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 in different fields of applied econometrics when one wants to analyze individual data, which may be :

- revealed preferences data which means that the data are observed choices of individual for example for a transport mode (car, plane and train for example),
- stated preferences data, for example three virtual train tickets with different characteristics proposed to travelers
  - 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 the utility of each alternative depends linearly on cost (x) and price (z)

$$\left\{\begin{array}{lcl} U_1 & = & \alpha_1 + \beta x_1 + \gamma z_1 \\ U_2 & = & \alpha_2 + \beta x_2 + \gamma z_2 \\ U_3 & = & \alpha_3 + \beta x_3 + \gamma z_3 \end{array}\right.$$

The multinomial logit model is obtained simply by applying a specific transformation to the utility level so that the results may be interpreted as probabilities of choosing each alternative :

$$\left\{\begin{array}{lcl} \mathrm{P}_1 & = & \frac{e^{U_1}}{e^{U_1}+e^{U_2}+e^{U_3}} \\ \mathrm{P}_2 & = & \frac{e^{U_2}}{e^{U_1}+e^{U_2}+e^{U_3}} \\ \mathrm{P}_3 & = & \frac{e^{U_3}}{e^{U_1}+e^{U_2}+e^{U_3}} \end{array}\right.$$

The two characteristics of probabilities are satisfied :

- $0 \leq \mathrm{P}_j \leq 1$,
- $\sum_{j=1}^{3} \mathrm{P}_j = 1$

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 for example). 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 travelers, on 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 prices 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 several choices between several alternatives, the determinants of these choices being variables that can be alternative specific or purely individual specific. Such data have therefore a specific structure which can be characterized 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 column for each alternative.

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 variable. 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)
```

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

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

