Estimation of multinomial logit models in R : The mlogit Packages

Yves Croissant
Université de la Réunion

Abstract

mlogit is a package for R which enables the estimation the multinomial logit models with individual and/or alternative specific variables. The main extensions of the basic multinomial model (heteroscedastic, nested and random parameter models) are implemented.

Keywords: discrete choice models, maximum likelihood estimation, R, econometrics.

An introductory example

The logit model is useful when one tries to explain discrete choices, *i.e.* choices of one among several mutually exclusive alternatives. There are many useful applications of discrete choice modelling in different fields of applied econometrics, using individual data, which may be :

– *revealed preferences data* which means that the data are observed choices of individuals for, say, a transport mode (car, plane and train for example),
– *stated preferences data* ; in this case, individuals face a virtual situation of choice, for example the choice between three train tickets with different characteristics :
  – A : a train ticket which costs 10 euros, for a trip of 30 minutes and one change,
  – B : a train ticket which costs 20 euros, for a trip of 20 minutes and no change,
  – C : a train ticket which costs 22 euros, for a trip of 22 minutes and one change.

Suppose that we can define an index of satisfaction V_j for each alternative which depends linearly on cost (x) and price (z) :

$$\begin{cases} U_1 &=& \alpha_1 + \beta x_1 + \gamma z_1 \\ U_2 &=& \alpha_2 + \beta x_2 + \gamma z_2 \\ U_3 &=& \alpha_3 + \beta x_3 + \gamma z_3 \end{cases}$$

In this case, the choice of alternative j is increasing with V_j. For sake of estimation, one has to transform V_j which can take any real value so that it belongs to the unit interval and can be interpreted as a probability. The multinomial logit model is obtained

by applying such a transformation to the V_js. More specifically, we have :

$$
\begin{cases}
P_1 &= \frac{e^{V_1}}{e^{V_1}+e^{V_2}+e^{V_3}} \\
P_2 &= \frac{e^{V_2}}{e^{V_1}+e^{V_2}+e^{V_3}} \\
P_3 &= \frac{e^{V_3}}{e^{V_1}+e^{V_2}+e^{V_3}}
\end{cases}
$$

The two characteristics of probabilities are satisfied :

– $0 \leq P_j \leq 1$,
– $\sum_{j=1}^{3} P_j = 1$

Once fitted, a logit model is useful for predictions :
– enter new values for the explanatory variables,
– get
    – at an individual level the probabilities of choice,
    – at an aggregate level the market shares.

Consider, as an example interurban trips between two towns (Lyon and Paris). Suppose that there are three modes (car, plane and train) and that the characteristics of the modes and the market shares are as follow :

	price	time	share
car	50	4	20%
plane	150	1	25%
train	80	2	55%

With a sample of travellers, one can estimate the coefficients of the logit model, *i.e.* the coefficients of time and price in the utility function.

The fitted model can then be used to predict the impact of some chocks on the market shares, for example :

– the influence of train trips length on modal shares,
– the influence of the arrival of low cost companies.

To get the predictions, one just has to change the values of train time or plane price and compute the new probabilities, which can be interpreted at the aggregate level as predicted market shares.

1. Data management and model description

1.1. Data management

mlogit is loaded using :

```
R> library("mlogit")
```

It comes with several data sets that we'll use to illustrate the features of the library. Data sets used for multinomial logit estimation deals with some individuals, that make one or a sequential choice of one alternative among a set of several alternatives. The determinants of these choices are variables that can be alternative specific or purely individual specific. Such data have therefore a specific structure that can be characterised by three indexes :

– the alternative,
– the choice situation,
– the individual

the last one being only relevant if we have repeated observations for the same individual.

Data sets can have two different shapes :

– a *wide* shape : in this case, there is one row for each choice situation,
– a *long* shape : in this case, there is one row for each alternative and, therefore, as many rows as there are alternatives for one each choice situation

This can be illustrated with two data sets. The first one, Fishing comes with mlogit. The second one TravelMode is from the AER package.

```
R> data("Fishing", package = "mlogit")
R> head(Fishing, 3)

     mode price.beach price.pier price.boat price.charter catch.beach
1 charter     157.930    157.930    157.930       182.930      0.0678
2 charter      15.114     15.114     10.534        34.534      0.1049
3    boat     161.874    161.874     24.334        59.334      0.5333
  catch.pier catch.boat catch.charter    income
1     0.0503     0.2601        0.5391 7083.332
2     0.0451     0.1574        0.4671 1250.000
3     0.4522     0.2413        1.0266 3750.000
```

There are four fishing modes (beach, pier, boat, charter), two alternative specific variables (price and catch) and one choice/individual specific variable (income)[1]. This "wide" format is suitable to store individual specific variables. Otherwise, it is cumbersome for alternative specific variables because there are as many columns for such variables that there are alternatives.

```
R> data("TravelMode", package = "AER")
R> head(TravelMode)

  individual  mode choice wait vcost travel gcost income size
1          1   air     no   69    59    100    70     35    1
2          1 train     no   34    31    372    71     35    1
```

---

1. Note that the distinction between choice situation and individual is not relevant here as these data are not panel data.

```
3        1   bus    no   35   25   417   70   35   1
4        1   car   yes    0   10   180   30   35   1
5        2   air    no   64   58    68   68   30   2
6        2 train    no   44   31   354   84   30   2
```

There are four transport modes (air, train, bus and car)and most of the variable are alternative specific (wait, vcost, travel, gcost). The only individual specific variables are income and size. This advantage of this shape is that there are much fewer columns than in the wide format, the caveat being that values of income and size are repeated four times.

mlogit deals with both format. It provides a mlogit.data function that take as first argument a `data.frame` and returns a `data.frame` in "long" format with some information about the structure of the data.

For the Fishing data, we would use :

```
R> Fish <- mlogit.data(Fishing, shape = "wide", varying = 2:9, choice = "mode")
```

The mandatory arguments are `choice` , which is the variable that indicates the choice made, the shape of the original `data.frame` and, if there are some alternative specific variables, `varying` which is a numeric vector that indicates which columns contains alternative specific variables. This argument is then passed to reshape that coerced the original `data.frame` in "long" format. Further arguments may be passed to reshape . For example, if the names of the variables are of the form var :alt, one can add **sep = ':'** .

```
R> head(Fish, 5)

            mode    income      alt    price  catch chid
1.beach    FALSE 7083.332    beach  157.930 0.0678    1
1.boat     FALSE 7083.332     boat  157.930 0.2601    1
1.charter   TRUE 7083.332  charter  182.930 0.5391    1
1.pier     FALSE 7083.332     pier  157.930 0.0503    1
2.beach    FALSE 1250.000    beach   15.114 0.1049    2
```

```
R> head(attr(Fish, "index"), 5)

           chid      alt
1.beach       1    beach
1.boat        1     boat
1.charter     1  charter
1.pier        1     pier
2.beach       2    beach
```

The result is a `data.frame` in "long format" with one line for each alternative. The "choice" variable is now a logical variable and the individual specific variable (income) is repeated 4 times. An index attribute is added to the data, which contains the two relevant index : `chid` is the choice index and `alt` index.

For data in "long" format like TravelMode , the `shape` (here equal to `long`) and the `choice` arguments are still mandatory.

The information about the structure of the data can be explicitly indicated or, in part, guessed by the mlogit.data function. Here, we have 210 individuals which are indicated by a variable called individual. The information about individuals can also be guessed from the fact that the data frame is balanced (every individual face 4 alternatives) and that the rows are ordered first by individual and then by alternative.

Concerning the alternative, there are indicated by the mode variable and they can also be guessed thanks to the ordering of the rows and the fact that the data frame is balanced.

The first way to read correctly this data frame is to ignore completely the two index variables. In this case, the only supplementary argument to provide is the `alt.levels` argument which is a character vector that contains the name of the alternatives :

```
R> TM <- mlogit.data(TravelMode, choice = "choice", shape = "long",
+     alt.levels = c("air", "train", "bus", "car"))
```

It is also possible to provide an argument `alt.var` which indicates the name of the variable that contains the alternatives

```
R> TM <- mlogit.data(TravelMode, choice = "choice", shape = "long",
+     alt.var = "mode")
```

The name of the variable that contains the information about the choice can be indicated using the `chid.var` argument :

```
R> TM <- mlogit.data(TravelMode, choice = "choice", shape = "long",
+     chid.var = "individual", alt.levels = c("air", "train", "bus",
+         "car"))
```

Both alternative and choice variable can be provided :

```
R> TM <- mlogit.data(TravelMode, choice = "choice", shape = "long",
+     chid.var = "individual", alt.var = "mode")
```

and dropped from the data using the `drop.index` argument :

```
R> TM <- mlogit.data(TravelMode, choice = "choice", shape = "long",
+     chid.var = "individual", alt.var = "mode", drop.index = TRUE)
R> head(TM)
```

```
        choice wait vcost travel gcost income size
1.air    FALSE   69    59    100    70     35    1
1.train  FALSE   34    31    372    71     35    1
1.bus    FALSE   35    25    417    70     35    1
1.car     TRUE    0    10    180    30     35    1
2.air    FALSE   64    58     68    68     30    2
2.train  FALSE   44    31    354    84     30    2
```

The final example is a data set called Train which contains data from a stated preference study.

```
R> data("Train", package = "mlogit")
R> head(Train, 3)
```

```
  id choiceid  choice price1 time1 change1 comfort1 price2 time2 change2
1  1        1 choice1   2400   150       0        1   4000   150       0
2  1        2 choice1   2400   150       0        1   3200   130       0
3  1        3 choice1   2400   115       0        1   4000   115       0
  comfort2
1        1
2        1
3        0
```

These data are panel data, each individual has responded to several (up to 16) scenario. To take this panel dimension into account, one has to add an argument `id` which contains the individual variable. The index attribute has now a supplementary column, the individual index.

```
R> Tr <- mlogit.data(Train, shape = "wide", choice = "choice", varying = 4:11,
+      sep = "", alt.levels = c(1, 2), id = "id")
R> head(Tr, 3)
```

```
    id choiceid choice alt price time change comfort chid
1.1  1        1   TRUE   1  2400  150      0       1    1
1.2  1        1  FALSE   2  4000  150      0       1    1
2.1  1        2   TRUE   1  2400  150      0       1    2
```

```
R> head(attr(Tr, "index"), 3)
```

```
    chid alt id
1.1    1   1  1
1.2    1   2  1
2.1    2   1  1
```

1.2. Model description

mlogit use the standard `formula` , `data` interface to describe the model to be estimated. However, standard `formula`s are not very practical for such models. More precisely, when working with multinomial logit models, one has to consider three kinds of variables :

– alternative specific variables x_{ij} with a generic coefficient β,
– individual specific variables z_i with alternative specific coefficients γ_j,
– alternative specific variables w_{ij} with an alternative specific coefficient δ_j.

The satisfaction index for the alternative j is then :

$$V_{ij} = \alpha_j + \beta x_{ij} + \gamma_j z_i + \delta_j w_{ij}$$

Satisfaction being ordinal, only differences are relevant to modelize the choice for one alternative. This means that, for example, we'll be interested in the difference between the satisfaction index of two different alternatives j and k :

$$V_{ij} - V_{ik} = (\alpha_j - \alpha_k) + \beta(x_{ij} - x_{ik}) + (\gamma_j - \gamma_k)z_i + (\delta_j w_{ij} - \delta_k w_{ik})$$

It is clear from the previous expression that coefficients for individual specific variables (the intercept being one of those) should be alternative specific, otherwise they would disappear in the differentiation. Moreover, only differences of these coefficients are relevant and may be identified. For example, with three alternatives 1, 2 and 3, the three coefficients $\gamma_1, \gamma_2, \gamma_3$ associated to an individual specific variable cannot be identified, but only two linear combinations of them. Therefore, one has to make a choice of normalization and the most simple one is just to put $\gamma_1 = 0$.

Coefficients for alternative specific variables may (or may not) be alternative specific. For example, transport time is alternative specific, but may be 10 mn in public transport don't have the same impact on utility than 10 mn in a car. In this case, alternative specific coefficients are relevant. Monetary time is also alternative specific, but in this case, one can consider than 1 euro is 1 euro whatever it is spent in car or in public transports[2]. In this case, a generic coefficient is relevant.

A model with only individual specific variables is sometimes called a *multinomial logit model*, one with only alternative specific variables a *conditional logit model* and one with both kind of variables a *mixed logit model*. This is seriously misleading : *conditional logit model* is also a logit model for longitudinal data in the statistical literature and *mixed logit* is one of the names of a logit model with random parameters. Therefore, in what follow, we'll use the name *multinomial logit model* for the model we've just described whatever the nature of the explanatory variables included in the model.

mlogit package provides objects of class `mFormula` which are extended model formulas and which are build upon `Formula` objects provided by the **Formula** package[3].

---

2. At least if the monetary cost of using car is correctly calculated.
3. See Zeileis and Croissant (2010) for a description of the **Formula** package.

