_{i,s,t}}. \tag{22}$$

For an observation i that chooses a single alternative s (e.g. "None" or "Don't know" or any other single choice), the probability is the usual MNL probability

$$P^{(1)}_{i,s} = \frac{a_{is}}{\sum_{j=1}^{J} a_{ij}} = \frac{\exp(V_{is})}{\sum_{j=1}^{J} \exp(V_{ij})}. \tag{23}$$

1.4.2 Indicators and likelihood

Define indicator variables for dual and single choices:

- For each individual i and unordered pair (s, t) with $1 \leq s < t \leq J$, let $d_{i,s,t} = 1$ if individual i chooses the pair (s, t) and 0 otherwise.
- For each individual i and alternative s, let $d_{i,s} = 1$ if individual i chooses s as a single alternative and 0 otherwise.

For the full dataset of N observations (allowing both dual and single choices), the contribution of observation i to the likelihood is

$$L_i(\beta) = \prod_{1 \leq s < t \leq J} P_{i,s,t}^{d_{i,s,t}} \prod_{s=1}^{J} \left(P^{(1)}_{i,s}\right)^{d_{i,s}}.$$

The full likelihood is

$$L(\beta) = \prod_{i=1}^{N} L_i(\beta) = \prod_{i=1}^{N} \left[\prod_{1 \leq s < t \leq J} P_{i,s,t}^{d_{i,s,t}} \prod_{s=1}^{J} \left(P^{(1)}_{i,s}\right)^{d_{i,s}} \right], \tag{24}$$

and the log-likelihood is

$$\ell(\beta) = \log L(\beta) = \sum_{i=1}^{N} \left[\sum_{1 \leq s < t \leq J} d_{i,s,t} \log P_{i,s,t} + \sum_{s=1}^{J} d_{i,s} \log P^{(1)}_{i,s} \right]. \tag{25}$$

1.4.3 Using dual choices only

In some applications we may wish to use only the subset of the data in which each agent chooses exactly two options. Let N_d be the number of such observations, and let $S_d = \{(s, t) : 1 \leq s < t \leq D\}$ denote the set of all unordered pairs for the D alternatives considered in this reduced sample.

For this dual-choice-only sample, the log-likelihood simplifies to

$$\ell(\beta) = \sum_{i=1}^{N_d} \sum_{(s,t) \in S_d} d_{i,s,t} \, \log P_{i,s,t}, \tag{26}$$

with $P_{i,s,t}$ given by (22).

1.5 Maximum Likelihood Estimation

The maximum likelihood estimator is obtained by maximising the log-likelihood with respect to β. For the dual-choice-only case, using (26), we have

$$\hat{\beta} = \arg\max_{\beta} \ell(\beta) = \arg\max_{\beta} \sum_{i=1}^{N_d} \sum_{(s,t) \in S_d} d_{i,s,t} \, \log P_{i,s,t}, \tag{27}$$

where $P_{i,s,t}$ is defined in (22). For the full sample, we use the log-likelihood in (25), with the restriction that, for each i,

$$\sum_{1 \leq s < t \leq J} d_{i,s,t} + \sum_{s=1}^{J} d_{i,s} = 1. \tag{28}$$

1.6 Jacobian (gradient of the log-likelihood)

We derive the gradient (Jacobian) of the log-likelihood with respect to the parameters. It is informative to first obtain the derivatives with respect to the deterministic utilities V_{ij} and then apply the chain rule. Consider a single observation i that chooses the unordered pair (s,t). Its log-likelihood contribution is

$$\ell_i = \log P_{i,s,t}.$$

Recall the notation

$$a_{ij} = e^{V_{ij}}, \qquad R_{i,s,t} = \sum_{j \notin \{s,t\}} a_{ij}.$$

For simplicity in this subsection, we suppress the observation index i in the notation and write a_j, R, $P_{s,t}$ instead of a_{ij}, $R_{i,s,t}$, $P_{i,s,t}$.
From (22),

$$P_{s,t} = \frac{a_s}{a_s+R} + \frac{a_t}{a_t+R} - \frac{a_s+a_t}{a_s+a_t+R}.$$

We now compute $\partial P_{s,t}/\partial V_j$ for $j=s$, $j=t$, and $j \notin \{s,t\}$.

Derivative with respect to V_s. Using $a_s = e^{V_s}$ and $R = \sum_{j \notin \{s,t\}} a_j$, we have

$$\frac{\partial a_s}{\partial V_s} = a_s, \qquad \frac{\partial R}{\partial V_s} = 0.$$

Only the first and third terms in $P_{s,t}$ depend on a_s. Differentiating,

$$\frac{\partial}{\partial V_s}\left(\frac{a_s}{a_s+R}\right) = \frac{(a_s+R)\,a_s - a_s\,a_s}{(a_s+R)^2} = \frac{a_s R}{(a_s+R)^2},$$
$$\frac{\partial}{\partial V_s}\left(\frac{a_s+a_t}{a_s+a_t+R}\right) = \frac{(a_s+a_t+R)\,a_s - (a_s+a_t)\,a_s}{(a_s+a_t+R)^2} = \frac{a_s R}{(a_s+a_t+R)^2}.$$