```
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 boolean 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 faces 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 tanks to the ordering and 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 a 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 variable :

```
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)

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

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 `formulas` 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 utility for the alternative j (or more precisely the deterministic component of utility) is then :

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

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

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

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 value 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. 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 kind of variables introduced.

`mlogit` package provides objects of class `mFormula` which are extended model formulas and which are build upon `Formula` objects provided by 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 following 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 variable (`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[2]:

- 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 function 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 include 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 individual point of view. It is random form the researcher'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 :

$$\left\{\begin{array}{ccccc} U_1 & = & \beta_1^\top x_1 + \epsilon_1 & = & V_1 + \epsilon_1 \\ U_2 & = & \beta_1^\top x_1 + \epsilon_2 & = & V_2 + \epsilon_2 \\ & \vdots & & \vdots & \\ U_J & = & \beta_J^\top x_J + \epsilon_J & = & V_J + \epsilon_J \end{array}\right.$$

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 :

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

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 :

[2] when possible, we'll omit the individual index to simplify the notations.

$$\left\{\begin{array}{lcl}\epsilon_1 & < & (V_l - V_1) + \epsilon_l \\ \epsilon_2 & < & (V_l - V_2) + \epsilon_l \\ & \vdots & \\ \epsilon_J & < & (V_l - V_J) + \epsilon_l\end{array}\right.$$

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 :

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

And the conditional (1) and unconditional (2) probabilities are just :

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

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

which means that the evaluation of only 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.

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

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 and the variance of the ϵ_js are not identified. We can then, without loss of generality suppose that $\mu_j = 0 \;\; \forall j$ and that one of the θ_j equals 1.

$$U_l = \beta_l^\top x_l + \eta_l$$

$$\frac{U_l}{\sigma} = \frac{\beta_l}{\sigma}^\top x_l + \frac{\eta_l}{\sigma} = \frac{\beta_l}{\sigma}^\top x_l + \epsilon_l$$

with $\epsilon_l = \frac{\eta_l}{\sigma}$ follows a standard Gumbel distribution

H3 identically distributed errors

As, the location 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. This common scale parameter is not identified, and therefore, we can suppose that $\theta_j = 1 \; \forall j \in 1 \ldots J$.

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

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

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

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 logit 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 mean of the previous expression weighted by the Gumbell density of ϵ_l.

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

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

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

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

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 part of them is the common influence of some omitted variables on utility.