To illustrate the use of mFormula objects, let's use again the TravelMode data set. income and size (the size of the household) are individual specific variables. vcost (monetary cost) and travel (travel time) are alternative specific. We want to use a generic coefficient for the former and alternative specific coefficients for the latter. This is done using the mFormula function that build a three-parts formula :

```
R> f <- mFormula(choice ~ vcost | income + size | travel)
```

By default, an intercept is added to the model, it can be removed by using `+0` or `-1` in the second part. Some parts may be omitted when there are no ambiguity. For example, the following couples of `formula`s are identical :

```
R> f2 <- mFormula(choice ~ vcost + travel | income + size)
R> f2 <- mFormula(choice ~ vcost + travel | income + size | 0)

R> f3 <- mFormula(choice ~ 0 | income | 0)
R> f3 <- mFormula(choice ~ 0 | income)

R> f4 <- mFormula(choice ~ vcost + travel)
R> f4 <- mFormula(choice ~ vcost + travel | 1)
R> f4 <- mFormula(choice ~ vcost + travel | 1 | 0)
```

Finally, we show below some `formula`s that describe models without intercepts (which is generally hardly relevant)

```
R> f5 <- mFormula(choice ~ vcost | 0 | travel)
R> f6 <- mFormula(choice ~ vcost | income + 0 | travel)
R> f6 <- mFormula(choice ~ vcost | income - 1 | travel)
R> f7 <- mFormula(choice ~ 0 | income - 1 | travel)
```

model.matrix and model.frame methods are provided for mFormula objects. The former is of particular interest, as illustrated in the following example :

```
R> f <- mFormula(choice ~ vcost | income | travel)
R> head(model.matrix(f, TM))
```

	alttrain	altbus	altcar	vcost	alttrain:income	altbus:income
1.air	0	0	0	59	0	0
1.train	1	0	0	31	35	0
1.bus	0	1	0	25	0	35
1.car	0	0	1	10	0	0
2.air	0	0	0	58	0	0
2.train	1	0	0	31	30	0

	altcar:income	altair:travel	alttrain:travel	altbus:travel	altcar:travel
1.air	0	100	0	0	0
1.train	0	0	372	0	0
1.bus	0	0	0	417	0
1.car	35	0	0	0	180
2.air	0	68	0	0	0
2.train	0	0	354	0	0

The model matrix contains $J-1$ columns for every individual specific variables (income and the intercept), which means that the coefficient associated to the first alternative (air) is fixed to 0.

It contains only one column for vcost because we want a generic coefficient for this variable.

It contains J columns for travel, because it is an alternative specific variable for which we want an alternative specific coefficient.

2. Random utility model and the multinomial logit model

2.1. Random utility model

The individual must choose one alternative among J different and exclusive alternatives. A level of utility may be defined for each alternative and the individual is supposed to choose the alternative with the highest level of utility. Utility is supposed to be the sum of two components[4] :

– a systematic component, denoted V_j, which is a function of different observed variables x_j. For sake of simplicity, it will be supposed that this component is a linear combination of the observed explanatory variables : $V_j = \beta_j^\top x_j$,

– an unobserved component ϵ_j which, from the researcher point of view, can be represented as a random variable. This error term includes the impact of all the unobserved variables which have an impact on the utility of choosing a specific alternative.

It is very important to understand that the utility and therefore the choice is purely deterministic from the decision maker's point of view. It is random form the searcher's point of view, because some of the determinants of the utility are unobserved, which implies that the choice can only be analyzed in terms of probabilities.

We have, for each alternative, the following utility levels :

$$
\begin{cases}
U_1 &= \beta_1^\top x_1 + \epsilon_1 &= V_1 + \epsilon_1 \\
U_2 &= \beta_2^\top x_2 + \epsilon_2 &= V_2 + \epsilon_2 \\
&\vdots & \vdots \\
U_J &= \beta_J^\top x_J + \epsilon_J &= V_J + \epsilon_J
\end{cases}
$$

---

4. when possible, we'll omit the individual index to simplify the notations.

alternative l will be chosen if and only if $\forall \ j \neq l \ \ U_j > U_l$ which leads to the following $J - 1$ conditions :

$$\begin{cases} U_l - U_1 &= (V_l - V_1) + (\epsilon_l - \epsilon_1) > 0 \\ U_l - U_2 &= (V_l - V_2) + (\epsilon_l - \epsilon_2) > 0 \\ &\vdots \\ U_l - U_J &= (V_l - V_J) + (\epsilon_l - \epsilon_J) > 0 \end{cases}$$

As ϵ_j are not observed, choices can only be modeled in terms of probabilities from the researcher point of view. The $J - 1$ conditions can be rewritten in terms of upper bonds for the $J - 1$ remaining error terms :

$$\begin{cases} \epsilon_1 &< (V_l - V_1) + \epsilon_l \\ \epsilon_2 &< (V_l - V_2) + \epsilon_l \\ &\vdots \\ \epsilon_J &< (V_l - V_J) + \epsilon_l \end{cases}$$

The general expression of the probability of choosing alternative l is then :

$$(\mathrm{P}_l \mid \epsilon_l) = \mathrm{P}(U_l > U_1, \ldots, U_l > U_J)$$

$$(\mathrm{P}_l \mid \epsilon_l) = F_{-l}(\epsilon_1 < (V_l - V_1) + \epsilon_l, \ldots, \epsilon_J < (V_l - V_J) + \epsilon_l) \tag{1}$$

where F_{-l} is the multivariate distribution of $J - 1$ error terms (all the ϵ's except ϵ_l). Note that this probability is conditional on the value of ϵ_l.

The unconditional probability (which depends only on β and on the value of the observed explanatory variables is :

$$\mathrm{P}_l = \int (\mathrm{P}_l \mid \epsilon_l) f_l(\epsilon_l) d\epsilon_l$$

$$\mathrm{P}_l = \int F_{-l}((V_l - V_1) + \epsilon_l, \ldots, (V_l - V_J) + \epsilon_l) f_l(\epsilon_l) d\epsilon_l \tag{2}$$

where f_l is the marginal density function of ϵ_l.

2.2. The distribution of the error terms

The multinomial logit model (McFadden (1974)) is a special case of the model developed in the previous section. It relies on three hypothesis :

H1 : independence of errors

If the hypothesis of independence of errors is made, we have :

$$
\begin{cases}
\mathrm{P}(U_l > U_1) &=& F_1(V_l - V_1 + \epsilon_l) \\
\mathrm{P}(U_l > U_2) &=& F_2(V_l - V_2 + \epsilon_l) \\
&\vdots& \\
\mathrm{P}(U_l > U_J) &=& F_J(V_l - V_J + \epsilon_l)
\end{cases}
$$

where F_j is the cumulative density of ϵ_j.

The conditional (1) and unconditional (2) probabilities are then :

$$
(\mathrm{P}_l \mid \epsilon_l) = \prod_{j \neq l} F_j(V_l - V_j + \epsilon_l) \tag{3}
$$

$$
\mathrm{P}_l = \int \prod_{j \neq l} F_j(V_l - V_j + \epsilon_l) \; f_l(\epsilon_l) \; d\epsilon_l \tag{4}
$$

which means that the evaluation of only a one-dimensional integral is required to compute the probabilities.

H2 : Gumbel distribution

Each ϵ follows a GUMBEL distribution :

$$
f(z) = \frac{1}{\theta} e^{\frac{\mu - z}{\theta}} e^{-e^{\frac{\mu - z}{\theta}}}
$$

where μ is the location parameter and θ the scale parameter.

$$
P(z < t) = F(t) = \int_{-\infty}^{t} \frac{1}{\theta} e^{\frac{\mu - z}{\theta}} e^{-e^{\frac{\mu - z}{\theta}}} \, dz = e^{-e^{\frac{\mu - t}{\theta}}}
$$

The first two moments of the GUMBEL distribution are $\mathrm{E}(z) = \mu + \theta\gamma$, where γ is the Euler-Mascheroni constant (0.577) and $\mathrm{V}(z) = \frac{\pi^2}{6}\theta^2$.

The mean of ϵ_js is not identified if U_j contains an intercept. We can then, without loss of generality suppose that $\mu_j = 0 \quad \forall j$. Moreover, the overall scale of utility is not identified. Therefore, only $J - 1$ scale parameters may be identified, and a natural choice of normalisation is to impose that one of the θ_j is equal to 1.

H3 identically distributed errors

As, the location parameter is not identified for any error term, this hypothesis is essentially an homoscedasticity hypothesis, which means that the scale parameter of GUMBEL distribution is the same for all the alternatives. As one of them has been previously fixed to 1, we can therefore suppose that, without loss of generality, $\theta_j = 1 \; \forall j \in 1 \ldots J$ in case of homoscedasticity.

In this case, the conditional (3) and unconditional (4) probabilities further simplify to :

$$
(\mathrm{P}_l \mid \epsilon_l) = \prod_{j \neq l} F(V_l - V_j + \epsilon_l) \tag{5}
$$

$$P_l = \int \prod_{j \neq l} F(V_l - V_j + \epsilon_l) \, f(\epsilon_l) \, d\epsilon_l \tag{6}$$

with F and f respectively the cumulative and the density of the standard GUMBEL distribution (*i.e.* with position and scale parameters equal to 0 and 1).

2.3. Computation of the probabilities

With these hypothesis on the distribution of the error terms, we can now show that the probabilities have very simple, closed forms, which correspond to the logit transformation of the deterministic parts of the utility.

Let's start with the probability that the alternative l is better than one other alternative j. With hypothesis 2 and 3, it can be written :

$$P(\epsilon_j < V_l - V_j + \epsilon_l) = e^{-e^{-(V_l - V_j + \epsilon_l)}} \tag{7}$$

With hypothesis 1, the probability of choosing l is then simply the product of probabilities (7) for all the alternatives except l :

$$(P_l \mid \epsilon_l) = \prod_{j \neq l} e^{-e^{-(V_l - V_j + \epsilon_l)}} \tag{8}$$

The unconditional probability is the expected value of the previous expression with respect to ϵ_l.

$$P_l = \int_{-\infty}^{+\infty} (P_l \mid \epsilon_l) \, e^{-\epsilon_l} e^{-e^{-\epsilon_l}} d\epsilon_l = \int_{-\infty}^{+\infty} \left(\prod_{j \neq l} e^{-e^{-(V_l - V_j + \epsilon_l)}} \right) e^{-\epsilon_l} e^{-e^{-\epsilon_l}} d\epsilon_l \tag{9}$$

We first begin by writing the preceding expression for *all* alternatives, including the l alternative.

$$P_l = \int_{-\infty}^{+\infty} \left(\prod_{j} e^{-e^{-(V_l - V_j + \epsilon_l)}} \right) e^{-\epsilon_l} d\epsilon_l$$

$$P_l = \int_{-\infty}^{+\infty} e^{-\sum_j e^{-(V_l - V_j + \epsilon_l)}} e^{-\epsilon_l} d\epsilon_l = \int_{-\infty}^{+\infty} e^{-e^{-\epsilon_l} \sum_j e^{-(V_l - V_j)}} e^{-\epsilon_l} d\epsilon_l$$

We then use the following change of variable

$$t = e^{-\epsilon_l} \Rightarrow dt = -e^{-\epsilon_l} d\epsilon_l$$

The unconditional probability is therefore the following integral :

$$P_l = \int_0^{+\infty} e^{-t \sum_j e^{-(V_l - V_j)}} \, dt$$

which has a closed form :

$$P_l = \left[-\frac{e^{-t \sum_j e^{-(V_l - V_j)}}}{\sum_j e^{-(V_l - V_j)}} \right]_0^{+\infty} = \frac{1}{\sum_j e^{-(V_l - V_j)}}$$

and can be rewritten as the usual logit probability :

$$P_l = \frac{e^{V_i}}{\sum_j e^{V_j}} \tag{10}$$

2.4. IIA hypothesis

If we consider the probabilities of choice for two alternatives l and m, we have :

$$P_l = \frac{e^{V_l}}{\sum_j e^{V_j}}$$

$$P_m = \frac{e^{V_m}}{\sum_j e^{V_j}}$$

The ration of these two probabilities is :

$$\frac{P_l}{P_m} = \frac{e^{V_l}}{e^{V_m}}$$

This probability ratio for the two alternatives depends only on the characteristics of these two alternatives and not on those of other alternatives. This is called the IIA hypothesis (for independence of irrelevant alternatives).

If we use again the introductory example of urban trips between Lyon and Paris :

	price	time	share
car	50	4	20%
plane	150	1	20%
train	80	2	60%

Suppose that, because of low cost companies arrival, the price of plane is now 100$. The market share of plane will increase (for example up to 60%). With a logit model, share for train / share for car is 3 before the price change, and will remain the same after the price change. Therefore, the new predicted probabilities for car and train are 10 and 30%.

The *IIA* hypothesis relies on the hypothesis of independence of the error terms. It is not a problem by itself and may even be considered as a useful feature for a well specified model. However, this hypothesis may be in practice violated if some important variables are unobserved.

To see that, suppose that the utilities for two alternatives are :

$$U_{i1} = \alpha_1 + \beta_1 z_i + \gamma x_{i1} + \epsilon_{i1}$$

$$U_{i2} = \alpha_2 + \beta_2 z_i + \gamma x_{i2} + \epsilon_{i2}$$

with ϵ_{i1} and ϵ_{i2} uncorrelated. In this case, the logit model can be safely used, as the hypothesis of independence of the errors is satisfied.

If z_i is unobserved, the estimated model is :

$$U_{i1} = \alpha_1 + \gamma x_{i1} + \eta_{i1}$$

$$U_{i2} = \alpha_2 + \gamma x_{i2} + \eta_{i2}$$

$$\eta_{i1} = \epsilon_{i1} + \beta_1 z_i$$

$$\eta_{i2} = \epsilon_{i2} + \beta_2 z_i$$

The error terms are now correlated because of the common influence of omitted variables.

2.5. Estimation

The coefficients of the multinomial logit model are estimated by full information maximum likelihood.

The likelihood function

Let's start with a very simple example. Suppose there are four individuals. For given parameters and explanatory variables, we can calculate the probabilities. The likelihood for the sample is the probability associated to the sample :

	choice	P_{i1}	P_{i2}	P_{i3}	l_i
1	1	0.5	0.2	0.3	0.5
2	3	0.2	0.4	0.4	0.4
3	2	0.6	0.1	0.3	0.1
4	2	0.3	0.6	0.1	0.6

With random sample the joint probability for the sample is simply the product of the probabilities associated with every observation.

$$\mathrm{L} = 0.5 \times 0.4 \times 0.1 \times 0.6$$

A compact expression of the probabilities that enter the likelihood function is obtained by denoting y_{ij} a dummy variable which is equal to one if individual i made choice j and 0 otherwise.