Therefore

$$\frac{\partial P_{s,t}}{\partial V_s} = a_s R\left[\frac{1}{(a_s+R)^2} - \frac{1}{(a_s+a_t+R)^2}\right]. \tag{29}$$

Derivative with respect to V_t. By symmetry,

$$\frac{\partial P_{s,t}}{\partial V_t} = a_t R\left[\frac{1}{(a_t+R)^2} - \frac{1}{(a_s+a_t+R)^2}\right]. \tag{30}$$

Derivative with respect to V_r for $r \notin \{s,t\}$. For $r \notin \{s,t\}$, the dependence is via R only. Using

$$\frac{\partial a_r}{\partial V_r} = a_r, \qquad \frac{\partial R}{\partial V_r} = a_r,$$

and differentiating each term in $P_{s,t}$ with respect to R, we obtain

$$\frac{\partial}{\partial R}\left(\frac{a_s}{a_s+R}\right) = -\frac{a_s}{(a_s+R)^2},$$
$$\frac{\partial}{\partial R}\left(\frac{a_t}{a_t+R}\right) = -\frac{a_t}{(a_t+R)^2},$$
$$\frac{\partial}{\partial R}\left(\frac{a_s+a_t}{a_s+a_t+R}\right) = -\frac{a_s+a_t}{(a_s+a_t+R)^2}.$$

Therefore,

$$\frac{\partial P_{s,t}}{\partial V_r} = a_r\left[-\frac{a_s}{(a_s+R)^2} - \frac{a_t}{(a_t+R)^2} + \frac{a_s+a_t}{(a_s+a_t+R)^2}\right], \qquad r \notin \{s,t\}. \tag{31}$$

Log-likelihood derivatives with respect to V_j. The derivative of the log-likelihood contribution from observation i is

$$\frac{\partial \ell_i}{\partial V_j} = \frac{1}{P_{s,t}}\frac{\partial P_{s,t}}{\partial V_j},$$

for $j = 1,\ldots,J$, using the appropriate expression from (29), (30), or (31) depending on whether $j=s$, $j=t$, or $j \notin \{s,t\}$.
Importantly, even for $j \notin \{s,t\}$, the derivative $\partial \ell_i/\partial V_j$ is generally non-zero because the probability $P_{s,t}$ depends on all alternatives via the sum R.

From utilities to parameters. With the specification

$$V_{ij} = X_i' \beta_j,$$

we have

$$\frac{\partial V_{ij}}{\partial \beta_j} = X_i, \qquad \frac{\partial V_{ij}}{\partial \beta_{j'}} = 0 \quad \text{for } j' \neq j.$$

Thus, for each alternative j,

$$\frac{\partial \ell(\beta)}{\partial \beta_j} = \sum_{i=1}^{N_d} \sum_{(s,t)\in S_d} d_{i,s,t} \frac{1}{P_{i,s,t}} \frac{\partial P_{i,s,t}}{\partial V_{ij}} X_i, \tag{32}$$

where $\partial P_{i,s,t}/\partial V_{ij}$ is given by the appropriate version of (29), (30), or (31), with $a_{ij} = \exp(V_{ij})$ and $R_{i,s,t} = \sum_{k \notin \{s,t\}} a_{ik}$.
Equation (32) defines the Jacobian (gradient) of the log-likelihood with respect to the parameter vectors β_j. In practice, this gradient can be used with standard numerical optimisation algorithms (e.g. BFGS) to obtain the maximum likelihood estimates.

1.7 Issues and extensions: single vs. dual choices

The model derived above naturally applies to the case where each respondent chooses exactly two options. When respondents may instead choose a single option (or some combination of single and dual choices within the dataset), care must be taken to ensure that the probabilities of all possible response patterns sum to one. Simply adding single-choice probabilities (23) and dual-choice probabilities (22) side by side, without a joint model of the process determining whether a respondent chooses one or two options, will generally lead to probabilities that are not mutually exclusive and do not sum to one.
A principled approach is to model a two-stage process:

1. A selection mechanism (e.g. a binary or ordered choice model) that determines whether the respondent chooses one option or two (possibly depending on the utilities V_{ij} or additional covariates), and

2. Conditional on choosing one or two, a logit model for the choice over alternatives (single-choice or dual-choice probabilities as derived above).

This is conceptually similar to a Heckman-type selection model or a two-part model, and it ensures that the joint probabilities over the full space of outcomes (single vs. dual choices and which alternative(s) are chosen) are properly normalised.

References

Ben-Akiva, M. E., Lerman, S. R. and Lerman, S. R. (1985). *Discrete Choice Analysis: Theory and Application to Travel Demand*, volume 9. MIT Press.

Ophem, H. V., Stam, P. and Praag, B. V. (1999). *Multichoice Logit: Modeling Incomplete Preference Rankings of Classical Concerts. Journal of Business & Economic Statistics*, 17(1), 117–128.