2.5. Estimation

The coefficients of the multinomial logit model are estimated using 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.

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

y_{ij} is equal to one if individual i made choice j, 0 otherwise.
The probability of the choice made for one individual is :

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

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.

1. Start with a value β_t,
2. Approximate the function to optimize by a second order Taylor series : $l(x) = f(\beta_t) + (x - \beta_t)g(\beta_t) + 0.5(x - \beta_t)^2 h(\beta_t)$ where g and h are the first two derivatives of f,
3. find the maximum of $l(x)$. The first order condition is : $\frac{\partial l(x)}{\partial x} = g(\beta_t) + (x - \beta_t)h(\beta_t) = 0$. The solution is : $x = \beta_t - \frac{g(\beta_t)}{h(\beta_t)}$

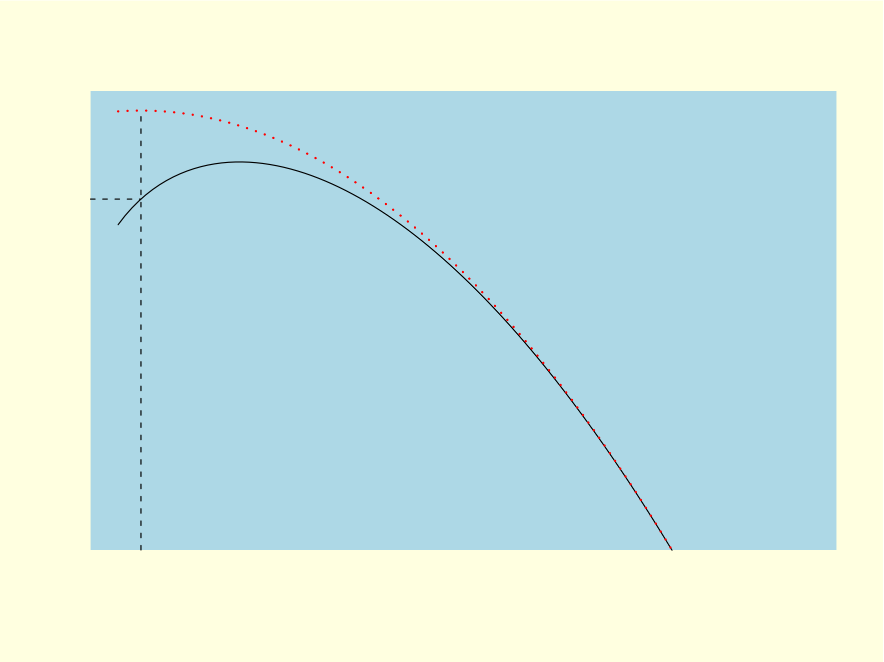


Figure 1: Numerical optimization

4. Go back to step one with that value.

Consider now a function of several variables $f(\beta)$. 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 :

$$l(x) = f(\beta_t) + (x - \beta_t)g(\beta_t) + 0.5(x - \beta_t)'H(\beta_t)(x - \beta_t)$$

The vector of first derivatives is :

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

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

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

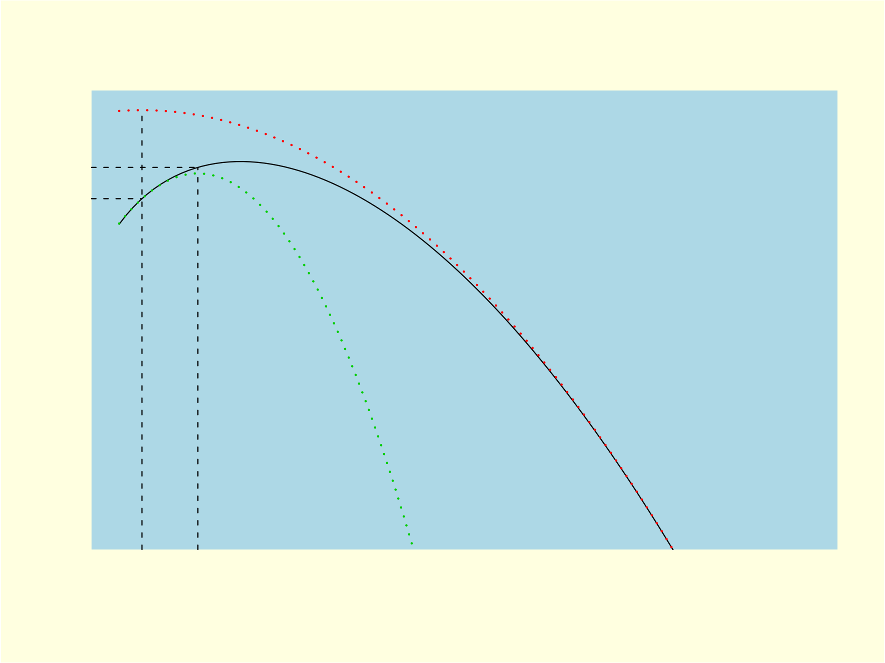


Figure 2: Numerical optimization

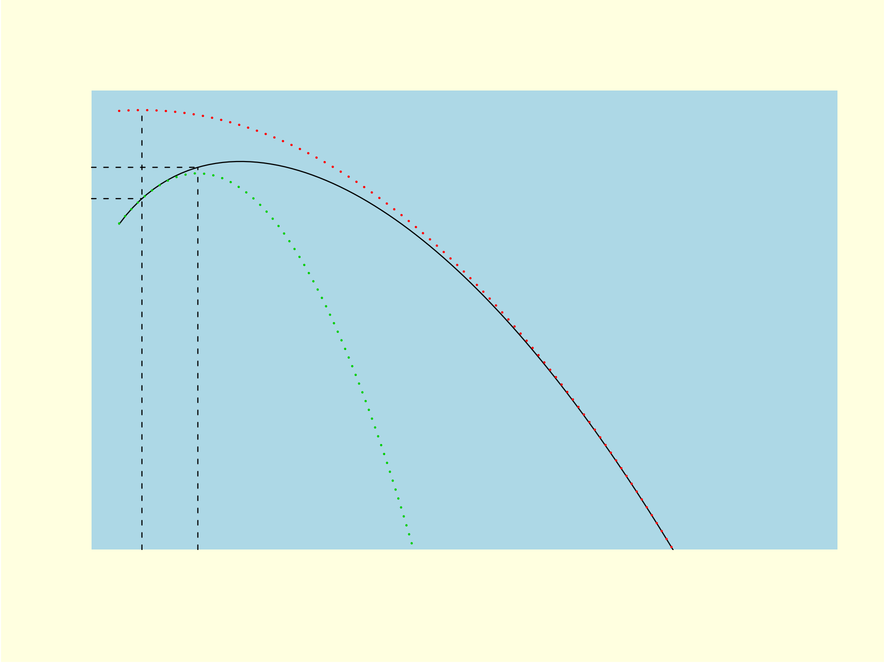


Figure 3: Numerical optimization

if the function is well-behaved, but it may performs poorly otherwise and scratch after a few iterations.

The second one, 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 (?) 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

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

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

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

2.6. Interpretation

Marginal effects

The coefficients are not directly interpretable. The marginal effects are obtained by deriving the probabilities with respect with the variables :

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

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

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

- For a choice specific variable, the sign of the coefficient is directly interpretable. The product of two probabilities is at most 0.25.

- For an individual specific variable, the sign of the coefficient is not necessarily the sign of the coefficient. Actually, it depends on the sign of $(\beta_j - \sum_l P_{il}\beta_l)$, which would be 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, 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

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 researcher'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 "sum log". 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 plan. Within this plan, some points corresponds to situations where alternative 1 is chosen and some where alternative 2 is chosen. 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.

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

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

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

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

with $a = 1 + e^{-(V_1-V_2)} = \frac{e^{V_1}+e^{V_2}}{e^{V_1}} = \frac{1}{P_1}$
Let define $z \mid e^{-z} = ae^{-\epsilon_1} \Leftrightarrow z = \epsilon_1 - \ln a$
We then have :

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

$$E_1 = (V1 + \ln a)/a + \mu/a$$

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

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 to 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:1s
g'(-H)^-1g = 1.28E-11
optimum reached

Coefficients :
          Estimate Std. Error  t-value  Pr(>|t|)
price   -0.0673581  0.0033933 -19.8506 < 2.2e-16 ***
time    -1.7205517  0.1603517 -10.7299 < 2.2e-16 ***
change  -0.3263410  0.0594892  -5.4857 4.118e-08 ***
comfort -0.9457257  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
25.543370  4.844869 14.040276
```