The probability of the choice made for one individual is then :

$$\mathrm{P}_i = \prod_j \mathrm{P}_{ij}^{y_{ij}}$$

Or in log :

$$\ln \mathrm{P}_i = \sum_j y_{ij} \ln \mathrm{P}_{ij}$$

which leads to the log-likelihood function :

$$\ln L = \sum_i \ln \mathrm{P}_i = \sum_i \sum_j y_{ij} \ln \mathrm{P}_{ij}$$

Numerical optimization

We seek to calculate the maximum of a function $f(x)$. This first order condition for a maximum is $f'(x_o) = 0$, but in general, there is no explicit solution for x_o, which then must be numerically approximated.

1. Start with a value x called x_t,
2. Approximate the function around x_t using a second order TAYLOR series : $l(x) = f(x_t) + (x - x_t)g(x_t) + 0.5(x - x_t)^2 h(x_t)$ where g and h are the first two derivatives of f,
3. find the maximum of $l(x)$. The first order condition is : $\frac{\partial l(x)}{\partial x} = g(x_t) + (x - x_t)h(x_t) = 0$. The solution is : $x_t - \frac{g(x_t)}{h(x_t)}$
4. call this value x_{t+1} and iterate until you get as close as required to the maximum.

Consider now a function of several variables $f(x)$. The vector of first derivatives (called the gradient) is denoted g and the matrix of second derivatives (called the hessian) is denoted H. The second order approximation is :

$$l(x) = f(x_t) + (x - x_t)g(x_t) + 0.5(x - x_t)'H(x_t)(x - x_t)$$
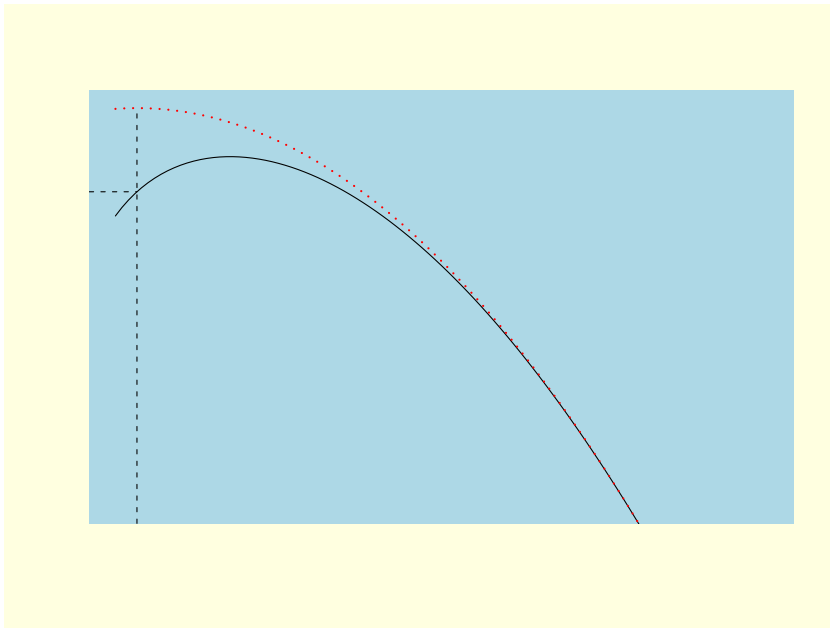
FIGURE 1 – Numerical optimization
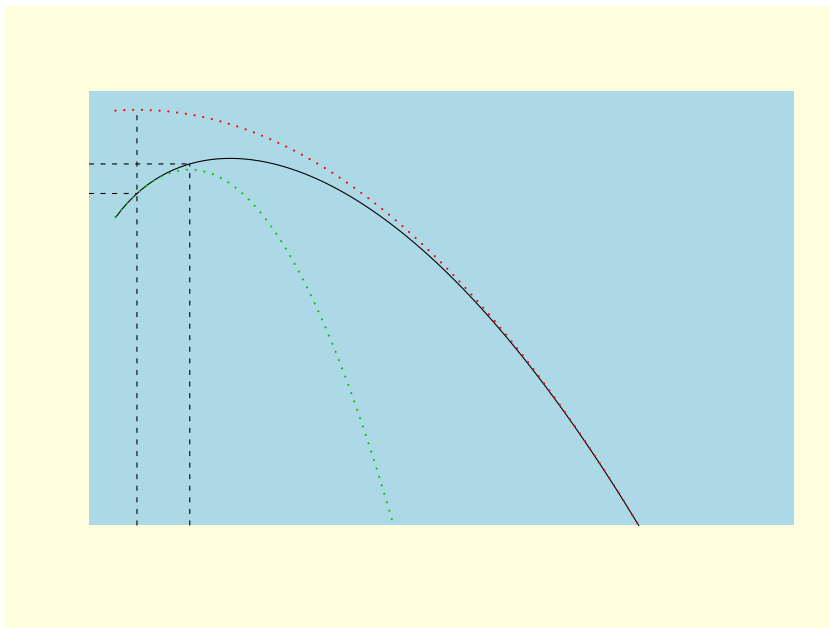
FIGURE 2 – Numerical optimization
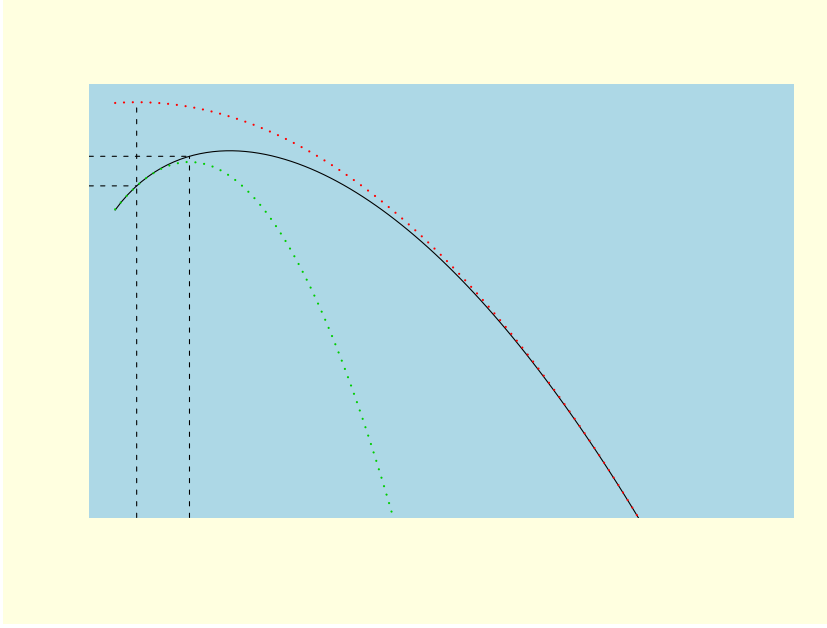
FIGURE 3 – Numerical optimization

The vector of first derivatives is :

$$\frac{\partial l(x)}{\partial x} = g(x_t) + H(x_t)(x - x_t)$$

$$x = x_t - H(x_t)^{-1}g(x_t)$$

Two kinds of routines are currently used for maximum likelihood estimation. The first one can be called "Newton-like" methods. In this case, at each iteration, an estimation of the hessian is calculated, whether using the second derivatives of the function (Newton-Ralphson method) or using the outer product of the gradient (BHHH). This approach is very powerful if the function is well-behaved, but it may performs poorly otherwise and scratch after a few iterations.

The second one, called BFGS, updates at each iteration the estimation of the hessian. It is often more robust and may performs well in cases where the first one doesn't work.

Two optimization functions are included in core R : nlm which use the Newton-Ralphson method and optim which use BFGS (among other methods). Recently, the **maxLik** package (Toomet, Henningsen, with contributions from Spencer Graves, and Croissant 2010) provides a unified approach. With a unique interface, all the previously described methods are available.

The behavior of maxLik can be controlled by the user using in the estimation function arguments like `print.level` (from 0-silent to 2-verbal), `iterlim` (the maximum number

of iterations), `methods` (the method used, one of `nr`, `bhhh` or `bfgs`) that are passed to `maxLik` .

Gradient and Hessian for the logit model

For the multinomial logit model, the gradient and the hessian have very simple expressions.

$$\frac{\partial \ln P_{ij}}{\partial \beta} = x_{ij} - \sum_l P_{il} x_{il}$$

$$\frac{\partial \ln L}{\partial \beta} = \sum_i \sum_j \left(y_{ij} - P_{ij} \right) x_{ij}$$

$$\frac{\partial^2 \ln L}{\partial \beta \partial \beta'} = \sum_i \sum_j P_{ij} \left(x_{ij} - \sum_l P_{il} x_{il} \right) \left(x_{ij} - \sum_l P_{il} x_{il} \right)'$$

Moreover, the log-likelihood function is globally concave, which mean that there is a unique optimum which is the global maximum. In this case, the Newton-Ralphson method is very efficient and the convergence is achieved after just a few iterations.

2.6. Interpretation

In a linear model, the coefficients can be directly interpretable as marginal effects of the explanatory variables on the explained variable. This is not the case for the multinomial models. However, meaningful results can be obtained using relevant transformations of the coefficients.

Marginal effects

The marginal effects are the derivatives of the probabilities with respect with the explanatory variables, which can be be individual-specific (z_i) or alternative specific (x_{ij}) :

$$\frac{\partial P_{ij}}{\partial z_i} = P_{ij} \left(\beta_j - \sum_l P_{il} \beta_l \right)$$

$$\frac{\partial P_{ij}}{\partial x_{ij}} = \gamma P_{ij} (1 - P_{ij})$$

$$\frac{\partial P_{ij}}{\partial x_{il}} = -\gamma P_{ij} P_{il}$$

– For an alternative-specific variable, the sign of the coefficient is directly interpretable. The marginal effect is obtained by multiplying the coefficient by the product of two probabilities which is at most 0.25. The rule of thumb is therefore to divide the coefficient by 4 in order to have an upper bound of the marginal effect.

– For an individual specific variable, the sign of the coefficient is not necessarily the sign of the coefficient. Actually, the sign of the marginal effect is given by $(\beta_j - \sum_l P_{il}\beta_l)$, which is positive if the coefficient for the j alternative is greater than a weighted average of the coefficients for all the alternative, the weights being the probabilities of choosing the alternatives.

Marginal rates of substitution

Coefficients are marginal utilities, which are not interpretable because utility is ordinal. However, ratios of coefficients are marginal rates of substitution, which are interpretable. For example, if the observable part of utility is : $V = \beta_o + \beta_1 x_1 + \beta x_2 + \beta x_3$; join variations of x_1 and x_2 which ensure the same level of utility are such that : $dV = \beta_1 dx_1 + \beta_2 dx_2 = 0$ so that :

$$-\frac{dx_2}{dx_1}\mid_{dV=0} = \frac{\beta_1}{\beta_2}$$

For example, if x_2 is transport cost (in euros), x_1 transport time (in hours), $\beta_1 = 1.5$ and $\beta_2 = 0.2$, $\frac{\beta_1}{\beta_2} = 30$ is the marginal rate of substitution of time in terms of euros and the value of 30 means that to reduce the travel time of one hour, the individual is willing to pay at most 30 euros more.

Consumer's surplus

Consumer's surplus is a very simple expression with multinomial logit models. It was first derived by Small and Rosen (1981).

The level of utility attained by an individual is $U_j = V_j + \epsilon_j$, j being the alternative chosen. The expected utility, from the searcher's point of view is then :

$$\mathrm{E}(\max_j U_j)$$

where the expectation is taken on the values of all the error terms. If the marginal utility of income (α) is known and constant, the expected surplus is simply $\mathrm{E}(\max_j U_j)/\alpha$.

This expected surplus is a very simple expression in the context of the logit model, which is called the "log-sum". We'll demonstrate this fact in the context of two alternatives.

With two alternatives, the values of ϵ_1 and ϵ_2 can be depicted in a plane. This plane contains all the possible combinations of (ϵ_1, ϵ_2). Some of them leads to the choice of alternative 1 and the other to the choice of alternative 2. More precisely, alternative 1 is chosen if $\epsilon_2 \leq V_1 - V_2 + \epsilon_1$ and alternative 2 is chosen if $\epsilon_1 \leq V_2 - V_1 + \epsilon_2$. The first expression is the equation of a straight line in the plan which delimits the choice for the two alternatives.

We can then write the expected utility as the sum of two terms E_1 and E_2, with :

$$E_1 = \int_{\epsilon_1=-\infty}^{\infty} \int_{-\infty}^{V_1-V_2+\epsilon_1} (V_1 + \epsilon_1) f(\epsilon_1) f(\epsilon_2) d\epsilon_1 d\epsilon_2$$

and

$$E_2 = \int_{\epsilon_2=-\infty}^{\infty} \int_{-\infty}^{V_2-V_1+\epsilon_1} (V_2 + \epsilon_2) f(\epsilon_1) f(\epsilon_2) d\epsilon_1 d\epsilon_2$$

with $f(z) = \exp(-e^{-z})$ the density of the Gumbell distribution.

We'll derive the expression for E_1, by symetry we'll guess the expression for E_2 and we'll then obtain the expected utility by summing E_1 and E_2.

$$E_1 = \int_{\epsilon_1=-\infty}^{\infty} (V_1 + \epsilon_1) \left(\int_{-\infty}^{V_1-V_2+\epsilon_1} f(\epsilon_2) d\epsilon_2 \right) f(\epsilon_1) d\epsilon_1$$

The expression in brackets is the cumulative density of ϵ_2. We then have :

$$E_1 = \int_{\epsilon_1=-\infty}^{\infty} (V_1 + \epsilon_1) e^{-e^{-(V_1-V_2)-\epsilon_1}} f(\epsilon_1) d\epsilon_1$$

$$E_1 = \int_{\epsilon_1=-\infty}^{\infty} (V_1 + \epsilon_1) e^{-\epsilon_1} e^{-ae^{-\epsilon_1}} d\epsilon_1$$

with $a = 1 + e^{-(V_1-V_2)} = \frac{e^{V_1}+e^{V_2}}{e^{V_1}} = \frac{1}{P_1}$

Let defines $z \mid e^{-z} = ae^{-\epsilon_1} \Leftrightarrow z = \epsilon_1 - \ln a$

We then have :

$$E_1 = \int_{\epsilon_1=-\infty}^{\infty} (V_1 + z + \ln a)/a e^{-z} e^{-e^{-z}} dz$$

$$E_1 = (V_1 + \ln a)/a + \mu/a$$

where μ is the expected value of a random variable which follows a standard Gumbell distribution, *i.e.* the Euler-Mascheroni constant.

$$E_1 = \frac{\ln(e^{V_1} + e^{V_2}) + \mu}{(e^{V_1} + e^{V_2})/e^{V_1}} = \frac{e^{V_1} \ln(e^{V_1} + e^{V_2}) + e^{V_1}\mu}{e^{V_1} + e^{V_2}}$$

By symmetry,

$$E_2 = \frac{e^{V_2} \ln(e^{V_1} + e^{V_2}) + e^{V_2}\mu}{e^{V_1} + e^{V_2}}$$

And then :

$$\mathrm{E}(U) = E_1 + E_2 = \ln(e^{V_1} + e^{V_2}) + \mu$$

More generally, in presence of J alternatives, we have :

$$\mathrm{E}(U) = \ln \sum_{j=1}^{J} e^{V_j} + \mu$$

and the expected surplus is, with α the constant marginal utility of income :

$$\mathrm{E}(U) = \frac{\ln \sum_{j=1}^{J} e^{V_j} + \mu}{\alpha}$$

2.7. Application

Train contains data about a stated preference survey in Netherlands. Users are asked to choose between two train trips characterized by four attributes :

– price : the price in cents of guilders,
– time : travel time in minutes,
– change : the number of changes,
– comfort : the class of comfort, 0, 1 or 2, 0 being the most comfortable class.

```
R> data("Train", package = "mlogit")
R> Tr <- mlogit.data(Train, shape = "wide", choice = "choice", varying = 4:11,
+       sep = "", alt.levels = c(1, 2), id = "id")
```

We first convert price and time in more meaningful unities, hours and euros (1 guilder is 2.20371 euros) :

```
R> Tr$price <- Tr$price/100 * 2.20371
R> Tr$time <- Tr$time/60
```

We then estimate the model : both alternatives being virtual train trips, it is relevant to use only generic coefficients and to remove the intercept :

```
R> m <- mlogit(choice ~ price + time + change + comfort | -1, Tr)
R> summary(m)

Call:
mlogit(formula = choice ~ price + time + change + comfort | -1,
    data = Tr, method = "nr", print.level = 0)

Frequencies of alternatives:
```

```
        1       2
0.50324 0.49676


nr method
5 iterations, 0h:0m:0s
g'(-H)^-1g = 0.00014
successive fonction values within tolerance limits


Coefficients :
          Estimate Std. Error  t-value  Pr(>|t|)
price   -0.0673580  0.0033933 -19.8506 < 2.2e-16 ***
time    -1.7205514  0.1603517 -10.7299 < 2.2e-16 ***
change  -0.3263409  0.0594892  -5.4857 4.118e-08 ***
comfort -0.9457256  0.0649455 -14.5618 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Log-Likelihood: -1724.2
```

All the coefficients are highly significant and have the predicted negative sign (remind than an increase in the variable comfort implies using a less comfortable class). The coefficients are not directly interpretable, but dividing them by the price coefficient, we get monetary values :

```
R> coef(m)[-1]/coef(m)[1]


    time    change  comfort
25.54337   4.84487 14.04028
```

We obtain the value of 26 euros for an hour of traveling, 5 euros for a change and 14 euros to access a more comfortable class.

The second example use the Fishing data. It illustrates the multi-part formula interface to describe the model, and the fact that it is not necessary to transform the data set using mlogit.data before the estimation, *i.e.* instead of using :

```
R> Fish <- mlogit.data(Fishing, shape = "wide", varying = 2:9, choice = "mode")
R> m <- mlogit(mode ~ price | income | catch, Fish)
```

it is possible to use mlogit with the original data.frame and the relevant arguments that will be internally passed to mlogit.data :

```
R> m <- mlogit(mode ~ price | income | catch, Fishing, shape = "wide",
+     varying = 2:9)
R> summary(m)
```

```
Call:
mlogit(formula = mode ~ price | income | catch, data = Fishing,
    shape = "wide", varying = 2:9, method = "nr", print.level = 0)

Frequencies of alternatives:
  beach    boat charter    pier
0.11337 0.35364 0.38240 0.15059

nr method
7 iterations, 0h:0m:0s
g'(-H)^-1g = 2.54E-05
successive fonction values within tolerance limits

Coefficients :
                    Estimate  Std. Error  t-value  Pr(>|t|)
altboat            8.4184e-01  2.9996e-01   2.8065 0.0050080 **
altcharter         2.1549e+00  2.9746e-01   7.2443 4.348e-13 ***
altpier            1.0430e+00  2.9535e-01   3.5315 0.0004132 ***
price             -2.5281e-02  1.7551e-03 -14.4046 < 2.2e-16 ***
altboat:income     5.5428e-05  5.2130e-05   1.0633 0.2876612
altcharter:income -7.2337e-05  5.2557e-05  -1.3764 0.1687088
altpier:income    -1.3550e-04  5.1172e-05  -2.6480 0.0080977 **
altbeach:catch     3.1177e+00  7.1305e-01   4.3724 1.229e-05 ***
altboat:catch      2.5425e+00  5.2274e-01   4.8638 1.152e-06 ***
altcharter:catch   7.5949e-01  1.5420e-01   4.9254 8.417e-07 ***
altpier:catch      2.8512e+00  7.7464e-01   3.6807 0.0002326 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-Likelihood: -1199.1
McFadden R^2:  0.19936
Likelihood ratio test : chisq = 597.16 (p.value=< 2.22e-16)
```

Several methods can be used to extract some results from the estimated model. `fitted`
returns the predicted probabilities for the outcome or for all the alternatives if outcome = FALSE
.

```
R> head(fitted(m))
```

```
[1] 0.3114002 0.4537956 0.4567631 0.3701758 0.4763721 0.4216448
```

```
R> head(fitted(m, outcome = FALSE))
```

```
          beach        boat   charter        pier
[1,] 0.09299769 0.5011740 0.3114002 0.09442817
[2,] 0.09151070 0.2749292 0.4537956 0.17976449
[3,] 0.01410358 0.4567631 0.5125571 0.01657625
[4,] 0.17065868 0.1947959 0.2643696 0.37017585
[5,] 0.02858215 0.4763721 0.4543225 0.04072324
[6,] 0.01029791 0.5572463 0.4216448 0.01081103
```

Finally, two further arguments can be usefully used while using mlogit

– `reflevel` indicates which alternative is the "reference" alternative, *i.e.* the one for which the coefficients are 0,
– `altsubset` indicates a subset on which the estimation has to be performed ; in this case, only the lines that corresponds to the selected alternatives are used and all the observations which corresponds to choices for unselected alternatives are removed :

```
R> m <- mlogit(mode ~ price | income | catch, Fish, reflevel = "charter",
+     alt.subset = c("beach", "pier", "charter"))
```

3. Tests

3.1. The three tests

As all models estimated by maximum likelihood, three testing procedures may be applied to test hypothesis about models fitted using mlogit . The hypothesis tested define two models :

– the unconstrained model that doesn't take these hypothesis into account,
– the constrained model that impose these hypothesis.

This in turns define three principles of tests :

– the Wald test is based only on the unconstrained model,
– the Lagrange multiplier tests is based only on the constrained model,
– the Likelihood ratio test is based on the comparison of both models.

The three principles of test are better understood using figure
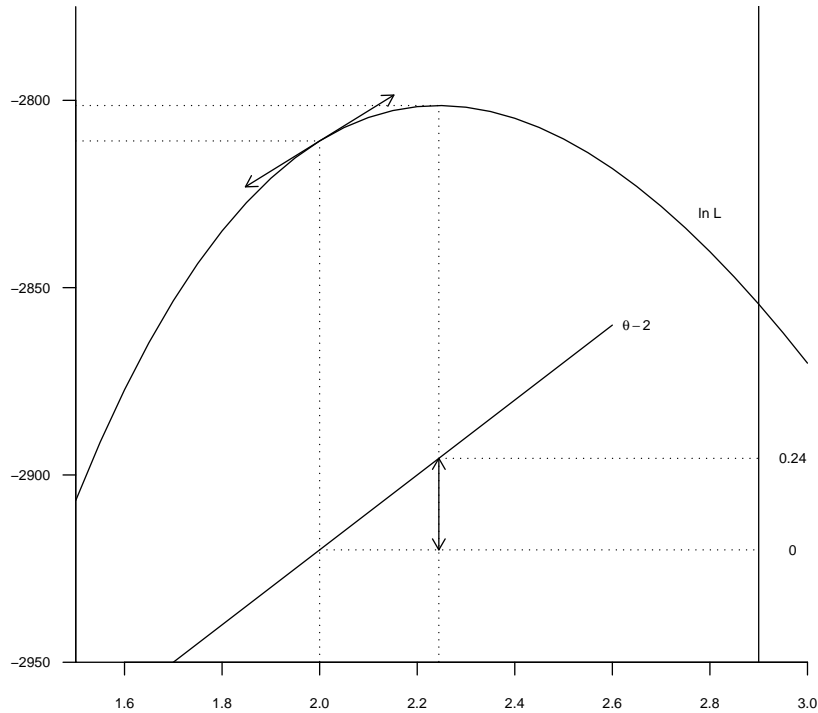
In this one dimensional setting, the hypothesis is of the form $\theta = \theta_o$, which can be written $f(\theta) = \theta - \theta_o$, with $f(\theta) = 0$ if the hypothesis is unforced. This is the equation of a straight line on 3.1 . The constrained model is just $\hat{\theta}_c = \theta_o$, *i.e.* the constrained model is not estimated. The unconstrained model corresponds to the maximum of the curve that represents the log-likelihood function.

– The Wald test is based on $f(\hat{\theta}_{nc})$ which is depicted by the arrow in 3.1. More generally, it is a vector of length J, whose expected value should be 0 if the hypothesis is true.

– The Lagrange multiplier is based on the slope of the likelihood curve evaluated at the constrained model : $\frac{\partial \ln L}{\theta}(\hat{\theta}_c)$. Here again, this should be a random vector with expected value equal to 0 if H_o is true.

– Finally, the likelihood ratio test compares both models. More specifically, the statistic is twice the value of the log-likelihood for the two models.

Two of these tests are implemented in the **lmtest** package : waldtest et lrtest . The wald test is also implemented in linearHypothesis from package **car** with a fairly different syntax. We provide special methods of `waldtest` and `lrtest` for `mlogit` objects and we also provide a function for the lagrange multiplier (or score) test called scoretest .

We'll see later on that the score test is especially usefull for `mlogit` objects when one is interested in extending the basic multinomial logit model. In this case, the unconstrained

model is much more difficult to estimate than the constrainted model which is the basic multinomial logit model. The score test, which is based on the constrained model is therefore very simple to compute.

For now, we'll just demonstrate the use of the testing in the usual setting where the two models are indicated. This can done by passing two fitted models to the testing function, or just one model and a formula which describes the second model.

We've previously estimated the following model :

```
R> m <- mlogit(mode ~ price | income | catch, Fishing, shape = "wide",
+      varying = 2:9)
```

The hypothesis that the income doesn't influence the choice for a fishing mode is a joint hypothesis that three coefficients are zero. The constrained model can be obtained by updated the previous model :

```
R> mc <- update(m, . ~ . | . - income | .)
```

The wald and likelihood ratio tests are then obtained by provinding the two models as arguments :

```
R> waldtest(m, mc)


Wald test

Model 1: mode ~ price | income | catch
Model 2: mode ~ price | 1 | catch
  Res.Df Df  Chisq Pr(>Chisq)
1   1171
2   1174 -3 28.613  2.701e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


R> lrtest(m, mc)


Likelihood ratio test

Model 1: mode ~ price | income | catch
Model 2: mode ~ price | 1 | catch
  #Df  LogLik Df  Chisq Pr(>Chisq)
1  11 -1199.1
2   8 -1214.2 -3 30.138  1.291e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
R> scoretest(mc, m)


        score test

data:  mode ~ price | income | catch
chisq = 29.7103, = 3, p-value = 1.588e-06
alternative hypothesis: unconstrained model
```

or just one of them and a formula that describes the second one :

```
R> lrtest(m, . ~ . | . - income | .)
R> lrtest(m, mode ~ price | 1 | catch)
R> lrtest(mc, . ~ . | . + income | .)
R> lrtest(mc, mode ~ price | income | catch)
R> waldtest(m, . ~ . | . - income | .)
R> waldtest(m, mode ~ price | 1 | catch)
R> waldtest(mc, . ~ . | . + income | .)
R> waldtest(mc, mode ~ price | income | catch)
R> scoretest(mc, . ~ . | . + income | .)
R> scoretest(mc, mode ~ price | income | catch)
```

4. Relaxing the iid hypothesis

With hypothesis 1 and 3, the error terms are *iid* (identically and independently distributed), *i.e.* not correlated and homoscedastic. Extensions of the basic multinomial logit model have been proposed by relaxing one of these two hypothesis while maintaining the second hypothesis of GUMBELL distribution.