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:1s
g'(-H)^-1g = 4.37E-12
optimum reached

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 0.2876609
```

```
altcharter:income -7.2337e-05  5.2557e-05  -1.3764 0.1687090
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. 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.

3.1. 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

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

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

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.

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

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

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

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

3.2. 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 K different nests. This implies the following multivariate distribution for the error terms.

$$\exp\left(-\sum_{k=1}^K \left(\sum_{j \in B_k} e^{-\epsilon_j/\lambda_k}\right)^{\lambda_k}\right)$$

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

$$P_j = \frac{e^{V_j/\lambda_l}\left(\sum_{j \in B_l} e^{V_j/\lambda_l}\right)^{\lambda_l - 1}}{\sum_{k=1}^K \left(\sum_{j \in B_k} e^{V_j/\lambda_k}\right)^{\lambda_k}}$$

Let write : $V_j = Z_j + W_l$

$$P_j = \frac{e^{(Z_j+W_l)/\lambda_l}}{\sum_{j \in B_l} e^{(Z_j+W_l)/\lambda_l}} \times \frac{\left(\sum_{j \in B_l} e^{(Z_j+W_l)/\lambda_l}\right)^{\lambda_l}}{\sum_{k=1}^K \left(\sum_{j \in B_k} e^{(Z_j+W_k)/\lambda_k}\right)^{\lambda_k}}$$

$$P_j = \frac{e^{Z_j/\lambda_l}}{\sum_{j \in B_l} e^{Z_j/\lambda_l}} \times \frac{\left(\sum_{j \in B_l} e^{(Z_j+W_l)/\lambda_l}\right)^{\lambda_l}}{\sum_{k=1}^K \left(\sum_{j \in B_k} e^{(Z_j+W_k)/\lambda_k}\right)^{\lambda_k}}$$

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

with $I_l = \ln \sum_{j \in B_l} e^{Z_j/\lambda_l}$ wich is often denoted as the inclusive value or inclusive utility.

We then can write the probability of choosing alternative j as :

$$P_j = \frac{e^{Z_j/\lambda_l}}{\sum_{j \in B_l} e^{Z_j/\lambda_l}} \times \frac{e^{W_l+\lambda_l I_l}}{\sum_{k=1}^K e^{W_k+\lambda_k I_k}}$$

The first term $\mathrm{P}_{j|l}$ 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*.

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

The inclusive values link 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 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 :

$$P_j = \frac{e^{V_j} \left(\sum_{j \in B_l} e^{V_j}\right)^{\lambda_l - 1}}{\sum_{k=1}^K \left(\sum_{j \in B_k} e^{V_j}\right)^{\lambda_k}}$$

The gradient is, for the first version of the model and denoting $A_j = e^{V_k/\lambda_k}$ and $N_k = \sum_{j \in B_k} A_j$:

$$\left\{\begin{array}{ccl}
\frac{\partial \ln P_j}{\partial \beta} & = & \frac{x_j}{\lambda_k} + \frac{\lambda_k - 1}{\lambda_k}\frac{1}{N_k}\sum_{j \in B_k} A_j x_j - \frac{1}{\sum_k N_k^{\lambda_k}} \sum_k N_k^{\lambda_k - 1} \sum_{j \in B_k} A_j x_j \\
\frac{\partial \ln P_j}{\partial \lambda_k} & = & -\frac{V_j}{\lambda_k^2} + \ln N_k - \frac{\lambda_k - 1}{\lambda_k^2}\frac{1}{N_k}\sum_{j \in B_k} V_j A_j \\
 & & -\frac{1}{\sum_k N_k^{\lambda_k}}\left(\ln N_k - \frac{1}{\lambda_k N_k}\sum_{j \in B_k} V_j A_j\right)
\end{array}\right.$$

For the unscaled version, $A_l = e^{V_l}$ and the gradient is :

$$\left\{\begin{array}{ccl}
\frac{\partial \ln P_j}{\partial \beta} & = & A_j x_j + (\lambda_l - 1)\frac{1}{N_l}\sum_{j \in B_l} A_j x_j - \frac{1}{\sum_k N_k^{\lambda_k}} \sum_k \lambda_k N_k^{\lambda_k - 1} \sum_{j \in B_k} A_j x_j \\
\frac{\partial \ln P_j}{\partial \lambda_l} & = & \ln N_l - \frac{1}{\sum_k N_k^{\lambda_k}} N_l^{\lambda_l} \ln N_l
\end{array}\right.$$

To illustrate the estimation of nested logit models, we use an application presented by Kenneth Train. The data consists on 250 newly built houses in California, and we seek to explain the heating system chosen. The data is available in `mlogit` under the name `HC`. Seven heating modes are available :

gcc gas central heat with cooling,

ecc electric central resistance heat with cooling,

erc electric room resistance heat with cooling,

hpc electric heat pump which provides cooling also,

gc gaz central heat without cooling,

ec electric central resistance heat without cooling,

er electric room resistance heat without cooling.

The covariates are the installation cost (`ich`), the operating cost (`och`) and the income of the household. This data set has a natural nesting structure, the first four modes providing also cooling whereas the three other modes being "pure" heating modes. For the cooling mode, the installation and operating cost for the cooling part (`icca` and `occa` should be added.

```
R> data("HC", package = "mlogit")
R> HC <- mlogit.data(HC, varying = c(2:8, 10:16), choice = "depvar",
+     shape = "wide")
R> cooling.modes <- attr(HC, "index")$alt %in% c("gcc", "ecc", "erc",
+     "hpc")
R> room.modes <- attr(HC, "index")$alt %in% c("erc", "er")
R> HC$icca[!cooling.modes] <- 0
R> HC$occa[!cooling.modes] <- 0
R> HC$icca <- HC$icca/100
R> HC$occa <- HC$occa/100
R> HC$ich <- HC$ich/100
R> HC$och <- HC$och/100
R> HC$inc.cooling <- HC$inc.room <- 0
R> HC$inc.cooling[cooling.modes] <- HC$income[cooling.modes]
R> HC$inc.room[room.modes] <- HC$income[room.modes]
R> HC$int.cooling <- as.numeric(cooling.modes)
R> nl <- mlogit(depvar ~ ich + och + icca + occa + inc.room + inc.cooling +
+     int.cooling | 0, HC, nests = list(cooling = c("gcc", "ecc",
+     "erc", "hpc"), other = c("gc", "ec", "er")), un.nest.el = TRUE)
```

```
Initial value of the function : 180.286442614203
iteration 1, step = 1, lnL = 179.72644009, chi2 = 3.89167036
iteration 2, step = 1, lnL = 178.49652314, chi2 = 4.54053221
iteration 3, step = 1, lnL = 178.18517025, chi2 = 0.42915312
iteration 4, step = 0.5, lnL = 178.14799363, chi2 = 0.20072201
iteration 5, step = 0.5, lnL = 178.13242191, chi2 = 0.10419337
iteration 6, step = 1, lnL = 178.1270763, chi2 = 0.01763203
iteration 7, step = 1, lnL = 178.12550886, chi2 = 0.00562972
iteration 8, step = 1, lnL = 178.12477236, chi2 = 0.00146782
iteration 9, step = 1, lnL = 178.12474273, chi2 = 4.802e-05
iteration 10, step = 1, lnL = 178.12473902, chi2 = 7.26e-06
iteration 11, step = 1, lnL = 178.12473901, chi2 = 2e-08
```

4. 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. G has the following characteristics :

- all of its arguments are non-negative,
- it is non negative,
- it is homogeneous of degree 1 in all its arguments,
- for all its argument, $\lim_{y_j \rightarrow +\infty} = G(y_1, \dots y_J) = +\infty$,
- for distinct arguments, $\frac{\partial^k G}{\partial y_i, \dots, y_j}$ 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 :

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

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 $\epsilon \Rightarrow -\infty$, we should have $F \Rightarrow 0$,
2. if F is a joint cumulative distribution of probability, for all $\epsilon \rightarrow +\infty$, we should have $F \rightarrow 1$,
3. if F is a multivariate extreme value distribution, the marginal distribution of any ϵ should be an extreme value distribution.