4.1. The heteroskedastic logit model

The heteroskedastic logit model was proposed by Bhat (1995).

The probability that $U_l > U_j$ is :

$$P(\epsilon_j < V_l - V_j + \epsilon_l) = e^{-e^{-\frac{(V_l - V_j + \epsilon_l)}{\theta_j}}}$$

which implies the following conditional and unconditional probabilities

$$(P_l \mid \epsilon_l) = \prod_{j \neq l} e^{-e^{-\frac{(V_l - V_j + \epsilon_l)}{\theta_j}}} \tag{11}$$

$$P_l = \int_{-\infty}^{+\infty} \prod_{j \neq l} \left(e^{-e^{-\frac{(V_l - V_j + \epsilon_l)}{\theta_j}}} \right) \frac{1}{\theta_l} e^{-\frac{\epsilon_l}{\theta_l}} e^{-e^{-\frac{\epsilon_l}{\theta_l}}} d\epsilon_l \qquad (12)$$

We then apply the following change of variable :

$$u = e^{-\frac{\epsilon_l}{\theta_l}} \;\Rightarrow\; du = -\frac{1}{\theta_l} e^{-\frac{\epsilon_l}{\theta_l}} d\epsilon_l$$

The unconditional probability (12) can then be rewritten :

$$P_l = \int_0^{+\infty} \prod_{j \neq l} \left(e^{-e^{-\frac{V_l - V_j - \theta_l \ln u}{\theta_j}}} \right) e^{-u} du = \int_0^{+\infty} \left(e^{-\sum_{j \neq l} e^{-\frac{V_l - V_j - \theta_l \ln u}{\theta_j}}} \right) e^{-u} du$$

There is no closed form for this integral but it can be written the following way :

$$P_l = \int_0^{+\infty} G_l e^{-u} du$$

with

$$G_l = e^{-A_l} \quad A_l = \sum_{j \neq l} \alpha_j \quad \alpha_j = e^{-\frac{V_l - V_j - \theta_l \ln u}{\theta_j}}$$

This one-dimensional integral can be efficiently computed using a Gauss quadrature method, and more precisely a Gauss-Laguerre quadrature method :

$$\int_0^{+\infty} f(u) e^{-u} du = \sum_t f(u_t) w_t$$

where u_t and w_t are respectively the nodes and the weights.

$$P_l = \sum_t G_l(u_t) w_t$$

$$\frac{\partial G_l}{\partial \beta_k} = \sum_{j \neq l} \frac{\alpha_j}{\theta_j} (x_{lk} - x_{jk}) G_l$$

$$\frac{\partial G_l}{\partial \theta_l} = -\ln u \sum_{j \neq l} \frac{\alpha_j}{\theta_j} G_l$$

$$\frac{\partial G_l}{\partial \theta_j} = \ln \alpha_j \frac{\alpha_j}{\theta_j} G_l$$

To illustrate the estimation of the heteroscedastic logit model, we use the data used by (Bhat 1995). This data set is called ModeCanada .

```
R> data("ModeCanada", package = "mlogit")
```

As done in the article, we first restrict the sample to the user who don't choose the bus and choose a mode among the four modes available (train, air, bus and car.

```
R> busUsers <- with(ModeCanada, case[choice == 1 & alt == "bus"])
R> Bhat <- subset(ModeCanada, !case %in% busUsers & alt != "bus" &
+     nchoice == 4)
R> Bhat$alt <- Bhat$alt[drop = TRUE]
R> Bhat <- mlogit.data(Bhat, shape = "long", chid.var = "case",
+     alt.var = "alt", choice = "choice", drop.index = TRUE)
```

This restricts the sample to 2769 users.

```
R> ml <- mlogit(choice ~ freq + cost + ivt + ovt | urban + income,
+     Bhat, reflevel = "car")
R> hl <- mlogit(choice ~ freq + cost + ivt + ovt | urban + income,
+     Bhat, reflevel = "car", heterosc = TRUE)
R> summary(hl)
```

```
Call:
mlogit(formula = choice ~ freq + cost + ivt + ovt | urban + income,
    data = Bhat, reflevel = "car", heterosc = TRUE)


Frequencies of alternatives:
    car    train      air
0.45757 0.16721 0.37523


bfgs method
10 iterations, 0h:0m:9s
g'(-H)^-1g = 2.89E-07
gradient close to zero


Coefficients :
                  Estimate Std. Error t-value  Pr(>|t|)
alttrain         0.6783934  0.3327626  2.0387  0.041483 *
altair           0.6567544  0.4681631  1.4028  0.160667
freq             0.0639247  0.0049168 13.0014 < 2.2e-16 ***
cost            -0.0269615  0.0042831 -6.2948 3.078e-10 ***
ivt             -0.0096808  0.0010539 -9.1859 < 2.2e-16 ***
ovt             -0.0321655  0.0035930 -8.9523 < 2.2e-16 ***
alttrain:urban   0.7971316  0.1207392  6.6021 4.054e-11 ***
altair:urban     0.4454726  0.0821609  5.4220 5.895e-08 ***
```

```
alttrain:income -0.0125979  0.0039942 -3.1541  0.001610 **
altair:income    0.0188600  0.0032159  5.8646 4.503e-09 ***
sp.train         1.2371829  0.1104610 11.2002 < 2.2e-16 ***
sp.air           0.5403239  0.1118353  4.8314 1.356e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Log-Likelihood: -1838.1
McFadden R^2:  0.35211
Likelihood ratio test : chisq = 1998 (p.value=< 2.22e-16)
```

The results obtained by Bhat (1995) can't be exactly reproduced because he uses some weights that are not available in the data set. However, we obtain very close values for the two estimated scale parameters for the train sp.train and for the air mode sp.air.

The homoscedascticity hypothesis can be tested using any of the three tests. A particular convenient syntax is provided in this case.

For the likelihood ratio and the wald test, one can pass only the fitted model as argument. In this case, it is guessed that the hypothesis that the user wants to test is the homoscedasticity hypothesis, *i.e.* sp.train=1 and sp.air=1 .

```
R> lrtest(hl)


Likelihood ratio test

Model 1: choice ~ freq + cost + ivt + ovt | urban + income
Model 2: choice ~ freq + cost + ivt + ovt | urban + income
  #Df  LogLik Df  Chisq Pr(>Chisq)
1  12 -1838.1
2  10 -1841.6 -2 6.8882    0.03193 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R> waldtest(hl)


        Wald test

data:  fixed uncorrelated
chisq = 25.1955, df = 2, p-value = 3.380e-06
```

The wald test can also be computed using the linearHypothesis function from the **car** package :

```
R> library("car")
R> linearHypothesis(hl, c("sp.air=1", "sp.train=1"))
```

```
Linear hypothesis test

Hypothesis:
sp.air = 1
sp.train = 1

Model 1: restricted model
Model 2: choice ~ freq + cost + ivt + ovt | urban + income

  Res.Df Df  Chisq Pr(>Chisq)
1   2759
2   2757  2 25.195  3.380e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

For the score test, we provide the constrained model as argument, which is the standard multinomial logit model and the supplementary argument which defines the unconstrained model, which is in this case heterosc = TRUE .

```
R> scoretest(ml, heterosc = TRUE)


        score test

data:  heterosc = TRUE
chisq = 9.4883, = 2, p-value = 0.008703
alternative hypothesis: heteroscedastic model
```

The second example we provide reproduces the first column of table 23.28 page 855 of Grenne (2008).

```
R> data("TravelMode", package = "AER")
R> TravelMode <- mlogit.data(TravelMode, choice = "choice", shape = "long",
+     alt.var = "mode", chid.var = "individual")
R> TravelMode$avinc <- with(TravelMode, (mode == "air") * income)
R> ml <- mlogit(choice ~ wait + gcost + avinc, TravelMode, reflevel = "car")
R> hl <- mlogit(choice ~ wait + gcost + avinc, TravelMode, reflevel = "car",
+     heterosc = TRUE)
R> summary(hl)


Call:
mlogit(formula = choice ~ wait + gcost + avinc, data = TravelMode,
    reflevel = "car", heterosc = TRUE)
```

```
Frequencies of alternatives:
    car     air   train     bus
0.28095 0.27619 0.30000 0.14286

bfgs method
43 iterations, 0h:0m:3s
g'(-H)^-1g = 3.77E-07
gradient close to zero

Coefficients :
          Estimate Std. Error t-value Pr(>|t|)
altair    7.832450  10.950706  0.7152   0.4745
alttrain  7.171867   9.135295  0.7851   0.4324
altbus    6.865775   8.829608  0.7776   0.4368
wait     -0.196843   0.288274 -0.6828   0.4947
gcost    -0.051562   0.069444 -0.7425   0.4578
avinc     0.040253   0.060680  0.6634   0.5071
sp.air    4.024020   5.977821  0.6732   0.5008
sp.train  3.854208   6.220456  0.6196   0.5355
sp.bus    1.648749   2.826916  0.5832   0.5597

Log-Likelihood: -195.66
McFadden R^2:  0.31047
Likelihood ratio test : chisq = 176.2 (p.value=< 2.22e-16)
```

Note that the ranking of the scale parameters differs from the previous example. In particular, the error of the air utility has the largest variance as it has the smallest one in the previous example.

The standard deviations print at the end of table 23.28 are obtained by multiplying the scale parameters by $\pi/\sqrt{6}$:

```
R> c(coef(hl)[7:9], sp.car = 1) * pi/sqrt(6)
```

```
  sp.air sp.train   sp.bus   sp.car
5.161007 4.943214 2.114603 1.282550
```

Note that the standard deviations of the estimated scale parameters are very high, which means that they are poorly identified. This is confirmed by the fact that the homoscedasticity hypothesis is not rejected :

```
R> waldtest(hl)
```

```
        Wald test
```

```
data:  fixed uncorrelated
chisq = 3.6356, df = 3, p-value = 0.3036

R> lrtest(hl)

Likelihood ratio test

Model 1: choice ~ wait + gcost + avinc
Model 2: choice ~ wait + gcost + avinc
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   9 -195.66
2   6 -199.13 -3 6.9357    0.07398 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R> scoretest(ml, heterosc = TRUE)

        score test

data:  heterosc = TRUE
chisq = 21.566, = 3, p-value = 8.03e-05
alternative hypothesis: heteroscedastic model
```

4.2. The nested logit model

The nested logit model was first proposed by McFadden (1978). It is a generalization of the multinomial logit model that rests on the idea that some alternatives may be joined in several groups (called nests). The error terms may then present some correlation in the same nest, whereas error terms of different nests are still uncorrelated.

We suppose that the alternatives can be put into M different nests. This implies the following multivariate distribution for the error terms.

$$\exp\left(-\sum_{m=1}^{M}\left(\sum_{j\in B_m} e^{-\epsilon_j/\lambda_m}\right)^{\lambda_m}\right)$$

The marginal distributions of the ϵs are still univariate extreme value, but there is now some correlation within nests. $1 - \lambda_m$ is a measure of the correlation, *i.e.* $\lambda_m = 1$ implies no correlation. It can then be shown that the probability of choosing alternative j that is part of nest l is :

$$P_j = \frac{e^{V_j/\lambda_l}\left(\sum_{k\in B_l} e^{V_k/\lambda_l}\right)^{\lambda_l-1}}{\sum_{m=1}^{M}\left(\sum_{k\in B_m} e^{V_k/\lambda_m}\right)^{\lambda_m}}$$

Let write : $V_j = Z_j + W_l$

$$P_j = \frac{e^{(Z_j+W_l)/\lambda_l}}{\sum_{k\in B_l} e^{(Z_k+W_l)/\lambda_l}} \times \frac{\left(\sum_{k\in B_l} e^{(Z_k+W_l)/\lambda_l}\right)^{\lambda_l}}{\sum_{m=1}^{M} \left(\sum_{k\in B_m} e^{(Z_k+W_m)/\lambda_m}\right)^{\lambda_m}}$$

$$P_j = \frac{e^{Z_j/\lambda_l}}{\sum_{k\in B_l} e^{Z_k/\lambda_l}} \times \frac{\left(\sum_{k\in B_l} e^{(Z_k+W_l)/\lambda_l}\right)^{\lambda_l}}{\sum_{m=1}^{M} \left(\sum_{k\in B_m} e^{(Z_k+W_m)/\lambda_m}\right)^{\lambda_m}}$$

$$\left(\sum_{k\in B_l} e^{(Z_k+W_l)/\lambda_l}\right)^{\lambda_l} = \left(e^{W_l/\lambda_l} \sum_{k\in B_l} e^{Z_k/\lambda_l}\right)^{\lambda_l} = e^{W_l+\lambda_l I_l}$$

with $I_l = \ln \sum_{k\in B_l} e^{Z_k/\lambda_l}$ which is often denoted as the inclusive value or inclusive utility. We then can write the probability of choosing alternative j as :

$$P_j = \frac{e^{Z_j/\lambda_l}}{\sum_{k\in B_l} e^{Z_k/\lambda_l}} \times \frac{e^{W_l+\lambda_l I_l}}{\sum_{m=1}^{M} e^{W_m+\lambda_m I_m}}$$

The first term $P_{j|l}$ is the conditional probability of choosing alternative j if nest l is chosen. It is often referred as the *lower model*. The second term P_l is the marginal probability of choosing the nest l and is referred as the *upper model*.

$W_m + \lambda_m I_m$ can be interpreted as the expected utility of choosing the best alternative of the nest m, W_m being the expected utility of choosing an alternative in this nest (whatever this alternative is) and $\lambda_m I_m$ being the expected extra utility he receives by being able to choose the best alternative in the nest.

The inclusive values links the two models.

It is then straightforward to show that IIA applies within nests, but not for two alternatives in different nests.

A slightly different version of the nested logit model (Daly 1987) is often used, but is not compatible with the random utility maximization hypothesis. Its difference with the previous expression is that the determinist parts of the utility for each alternative is not normalized by the nest elasticity :

$$P_j = \frac{e^{V_j} \left(\sum_{k\in B_l} e^{V_k}\right)^{\lambda_l-1}}{\sum_{m=1}^{M} \left(\sum_{k\in B_m} e^{V_k}\right)^{\lambda_m}}$$

The differences between the two versions have been discussed in Koppelman and Wen (1998) and Heiss (2002).

The gradient is, for the first version of the model and denoting $N_m = \sum_{k\in B_m} e^{V_k/\lambda_m}$:

$$
\begin{cases}
\frac{\partial \ln P_j}{\partial \beta} &= \frac{x_j}{\lambda_l} + \frac{\lambda_l - 1}{\lambda_l} \frac{1}{N_l} \sum_{k \in B_l} e^{V_k/\lambda_l} x_k - \frac{1}{\sum_m N_m^{\lambda_m}} \sum_m N_m^{\lambda_m - 1} \sum_{k \in B_m} e^{V_k/\lambda_m} x_k \\
\frac{\partial \ln P_j}{\partial \lambda_l} &= -\frac{V_j}{\lambda_l^2} + \ln N_l - \frac{\lambda_l - 1}{\lambda_l^2} \frac{1}{N_l} \sum_{k \in B_l} V_k e^{V_k/\lambda_l} \\
& \quad - \frac{N_l^{\lambda_l}}{\sum_m N_m^{\lambda_m}} \left(\ln N_l - \frac{1}{\lambda_l N_l} \sum_{k \in B_l} V_k e^{V_k/\lambda_l} \right) \\
\frac{\partial \ln P_j}{\partial \lambda_m} &= -\frac{N_m^{\lambda_m}}{\sum_m N_m^{\lambda_m}} \left(\ln N_m - \frac{1}{\lambda_m N_m} \sum_{k \in B_m} V_k e^{V_k/\lambda_m} \right)
\end{cases}
$$

Denoting $P_{j|l} = \frac{e^{V_j/\lambda_l}}{N_l}$ the conditional probability of choosing alternative j if nest l is chosen, $P_l = \frac{N_l^{\lambda_l}}{\sum_m N_m^{\lambda_m}}$ the probability of choosing nest l, $\bar{x}_l = \sum_{k \in B_l} P_{k|l} x_k$ the weight average value of x in nest l, $\bar{x} = \sum_{m=1}^M P_m \bar{x}_m$ the weight average of x for all the nests and $\bar{V}_l = \sum_{k \in B_l} P_{k|l} V_k$

$$
\begin{cases}
\frac{\partial \ln P_j}{\partial \beta} &= \frac{1}{\lambda_l} [x_j - (1 - \lambda_l)\bar{x}_l] - \bar{x} \\
\frac{\partial \ln P_j}{\partial \lambda_l} &= -\frac{1}{\lambda_l^2} [V_j - \lambda_l^2 \ln N_l - (1 - \lambda_l)\bar{V}_l] - \frac{P_l}{\lambda_l^2} [\lambda_l^2 \ln N_l - \lambda_l \bar{V}_l] \\
\frac{\partial \ln P_j}{\partial \lambda_m} &= \frac{P_m}{\lambda_m} [\bar{V}_m - \lambda_m \ln N_m]
\end{cases}
$$

$$
\begin{cases}
\frac{\partial \ln P_j}{\partial \beta} &= \frac{x_j - \{(1-\lambda_l)\bar{x}_l + \lambda_l \bar{x}\}}{\lambda_l} \\
\frac{\partial \ln P_j}{\partial \lambda_l} &= -\frac{V_j - \{\lambda_l(1 - P_l)\lambda_l \ln N_l + (1 - \lambda_l(1 - P_l))\bar{V}_l\}}{\lambda_l^2} \\
\frac{\partial \ln P_j}{\partial \lambda_m} &= \frac{P_m}{\lambda_m} [\bar{V}_m - \lambda_m \ln N_m]
\end{cases}
$$

For the unscaled version, the gradient is :

$$
\begin{cases}
\frac{\partial \ln P_j}{\partial \beta} &= x_j - (1 - \lambda_l)\bar{x}_l - \sum_m \lambda_m P_m \bar{x}_m \\
\frac{\partial \ln P_j}{\partial \lambda_l} &= (1 - P_l) \ln N_l \\
\frac{\partial \ln P_j}{\partial \lambda_m} &= -P_m \ln N_m
\end{cases}
$$

With overlapping nests, the notations are slightly modified :

$$
P_j = \frac{\sum_{l | j \in B_l} e^{V_j/\lambda_l} N_l^{\lambda_l - 1}}{\sum_m N_m^{\lambda_m}}
$$

$$
P_j = \sum_{l | j \in B_l} \frac{e^{V_j/\lambda_l}}{N_l} \frac{N_l^{\lambda_l}}{\sum_m N_m^{\lambda_m}} = \sum_{l | j \in B_l} P_{j|l} P_l
$$

$$
\begin{cases}
\frac{\partial \ln P_j}{\partial \beta} &= \sum_{l | j \in B_l} \frac{P_{j|l} P_l}{P_j} \frac{x_j - \{(1-\lambda_l)\bar{x}_l + \lambda_l \bar{x}\}}{\lambda_l} \\
\frac{\partial \ln P_j}{\partial \lambda_l} &= -\frac{P_{j|l} P_l}{P_j} \frac{V_j - \{\lambda_l(1 - P_j/P_{j|l})\lambda_l \ln N_l + (1 - \lambda_l(1 - P_j/P_{j|l}))\bar{V}_l\}}{\lambda_l^2} \\
\frac{\partial \ln P_j}{\partial \lambda_m} &= \frac{P_m}{\lambda_m} [\bar{V}_m - \lambda_m \ln N_m]
\end{cases}
$$

For the unscaled version, the gradient is :

$$
\begin{cases}
\frac{\partial \ln P_j}{\partial \beta} &= \sum_{l|j \in B_l} \frac{P_{j|l} P_l}{P_j} (x_j - (1 - \lambda_l)\bar{x}_l) - \sum_m \lambda_m P_m \bar{x}_m \\
\frac{\partial \ln P_j}{\partial \lambda_l} &= P_l \left(\frac{P_{jl}}{P_j} - 1\right) \ln N_l \\
\frac{\partial \ln P_j}{\partial \lambda_m} &= -P_m \ln N_m
\end{cases}
$$

We illustrate the estimation of the unscaled nested logit model with an example used in (Grenne 2008). The dataset, called TravelMode has already been used. Four transport modes are available and two nests are considered :

– the *ground* nest with bus, train and car modes,
– the *fly* nest with the air modes.

Note that the second nest is a "degenerate" nest, which means that it contains only one alternative. In this case, the nest elasticity is somewhat meaningless, as it is related to the degree of correlation of the alternatives within the nests and that there is only on alternative in this nest. This parameter can only be identified in a very special case : the use of the unscaled version of the nested logit model with generic variable. This is exactly the situation considered by (Grenne 2008) and presented in the table 21.11 p. 730.

```
R> data("TravelMode", package = "AER")
R> TravelMode <- mlogit.data(TravelMode, choice = "choice", shape = "long",
+     alt.var = "mode", chid.var = "individual")
R> TravelMode$avinc <- with(TravelMode, (mode == "air") * income)
R> nl <- mlogit(choice ~ wait + gcost + avinc, TravelMode, reflevel = "car",
+     nests = list(fly = "air", ground = c("train", "bus", "car")),
+     unscaled = TRUE)
R> summary(nl)


Call:
mlogit(formula = choice ~ wait + gcost + avinc, data = TravelMode,
    reflevel = "car", nests = list(fly = "air", ground = c("train",
        "bus", "car")), unscaled = TRUE)


Frequencies of alternatives:
    car     air   train     bus
0.28095 0.27619 0.30000 0.14286


bfgs method
17 iterations, 0h:0m:0s
g'(-H)^-1g = 1.02E-07
gradient close to zero
```

```
Coefficients :
          Estimate Std. Error t-value  Pr(>|t|)
altair    6.042373   1.331325   4.5386 5.662e-06 ***
alttrain  5.064620   0.676010   7.4919 6.795e-14 ***
altbus    4.096325   0.628870   6.5138 7.328e-11 ***
wait     -0.112618   0.011826  -9.5232 < 2.2e-16 ***
gcost    -0.031588   0.007434  -4.2491 2.147e-05 ***
avinc     0.026162   0.019842   1.3185   0.18732
iv.fly    0.586009   0.113056   5.1833 2.180e-07 ***
iv.ground 0.388962   0.157904   2.4633   0.01377 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-Likelihood: -193.66
McFadden R^2:  0.31753
Likelihood ratio test : chisq = 180.21 (p.value=< 2.22e-16)
```

Recall that the unscaled version version of the nested logit model is not compatible with the hypothesis of random utility maximisation. The parameters associated with the inclusive values are far more difficult to interpret in this case.

The second example deals with a choice of a heating mode. The data set is called HC . There are five alternatives, four of them provide also cooling : gaz central with cooling gcc, electric central with cooling ecc, electric room with cooling erc and heat pump with cooling hpc the other three provide only heating, these are electric central ec, electric room er and gaz central gc.

```
R> data("HC", package = "mlogit")
R> HC <- mlogit.data(HC, varying = c(2:8, 10:16), choice = "depvar",
+     shape = "wide")
R> cooling.modes <- HC$alt %in% c("gcc", "ecc", "erc", "hpc")
R> head(HC)

      depvar  icca occa income alt   ich  och chid
1.ec   FALSE 27.28 2.95     20  ec 24.50 4.09    1
1.ecc  FALSE 27.28 2.95     20 ecc  7.86 4.09    1
1.er   FALSE 27.28 2.95     20  er  7.37 3.85    1
1.erc   TRUE 27.28 2.95     20 erc  8.79 3.85    1
1.gc   FALSE 27.28 2.95     20  gc 24.08 2.26    1
1.gcc  FALSE 27.28 2.95     20 gcc  9.70 2.26    1
```

icca and occa are the investment and the operating cost of the cooling part of the system. This is only relevant for the cooling modes.

```
R> HC$icca[!cooling.modes] <- HC$occa[!cooling.modes] <- 0
```

We now estimate a nested logit model with two nests : the cooling / non-cooling systems :

```
R> ml <- mlogit(depvar ~ occa + icca + och + ich, HC)
R> nl <- mlogit(depvar ~ occa + icca + och + ich, HC, nests = list(cooling = c("ecc",
+     "erc", "gcc", "hpc"), noncool = c("ec", "gc", "er")))
R> summary(nl)

Call:
mlogit(formula = depvar ~ occa + icca + och + ich, data = HC,
    nests = list(cooling = c("ecc", "erc", "gcc", "hpc"), noncool = c("ec",
        "gc", "er")))

Frequencies of alternatives:
    ec   ecc    er   erc    gc   gcc   hpc
0.004 0.016 0.032 0.004 0.096 0.744 0.104

bfgs method
18 iterations, 0h:0m:0s
g'(-H)^-1g = 2.24E-07
gradient close to zero

Coefficients :
            Estimate Std. Error t-value Pr(>|t|)
altecc      2.171367   3.401923  0.6383  0.52329
alter      -2.455199   1.071462 -2.2914  0.02194 *
alterc      1.756250   3.547708  0.4950  0.62057
altgc      -0.208090   0.469091 -0.4436  0.65733
altgcc      2.234177   3.383645  0.6603  0.50907
althpc      1.272654   3.618232  0.3517  0.72504
occa       -0.966387   0.708161 -1.3646  0.17237
icca       -0.051249   0.081461 -0.6291  0.52927
och        -0.868681   0.445484 -1.9500  0.05118 .
ich        -0.205005   0.090851 -2.2565  0.02404 *
iv.cooling  0.333827   0.172073  1.9400  0.05238 .
iv.noncool  0.328934   0.212062  1.5511  0.12087
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-Likelihood: -188.03
McFadden R^2:  0.16508
Likelihood ratio test : chisq = 74.354 (p.value=5.2125e-14)
```

The two nest elasticities are about 0.3, which implies a correlation of 0.7, which is quite high. The results of the three tests are given bellow :

```
R> lrtest(nl)

Likelihood ratio test

Model 1: depvar ~ occa + icca + och + ich
Model 2: depvar ~ occa + icca + och + ich
  #Df  LogLik Df  Chisq Pr(>Chisq)
1  12 -188.03
2  10 -192.88 -2 9.6853   0.007886 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R> waldtest(nl)

        Wald test

data:  fixed uncorrelated
chisq = 15.3069, df = 2, p-value = 0.0004744

R> scoretest(ml, nests = list(cooling = c("ecc", "erc", "gcc", "hpc"),
+     noncool = c("ec", "gc", "er")))

        score test

data:  rien
chisq = 15.1762, = 2, p-value = 0.0005065
alternative hypothesis: nested model
```

The three tests reject the null hypothesis of no correlation at the 1% level. The two nests elasticities being very closed, we'd like to test equality between both elasticities. linearHypothesis provides a particular convenient interface for that :

```
R> linearHypothesis(nl, "iv.cooling = iv.noncool")

Linear hypothesis test

Hypothesis:
iv.cooling - iv.noncool = 0

Model 1: restricted model
Model 2: depvar ~ occa + icca + och + ich

  Res.Df Df  Chisq Pr(>Chisq)
1    239
2    238  1 0.0011      0.9739
```

Therefore the hypothesis is not rejected and can be unforced by updating the model with the argument `un.nest.el` equal to `TRUE`.