For point 1, if $\epsilon_j \rightarrow -\infty$, $y_j \rightarrow +\infty$, $G \rightarrow +\infty$ and then $F \rightarrow 0$.
For point 2, if $(\epsilon_1, \dots, \epsilon_J) \rightarrow +\infty$, $G \rightarrow 0$ and then $F \rightarrow 1$.
To demonstrate the third point, we compute the marginal cumulative distribution of ϵ_l which is :

$$F(\epsilon_l) = \lim_{\epsilon_j \rightarrow +\infty \forall j \neq l} F(\epsilon_1, \dots, \epsilon_l, \dots \epsilon_J) =$$

$$F(\epsilon_l) = \exp\left(-G\left(0, \dots, e^{-\epsilon_l}, \dots, 0\right)\right)$$

with G being homogeneous of degree one, we have :

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

with $a_l = G(0, \dots, 1, \dots, 0)$. 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 l^{th} argument. The derivative of F respective to the ϵ_l is then :

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

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 :

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

G being homogeneous of degree one, one can write :

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

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

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

The probability of choosing alternative i is then :

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

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

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

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

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

5. 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.

5.1. 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.

5.2. Panel data

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

Joint probabilities for each individual are computed.

$$P_{ikl}^r = \frac{e^{\beta_i^r x_{ikl}}}{\sum_j e^{\beta_i^r x_{ikj}}}$$

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

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

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

5.3. 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 : $P_{il}^r = \frac{e^{\beta_i^r x_{il}}}{\sum_j e^{\beta_i^r x_{ij}}}$
- compute the average of these probabilities : $\bar{\mathrm{P}}_{il} = \sum_{r=1}^n \mathrm{P}_{il}/R$
- compute the log–likelihood for these probabilities,
- iterate until the maximum.

Drawing from densities

- use `runif` to 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 :

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

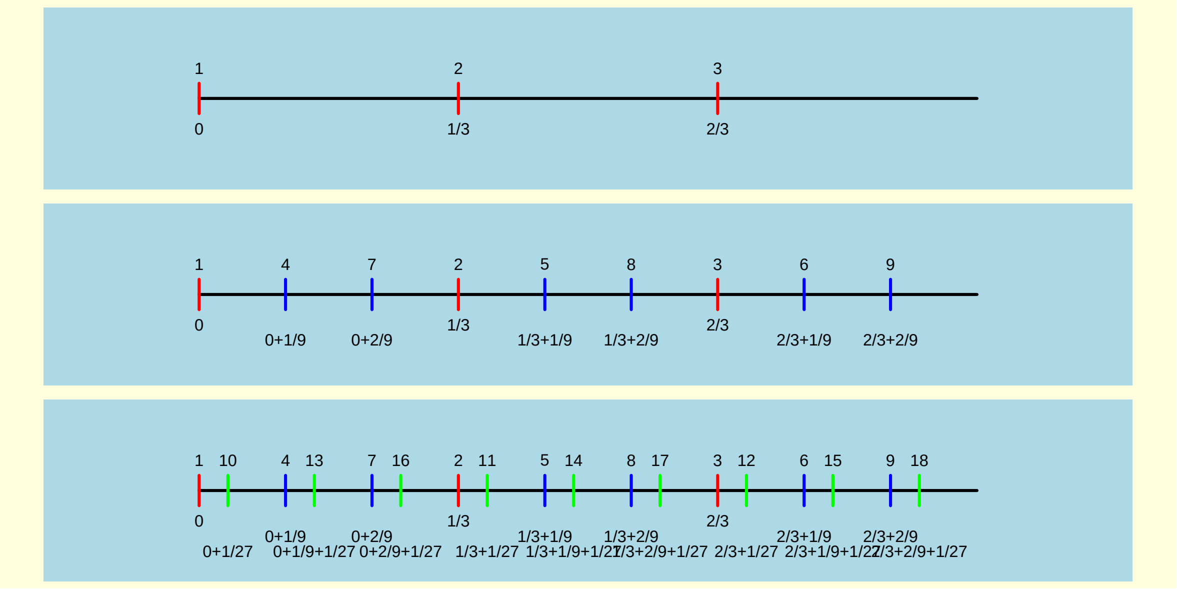
Problem : not good coverage of the relevant interval instead numerous draws are made. More deterministic methods like Halton draws may be used instead.

Halton sequence

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

- 0 — 1/3 — 2/3,

Figure 4: Halton sequences



- 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

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 $\mathrm{V}(\epsilon_1, \epsilon_2) = I$, then the variance of $(\epsilon_1 \epsilon_2)C$ is Ω

ex :

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

Figure 5: Halton sequences vs random numbers in two dimensions

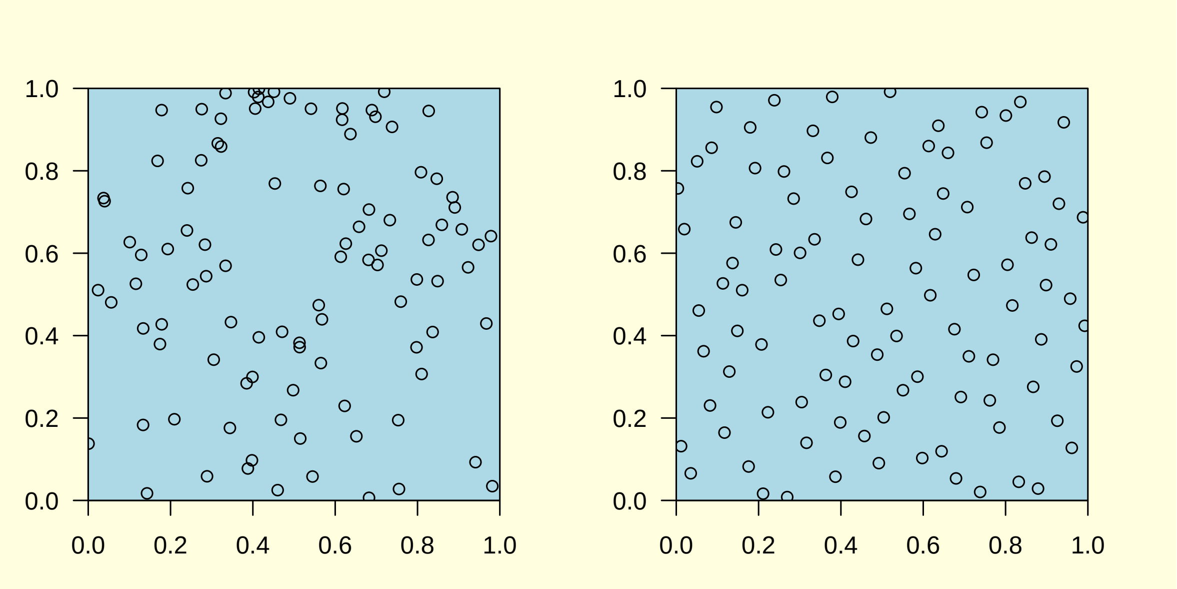
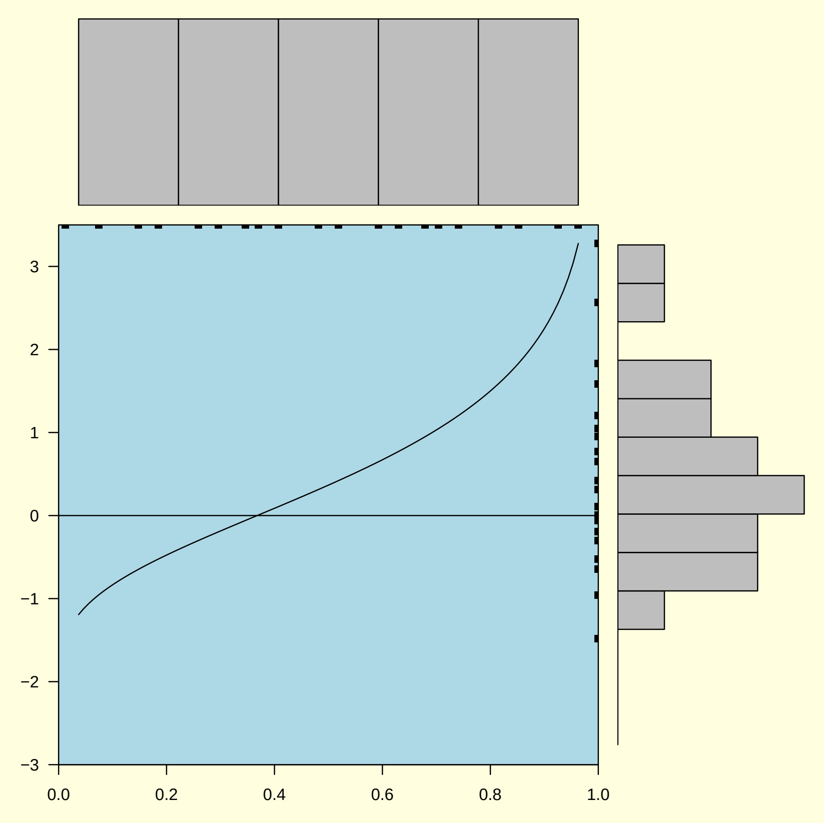