```
R> nl2 <- update(nl, un.nest.el = TRUE)
```

A likelihood ratio test confirms that the hypothesis is not rejected :

```
R> lrtest(nl, nl2)

Likelihood ratio test

Model 1: depvar ~ occa + icca + och + ich
Model 2: depvar ~ occa + icca + och + ich
  #Df  LogLik Df  Chisq Pr(>Chisq)
1  12 -188.03
2  11 -188.03 -1 0.0012     0.9723
```

5. The general extreme value model

McFadden (1978) developed a general model that suppose that the join distribution of the error terms follow a a multivariate extreme value distribution. Let G be a function with J arguments $y_j \geq 0$. G has the following characteristics :

 i) it is non negative $G(y_1, \ldots, y_J) \geq 0 \forall j$,

 ii) it is homogeneous of degree 1 in all its arguments $G(\lambda y_1, \ldots \lambda y_J) = \lambda G(y_1, \ldots, y_J)$,

iii) for all its argument, $\lim_{y_j \to +\infty} = G(y_1, \ldots y_J) = +\infty$,

iv) for distinct arguments, $\frac{\partial^k G}{\partial y_i, \ldots, y_j}$ is non-negative if k is odd and non-positive if k is even.

Assume now that the joint cumulative distribution of the error terms can be written :

$$F(\epsilon_1, \epsilon_2, \ldots, \epsilon_J) = \exp\left(-G\left(e^{-\epsilon_1}, e^{-\epsilon_2}, \ldots, e^{-\epsilon_J}\right)\right)$$

We first show that this is a multivariate extreme value distribution. This implies :

1. if F is a joint cumulative distribution of probability, for any $\lim_{\epsilon_j \Rightarrow -\infty} F(\epsilon_1 \ldots \epsilon_J) = 0$,

2. if F is a joint cumulative distribution of probability, $\lim_{\epsilon_1, \ldots \epsilon_J \to +\infty} F(\epsilon_1 \ldots \epsilon_J) = 1$,

3. all the cross-derivates of any order of F should be non-negative,

4. if F is a multivariate extreme value distribution, the marginal distribution of any ϵ_j, which is $\lim_{\epsilon_k \to +\infty \forall k \neq j} F(\epsilon_1 \ldots \epsilon_J)$ should be an extreme value distribution.

For point 1, if $\epsilon_j \to -\infty$, $y_j \to +\infty$, $G \to +\infty$ and then $F \to 0$.

For point 2, if $(\epsilon_1, \ldots, \epsilon_J) \to +\infty$, $G \to 0$ and then $F \to 1$.

For point 3, let denote [5] :

$$Q_k = Q_{k-1} G_k - \frac{\partial Q_{k-1}}{\partial y_k} \text{ and } Q_1 = G_1$$

Q_k is a sum of signed terms that are products of cross derivates of G of various order. If each term of Q_{k-1} are non-negative, so is $Q_{k-1} G_k$ (from iv, the first derivates are non-negative. Moreover "each term in $\frac{\partial Q_{k-1}}{\partial y_k}$ is non positive, since one of the derivates within each term has increased in order, changing from even to odd or vice versa, with a hypothesized change in sign (hypothesis iv). Hence each term in Q_k is non negative and, by induction, Q_k is non-negative for $k = 1, 2, \ldots J$.

Suppose that the $k - 1$-order cross-derivate of F can be writen :

$$\frac{\partial^{k-1} F}{\partial \epsilon_1 \ldots \partial \epsilon_{k-1}} = e^{-\epsilon_1} \ldots e^{-\epsilon_k} Q_{k-1} F$$

Then , the k-order derivate is :

$$\frac{\partial^k F}{\partial \epsilon_1 \ldots \partial \epsilon_k} = e^{-\epsilon_1} \ldots e^{-\epsilon_k} Q_k F$$

$Q_1 = G_1$ is non-negative, so are $Q_2, Q_3, \ldots Q_k$ and therefore all the cross-derivates of any order are non-negatives.

To demonstrate the fourth point, we compute the marginal cumulative distribution of ϵ_l which is :

$$F(\epsilon_l) = \lim_{\epsilon_j \to +\infty \forall j \neq l} F(\epsilon_1, \ldots, \epsilon_l, \ldots \epsilon_J) = \exp\left(-G\left(0, \ldots, e^{-\epsilon_l}, \ldots, 0\right)\right)$$

with G being homogeneous of degree one, we have :

$$G\left(0, \ldots, e^{-\epsilon_l}, \ldots, 0\right) = a_l e^{-\epsilon_l}$$

with $a_l = G(0, \ldots, 1, \ldots, 0)$. The marginal distribution of ϵ_l is then :

$$F(\epsilon_l) = \exp\left(-a_l e^{-\epsilon_l}\right)$$

which is an uni-variate extreme value distribution.

We note compute the probabilities of choosing an alternative :

We denote G_l the derivative of G respective to the lth argument. The derivative of F respective to the ϵ_l is then :

---

5. citation from Mc Fadden.

$$F_l(\epsilon_1, \epsilon_2, \ldots, \epsilon_J) = e^{-\epsilon_l} G_l \left(e^{-\epsilon_1}, e^{-\epsilon_2}, \ldots, e^{-\epsilon_J}\right) \exp\left(-G\left(e^{-\epsilon_1}, e^{-\epsilon_2}, \ldots, e^{-\epsilon_J}\right)\right)$$

which is the density of ϵ_l for given values of the other $J - 1$ error terms.

The probability of choosing alternative l is the probability that $U_l > U_j \; \forall j \neq l$ which is equivalent to $\epsilon_j < V_l - V_j + \epsilon_l$.

This probability is then :

$$
\begin{aligned}
P_l &= \int_{-\infty}^{+\infty} F_l(V_l - V_1 + \epsilon_l, V_l - V_2 + \epsilon_l, \ldots, V_l - V_J + \epsilon_l) d\epsilon_l \\
&= \int_{-\infty}^{+\infty} e^{-\epsilon_l} G_l \left(e^{-V_l + V_1 - \epsilon_l}, e^{-V_l + V_2 - \epsilon_l}, \ldots, e^{-V_l + V_J - \epsilon_l}\right) \\
&\times \exp\left(-G\left(e^{-V_l + V_1 - \epsilon_l}, e^{-V_l + V_2 - \epsilon_l}, \ldots, e^{-V_l + V_J - \epsilon_l}\right)\right) d\epsilon_l
\end{aligned}
$$

G being homogeneous of degree one, one can write :

$$G\left(e^{-V_l + V_1 - \epsilon_l}, e^{-V_l + V_2 - \epsilon_l}, \ldots, e^{-V_l + V_J - \epsilon_l}\right) = e^{-V_l} e^{-\epsilon_l} \times G\left(e^{V_1}, e^{V_2}, \ldots, e^{V_J}\right)$$

Homogeneity of degree one implies homogeneity of degree 0 of the first derivative :

$$G_l\left(e^{-V_l + V_1 - \epsilon_l}, e^{-V_l + V_2 - \epsilon_l}, \ldots, e^{-V_l - V_J - \epsilon_l}\right) = G_l\left(e^{V_1}, e^{V_2}, \ldots, e^{V_J}\right)$$

The probability of choosing alternative i is then :

$$P_l = \int_{-\infty}^{+\infty} e^{-\epsilon_l} G_l\left(e^{V_1}, e^{V_2}, \ldots, e^{V_J}\right) \exp\left(-e^{-\epsilon_l} e^{-V_l} G\left(e^{V_1}, e^{V_2}, \ldots, e^{V_J}\right)\right) d\epsilon_l$$

$$P_l = G_l \int_{-\infty}^{+\infty} e^{-\epsilon_l} \exp\left(-e^{-\epsilon_l} e^{-V_l} G\right) d\epsilon_l$$

$$P_l = G_l \frac{1}{e^{-V_l} G} \left[\exp\left(-e^{-\epsilon_l} e^{-V_l} G\right)\right]_{-\infty}^{+\infty} = \frac{G_l}{e^{-V_l} G}$$

Finally, the probability of choosing alternative i can be written :

$$P_l = \frac{e^{V_l} G_l\left(e^{V_1}, e^{V_2}, \ldots, e^{V_J}\right)}{G\left(e^{V_1}, e^{V_2}, \ldots, e^{V_J}\right)}$$

Among this vast family of models, several authors have proposed some nested logit models with overlapping nests Koppelman and Wen (2000) and Wen and Koppelman (2001)

6. The random parameters (or mixed) logit model

A mixed logit model or random parameters logit model is a logit model for which the parameters are assumed to vary from one individual to another.

6.1. The probabilities

The standard logit model is :

$$P_{il} = \frac{e^{\beta' x_{il}}}{\sum_j e^{\beta' x_{ij}}}$$

The mixed logit model is :

$$P_{il} = \frac{e^{\beta'_i x_{il}}}{\sum_j e^{\beta'_i x_{ij}}}$$

Two strategies of estimation may be considered :

– estimate the coefficients for each individual in the sample,
– consider the coefficients as random variables.

The first approach is of limited interest, because it would requires numerous observations for each individual.

The second approach leads to the mixed logit model.

The probability that individual i will choose alternative l is :

$$P_{il} \mid \beta_i = \frac{e^{\beta'_i x_{il}}}{\sum_j e^{\beta'_i x_{ij}}}$$

This is the probability for individual i conditional on the vector of coefficients β_i. To get the unconditional probability, we have the average probability for the different values of β_i.

If $V_{il} = \alpha_i + \beta_i x_{il}$ and the density of β_i is $f(\beta_i, \theta)$:

$$P_{il} = \mathrm{E}(P_{il} \mid \beta_i) = \int_{-\infty}^{+\infty} (P_{il} \mid \beta_i) f(\beta_i, \theta) d\beta_i$$

which can be estimated efficiently by quadrature methods.

If $V_{il} = \alpha_i + \beta_i x_{il} + \gamma_i v_{il}$ and the density of β_i and γ_i is $f(\beta_i, \gamma_i, \theta)$

$$P_{il} = \mathrm{E}(P_{il} \mid \beta_i, \gamma_i) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (P_{il} \mid \beta_i, \gamma_i) f(\beta_i, \gamma_i \theta) d\beta_i d\gamma_i$$

can be estimated by simulations.

6.2. Panel data

Especially important for stated preference survey were several questions are asked to every individual.

Joint probabilities for each individual are computed.

$$P^r_{ikl} = \frac{e^{\beta^r_i x_{ikl}}}{\sum_j e^{\beta^r_i x_{ikj}}}$$

$$P_{ik}^r = \prod_l {P_{ikl}^r}^{y_{ikl}}$$

$$P_i^r = \prod_i \prod_l {P_{ikl}^r}^{y_{ikl}}$$

$$\bar{P}_i = \frac{1}{R} P_i^r$$

6.3. Simulations

The probabilities for the random parameter logit are integrals with no closed form. Moreover, the degree of integration is the number of random parameters. In practice, these models are estimated using simulation techniques, *i.e.* the expected value is replaced by an arithmetic mean. More precisely :

– make an initial hypothesis about the distribution of the random parameter : β_i follows a normal distribution with mean μ and standard deviation σ,
– draw R numbers on this distribution,
– for each draw β_i^r, compute the probability : $P_{il}^r = \frac{e^{\beta_i^T x_{il}}}{\sum_j e^{\beta_i^r x_{ij}}}$
– compute the average of these probabilities : $\bar{P}_{il} = \sum_{r=1}^n P_{il}/R$
– compute the log–likelihood for these probabilities,
– iterate until the maximum.

Drawing from densities

– generate pseudo random-draws from a uniform distribution,
– transform this random numbers with the quantile function of the required distribution.

ex : for the Gumbell distribution :

$$F(x) = e^{-e^{-x}} \Rightarrow F^{-1}(x) = -\ln(-\ln(x))$$

Problem : not good coverage of the relevant interval instead numerous draws are made. More deterministic methods like Halton draws may be used instead.

Halton sequence

To generate a Halton sequence, use a prime (e.g. 3). The sequence is then :

– 0 — 1/3 — 2/3,
– 0+1/9 — 1/3+1/9 — 2/3+1/9 — 0+2/9 — 1/3+2/9 — 2/3+2/9,
– 0+1/27 — 1/3++1/27 — 2/3+1/9+1/27 — 1/3+2/9+1/27 — 2/3+2/9+1/27 — 1/3+1/9+2/27 — 2/3+1/9+2/27 — 1/3+2/9+2/27 — 2/3+2/9+2/27
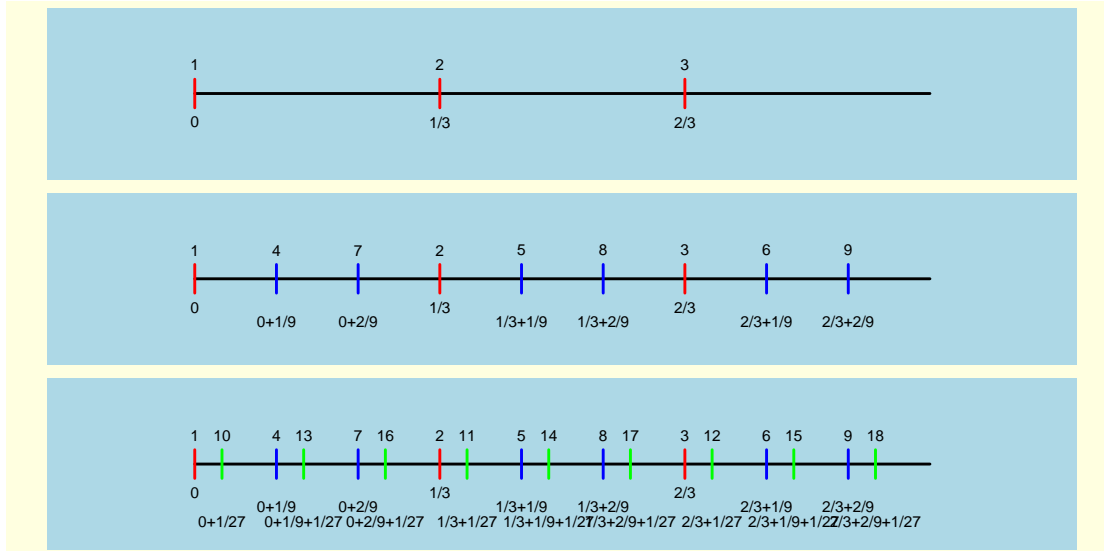
FIGURE 4 – Halton sequences

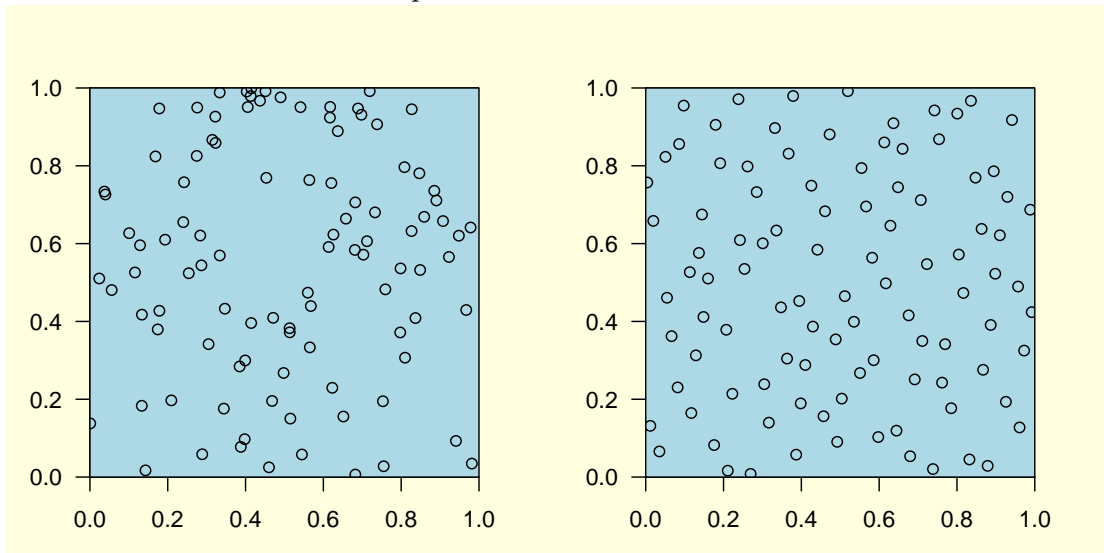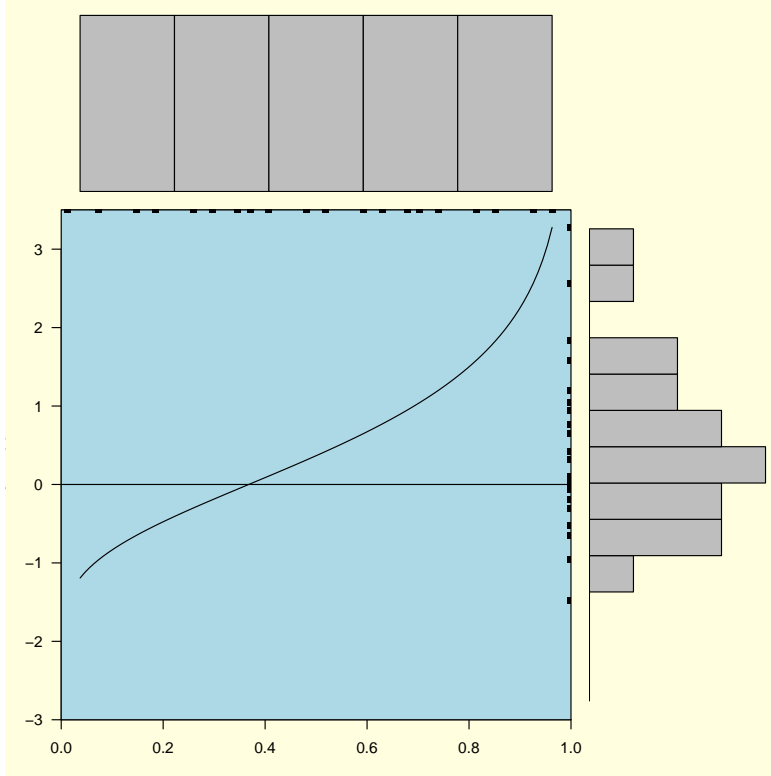FIGURE 5 – Halton sequences vs random numbers in two dimensions

FIGURE 6 – uniform to Gumbell deviates

Correlation

Cholesky decomposition is used :

Ω is the covariance matrix of two random parameters.

The Cholesky matrix is :

$$C = \begin{pmatrix} c_{11} & c_{12} \\ 0 & c_{22} \end{pmatrix}$$

so that

$$C^\top C = \begin{pmatrix} c_{11}^2 & c_{11}c_{12} \\ c_{11}c_{12} & c_{12}^2 + c_{22}^2 \end{pmatrix} = \Omega$$

if $V(\epsilon_1, \epsilon_2) = I$, then the variance of $(\epsilon_1\epsilon_2)C$ is Ω
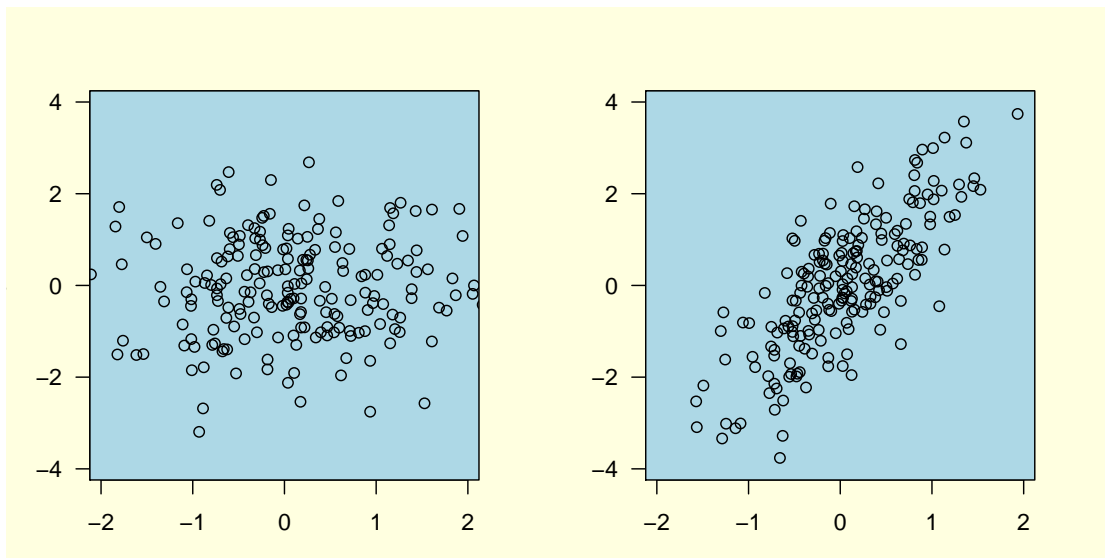
FIGURE 7 – Correlation

ex :

$$\Omega = \begin{pmatrix} 0.5 & 0.8 \\ 0.8 & 2.0 \end{pmatrix} \text{ and } C = \begin{pmatrix} 0.71 & 1.13 \\ 0 & 0.85 \end{pmatrix}$$

$$\begin{cases} \beta_1 &= 0.71\epsilon_1 \\ \beta_2 &= 1.13\epsilon_1 + 0.85\epsilon_2 \end{cases}$$

6.4. Application

The random parameter logit model is estimated by providing a rpar argument to mlogit
.

```
R> data("Train", package = "mlogit")
R> Tr <- mlogit.data(Train, shape = "wide", varying = 4:11, choice = "choice",
+     sep = "", opposite = c("price", "time", "change", "comfort"),
+     alt.levels = c("choice1", "choice2"), id = "id")
R> ml <- mlogit(choice ~ price + time + change + comfort, Tr, panel = TRUE,
+     rpar = c(time = "cn", change = "n", comfort = "ln"), correlation = TRUE,
+     R = 20, tol = 10, halton = NA)
R> summary(ml)


Call:
mlogit(formula = choice ~ price + time + change + comfort, data = Tr,
```

```
    rpar = c(time = "cn", change = "n", comfort = "ln"), R = 20,
    correlation = TRUE, halton = NA, panel = TRUE, tol = 10)

Frequencies of alternatives:
choice1 choice2
0.50324 0.49676

bfgs method
8 iterations, 0h:0m:10s
g'(-H)^-1g =  3.77
gradient close to zero

Coefficients :
                  Estimate  Std. Error  t-value  Pr(>|t|)
altchoice2      -0.02676822  0.05603835  -0.4777 0.6328803
price            0.00252142  0.00011766  21.4295 < 2.2e-16 ***
time             0.01662518  0.01041740   1.5959 0.1105099
change           0.78750820  0.08713793   9.0375 < 2.2e-16 ***
comfort          0.17129088  0.09921321   1.7265 0.0842588 .
time.time        0.16668469  0.01732761   9.6196 < 2.2e-16 ***
time.change      0.08846536  0.11847681   0.7467 0.4552511
time.comfort     0.74176133  0.07116264  10.4235 < 2.2e-16 ***
change.change    1.04445633  0.13102693   7.9713 1.554e-15 ***
change.comfort  -0.27106862  0.07498506  -3.6150 0.0003004 ***
comfort.comfort -1.39447139  0.09176197 -15.1966 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-Likelihood: -1531
McFadden R^2:  0.24587
Likelihood ratio test : chisq = 998.33 (p.value=< 2.22e-16)

random coefficients
        Min.     1st Qu.      Median        Mean    3rd Qu. Max.
time       0 0.00000000 0.01662518 0.07514065 0.1290523  Inf
change  -Inf 0.08051064 0.78750820 0.78750820 1.4945057  Inf
comfort    0 0.40267406 1.18683592 4.28624104 3.4980637  Inf
```

The summary method supplies the usual table of coefficients, and also some statistics about the random parameters. Random parameters may be extracted using the function `rpar` which take as first argument a `mlogit` object, as second argument the parameter(s) to be extracted and as third argument the coefficient that should be used for normalization. This is usually the coefficient of the price (taken as a non random parameter), so that the effects can be interpreted as monetary values. This function returns a `rpar`

object, and several methods/functions are provided to describe it :

```
R> tvalue <- rpar(ml, "time", norm = "price")
R> summary(tvalue)

     Min.   1st Qu.    Median      Mean   3rd Qu.       Max.
 0.000000  0.000000  6.593572 29.800905 51.182351        Inf

R> med(tvalue)

[1] 0.01662518

R> mean(tvalue)

[1] 0.07514065

R> stdev(tvalue)

[1] 0.1029500
```

In case of correlated random parameters further functions are provided to analyse the correlation of the coefficients :

```
R> cor.mlogit(ml)


               time      change     comfort
time     1.00000000  0.08439772   0.4628567
change   0.08439772  1.00000000  -0.1294784
comfort  0.46285666 -0.12947840   1.0000000


R> cov.mlogit(ml)


               time      change     comfort
time     0.02778379  0.01474582   0.1236403
change   0.01474582  1.09871515  -0.2174991
comfort  0.12364026 -0.21749914   2.5682385


R> stdev(ml)


      time     change    comfort
 0.1029500  1.0481961 14.8744480
```

Références

Bhat C (1995). "A heterocedastic extreme value model of intercity travel mode choice." *Transportation Research B*, **29**(6), 471–483.

Daly A (1987). "Estimating 'tree' logit models." *Transportation Research B*, pp. 251–267.

Grenne W (2008). *Econometric Analysis*. 6 edition. Prentice Hall.

Heiss F (2002). "Structural choice analysis with nested logit models." *The Stata Journal*, **2**(3), 227–252.

Koppelman FS, Wen CH (1998). "Alternative nested logit models : structure, properties and estimation." *Transportation Research B*, **32**(5), 289–298.

Koppelman FS, Wen CH (2000). "The paired combinatorial logit model : properties, estimation and application." *Transportation Research B*, **34**, 75–89.

McFadden D (1974). "The measurment of urban travel demand." *Journal of public economics*, **3**, 303–328.

McFadden D (1978). "Spatial interaction theory and planning models." In A Karlqvist (ed.), *Modeling the choice of residential location*, pp. 75–96. North-Holland, Amsterdam.

Small KA, Rosen HS (1981). "Applied welfare exonomics with discrete choice models." *Econometrica*, **49**, 105–130.

Toomet O, Henningsen A, with contributions from Spencer Graves, Croissant Y (2010). *maxLik : Maximum Likelihood Estimation*. R package version 0.8-0, URL http://CRAN.R-project.org/package=maxLik.

Wen CH, Koppelman FS (2001). "The generalized nested logit model." *Transportation Research part B*, **35**, 627–641.

Zeileis A, Croissant Y (2010). "Extended Model Formulas in R : Multiple Parts and Multiple Responses." *Journal of Statistical Software*, **34**(1), 1–13. URL http://www.jstatsoft.org/v34/i01/.

Index

Affiliation:

Yves Croissant
Faculté de Droit et d'Economie
Université de la Réunion
15, avenue René Cassin
BP 7151
F-97715 Saint-Denis Messag Cedex 9
Telephone: +33/262/938446
E-mail: yves.croissant@univ-reunion.fr