Figure 6: uniform to Gumbell deviates



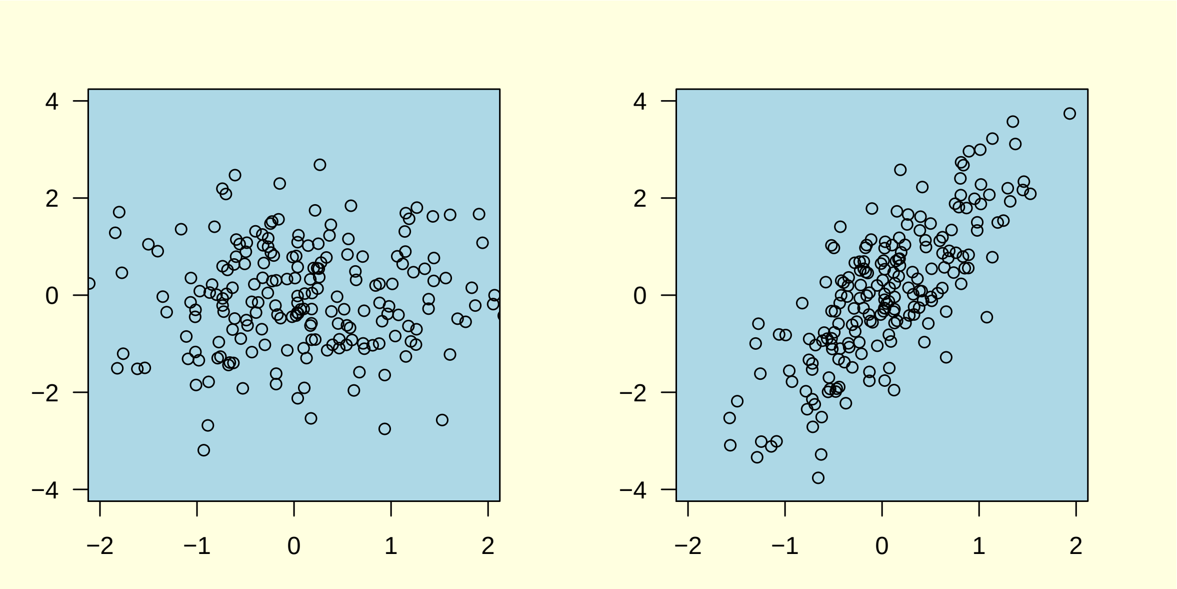


Figure 7: Correlation

$$\left\{ \begin{array}{lcl} \beta_1 & = & 0.71\epsilon_1 \\ \beta_2 & = & 1.13\epsilon_1 + 0.85\epsilon_2 \end{array} \right.$$

References

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

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.

Affiliation:

Yves Croissant
LET-ISH
Avenue Berthelot
F-69363 Lyon cedex 07
Telephone: +33/4/78727249
Fax: +33/4/78727248
E-mail: yves.croissant@let.ish-lyon.cnrs.fr